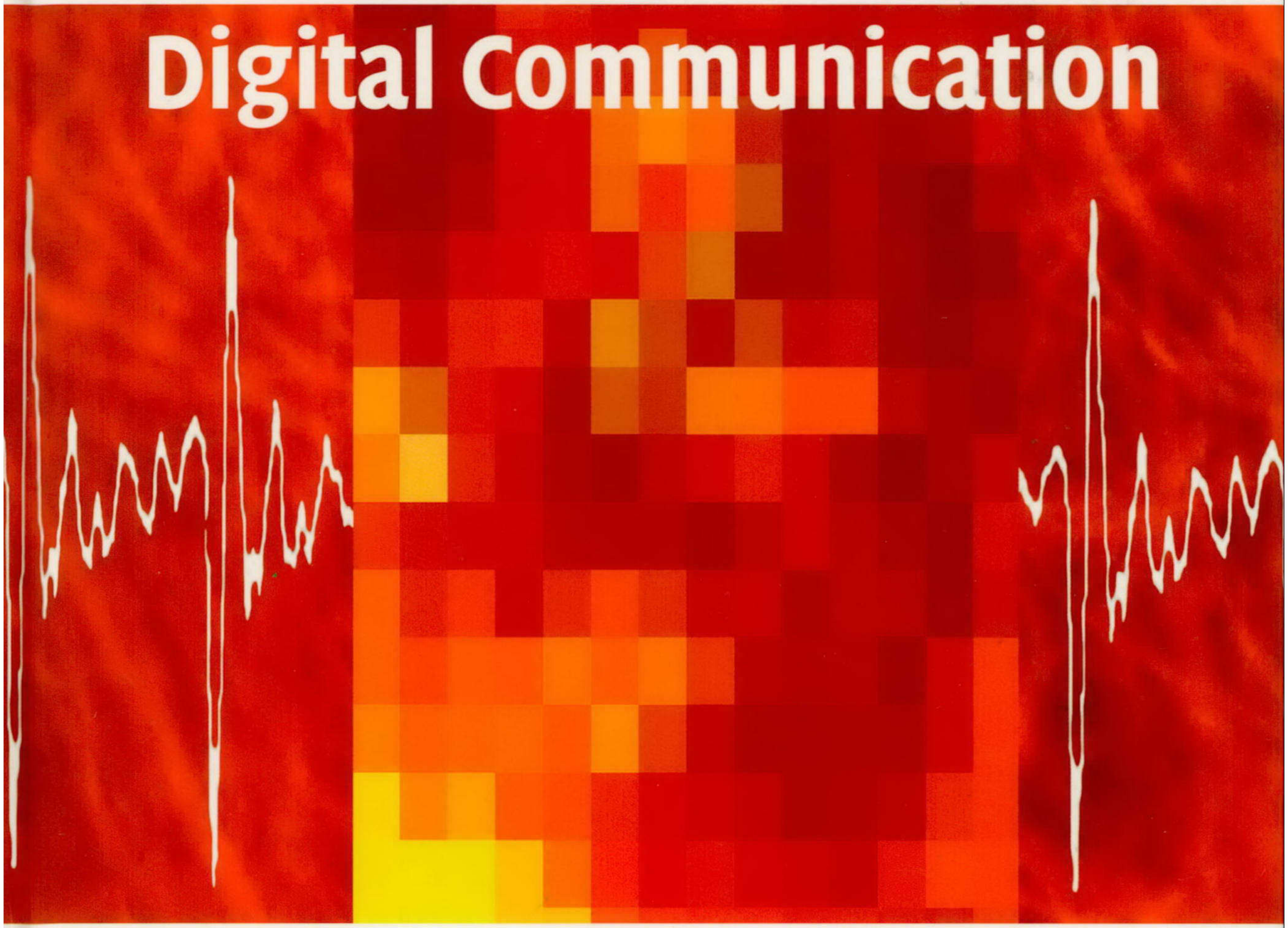
Principles of Digital Communication



Robert G. Gallager

CAMBRIDGE

Gallager

Principles of **Digital Communication**

CAMBRIDGE

Principles of Digital Communication

The renowned communication theorist Robert Gallager brings his lucid writing style to this first-year graduate textbook on the fundamental system aspects of digital communication. With the clarity and insight that have characterized his teaching and earlier textbooks he develops a simple framework and then combines this with careful proofs to help the reader understand modern systems and simplified models in an intuitive yet precise way. Although many features of various modern digital communication systems are discussed, the focus is always on principles explained using a hierarchy of simple models.

A major simplifying principle of digital communication is to separate source coding and channel coding by a standard binary interface. Data compression, i.e., source coding, is then treated as the conversion of arbitrary communication sources into binary data streams. Similarly digital modulation, i.e., channel coding, becomes the conversion of binary data into waveforms suitable for transmission over communication channels. These waveforms are viewed as vectors in signal space, modeled mathematically as Hilbert space.

A self-contained introduction to random processes is used to model the noise and interference in communication channels. The principles of detection and decoding are then developed to extract the transmitted data from noisy received waveforms. An introduction to coding and coded modulation then leads to Shannon's noisy-channel coding theorem. The final topic is wireless communication. After developing models to explain various aspects of fading, there is a case study of cellular CDMA communication which illustrates the major principles of digital communication.

Throughout, principles are developed with both mathematical precision and intuitive explanations, allowing readers to choose their own mix of mathematics and engineering. An extensive set of exercises ranges from confidence-building examples to more challenging problems. Instructor solutions and other resources are available at www.cambridge.org/9780521879071.

'Prof. Gallager is a legendary figure . . . known for his insights and excellent style of exposition'
Professor Lang Tong, Cornell University

'a compelling read'
Professor Emre Telatar, EPFL

'It is surely going to be a classic in the field'
Professor Hideki Imai, University of Tokyo

Robert G. Gallager has had a profound influence on the development of modern digital communication systems through his research and teaching. As a Professor at M.I.T. since 1960 in the areas of information theory, communication technology, and data networks. He is a member of the U.S. National Academy of Engineering, the U.S. National Academy of Sciences, and, among many honors, received the IEEE Medal of Honor in 1990 and the Marconi prize in 2003. This text has been his central academic passion over recent years.

Principles of Digital Communication

ROBERT G. GALLAGER
Massachusetts Institute of Technology

CAMBRIDGE
UNIVERSITY PRESS

CAMBRIDGE UNIVERSITY PRESS
Cambridge, New York, Melbourne, Madrid, Cape Town, Singapore, São Paulo, Delhi

Cambridge University Press
The Edinburgh Building, Cambridge CB2 8RU, UK

Published in the United States of America by Cambridge University Press, New York

www.cambridge.org

First published 2008

Printed in the United Kingdom at the University Press, Cambridge

A catalog record for this publication is available from the British Library

ISBN 978-0-521-87907-1 hardback

1 Introduction to digital communication

Communication has been one of the deepest needs of the human race throughout recorded history. It is essential to forming social unions, to educating the young, and to expressing a myriad of emotions and needs. Good communication is central to a civilized society.

The various communication disciplines in engineering have the purpose of providing technological aids to human communication. One could view the smoke signals and drum rolls of primitive societies as being technological aids to communication, but communication technology as we view it today became important with telegraphy, then telephony, then video, then computer communication, and today the amazing mixture of all of these in inexpensive, small portable devices.

Initially these technologies were developed as separate networks and were viewed as having little in common. As these networks grew, however, the fact that all parts of a given network had to work together, coupled with the fact that different components were developed at different times using different design methodologies, caused an increased focus on the underlying principles and architectural understanding required for continued system evolution.

This need for basic principles was probably best understood at American Telephone and Telegraph (AT&T), where Bell Laboratories was created as the research and development arm of AT&T. The Math Center at Bell Labs became the predominant center for communication research in the world, and held that position until quite recently. The central core of the principles of communication technology were developed at that center.

Perhaps the greatest contribution from the Math Center was the creation of Information Theory [27] by Claude Shannon (Shannon, 1948). For perhaps the first 25 years of its existence, Information Theory was regarded as a beautiful theory but not as a central guide to the architecture and design of communication systems. After that time, however, both the device technology and the engineering understanding of the theory were sufficient to enable system development to follow information theoretic principles.

A number of information theoretic ideas and how they affect communication system design will be explained carefully in subsequent chapters. One pair of ideas, however, is central to almost every topic. The first is to view all communication sources, e.g., speech waveforms, image waveforms, and text files, as being representable by binary sequences. The second is to design communication systems that first convert the

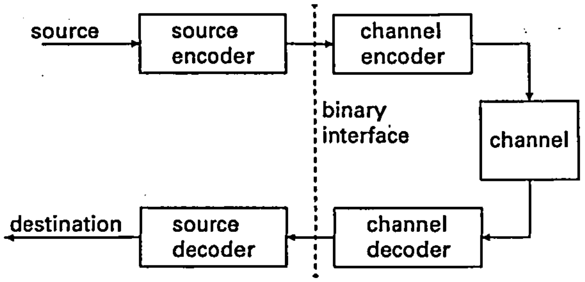


Figure 1.1. Placing a binary interface between source and channel. The source encoder converts the source output to a binary sequence and the channel encoder (often called a modulator) processes the binary sequence for transmission over the channel. The channel decoder (demodulator) recreates the incoming binary sequence (hopefully reliably), and the source decoder recreates the source output.

source output into a binary sequence and then convert that binary sequence into a form suitable for transmission over particular physical media such as cable, twisted wire pair, optical fiber, or electromagnetic radiation through space.

Digital communication systems, by definition, are communication systems that use such a digital[1] sequence as an interface between the source and the channel input (and similarly between the channel output and final destination) (see Figure 1.1).

The idea of converting an analog source output to a binary sequence was quite revolutionary in 1948, and the notion that this should be done before channel processing was even more revolutionary. Today, with digital cameras, digital video, digital voice, etc., the idea of digitizing any kind of source is commonplace even among the most technophobic. The notion of a binary interface before channel transmission is almost as commonplace. For example, we all refer to the speed of our Internet connection in bits per second.

There are a number of reasons why communication systems now usually contain a binary interface between source and channel (i.e., why digital communication systems are now standard). These will be explained with the necessary qualifications later, but briefly they are as follows.

- Digital hardware has become so cheap, reliable, and miniaturized that digital interfaces are eminently practical.
- A standardized binary interface between source and channel simplifies implementation and understanding, since source coding/decoding can be done independently of the channel, and, similarly, channel coding/decoding can be done independently of the source.

[1] A digital sequence is a sequence made up of elements from a finite alphabet (e.g. the binary digits $\{0, 1\}$, the decimal digits $\{0, 1, \ldots, 9\}$, or the letters of the English alphabet). The binary digits are almost universally used for digital communication and storage, so we only distinguish digital from binary in those few places where the difference is significant.

- A standardized binary interface between source and channel simplifies networking, which now reduces to sending binary sequences through the network.
- One of the most important of Shannon's information theoretic results is that if a source can be transmitted over a channel in any way at all, it can be transmitted using a binary interface between source and channel. This is known as the *source/channel separation theorem.*

In the remainder of this chapter, the problems of source coding and decoding and channel coding and decoding are briefly introduced. First, however, the notion of layering in a communication system is introduced. One particularly important example of layering was introduced in Figure 1.1, where source coding and decoding are viewed as one layer and channel coding and decoding are viewed as another layer.

1.1 Standardized interfaces and layering

Large communication systems such as the Public Switched Telephone Network (PSTN) and the Internet have incredible complexity, made up of an enormous variety of equipment made by different manufacturers at different times following different design principles. Such complex networks need to be based on some simple architectural principles in order to be understood, managed, and maintained. Two such fundamental architectural principles are *standardized interfaces* and *layering.*

A standardized interface allows the user or equipment on one side of the interface to ignore all details about the other side of the interface except for certain specified interface characteristics. For example, the binary interface[2] in Figure 1.1 allows the source coding/decoding to be done independently of the channel coding/decoding.

The idea of layering in communication systems is to break up communication functions into a string of separate layers, as illustrated in Figure 1.2.

Each layer consists of an input module at the input end of a communcation system and a "peer" output module at the other end. The input module at layer i processes the information received from layer $i+1$ and sends the processed information on to layer $i-1$. The peer output module at layer i works in the opposite direction, processing the received information from layer $i-1$ and sending it on to layer i.

As an example, an input module might receive a voice waveform from the next higher layer and convert the waveform into a binary data sequence that is passed on to the next lower layer. The output peer module would receive a binary sequence from the next lower layer at the output and convert it back to a speech waveform.

As another example, a *modem* consists of an input module (a modulator) and an output module (a demodulator). The modulator receives a binary sequence from the next higher input layer and generates a corresponding modulated waveform for transmission over a channel. The peer module is the remote demodulator at the other end of the channel. It receives a more or less faithful replica of the transmitted

[2] The use of a binary sequence at the interface is not quite enough to specify it, as will be discussed later.

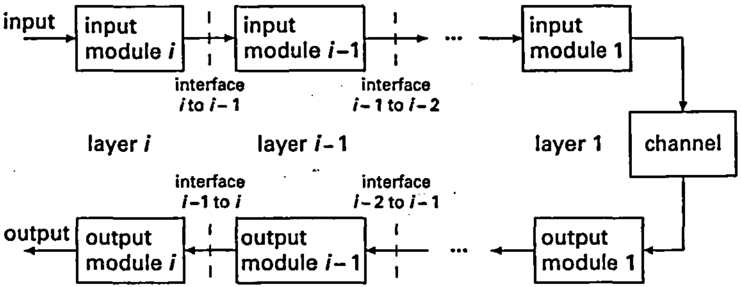


Figure 1.2. Layers and interfaces. The specification of the interface between layers i and $i-1$ should specify how input module i communicates with input module $i-1$, how the corresponding output modules communicate, and, most important, the input/output behavior of the system to the right of the interface. The designer of layer $i-1$ uses the input/output behavior of the layers to the right of $i-1$ to produce the required input/output performance to the right of layer i. Later examples will show how this multilayer process can simplify the overall system design.

waveform and reconstructs a typically faithful replica of the binary sequence. Similarly, the local demodulator is the peer to a remote modulator (often collocated with the remote demodulator above). Thus a modem is an input module for communication in one direction and an output module for independent communication in the opposite direction. Later chapters consider modems in much greater depth, including how noise affects the channel waveform and how that affects the reliability of the recovered binary sequence at the output. For now, however, it is enough simply to view the modulator as converting a binary sequence to a waveform, with the peer demodulator converting the waveform back to the binary sequence.

As another example, the source coding/decoding layer for a waveform source can be split into three layers, as shown in Figure 1.3. One of the advantages of this layering is that discrete sources are an important topic in their own right (discussed in Chapter 2) and correspond to the inner layer of Figure 1.3. Quantization is also an important topic in its own right (discussed in Chapter 3). After both of these are understood, waveform sources become quite simple to understand.

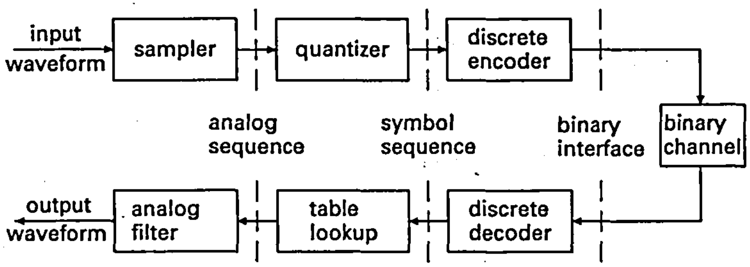


Figure 1.3. Breaking the source coding/decoding layer into three layers for a waveform source. The input side of the outermost layer converts the waveform into a sequence of samples and the output side converts the recovered samples back to the waveform. The quantizer then converts each sample into one of a finite set of symbols, and the peer module recreates the sample (with some distortion). Finally the inner layer encodes the sequence of symbols into binary digits.

The channel coding/decoding layer can also be split into several layers, but there are a number of ways to do this which will be discussed later. For example, binary error-correction coding/decoding can be used as an outer layer with modulation and demodulation as an inner layer, but it will be seen later that there are a number of advantages in combining these layers into what is called coded modulation.[3] Even here, however, layering is important, but the layers are defined differently for different purposes.

It should be emphasized that layering is much more than simply breaking a system into components. The input and peer output in each layer encapsulate all the lower layers, and all these lower layers can be viewed in aggregate as a communication channel. Similarly, the higher layers can be viewed in aggregate as a simple source and destination.

The above discussion of layering implicitly assumed a point-to-point communication system with one source, one channel, and one destination. Network situations can be considerably more complex. With broadcasting, an input module at one layer may have multiple peer output modules. Similarly, in multiaccess communication a multiplicity of input modules have a single peer output module. It is also possible in network situations for a single module at one level to interface with multiple modules at the next lower layer or the next higher layer. The use of layering is at least as important for networks as it is for point-to-point communications systems. The physical layer for networks is essentially the channel encoding/decoding layer discussed here, but textbooks on networks rarely discuss these physical layer issues in depth. The network control issues at other layers are largely separable from the physical layer communication issues stressed here. The reader is referred to Bertsekas and Gallager (1992), for example, for a treatment of these control issues.

The following three sections provide a fuller discussion of the components of Figure 1.1, i.e. of the fundamental two layers (source coding/decoding and channel coding/decoding) of a point-to-point digital communication system, and finally of the interface between them.

1.2 Communication sources

The source might be discrete, i.e. it might produce a sequence of discrete symbols, such as letters from the English or Chinese alphabet, binary symbols from a computer file, etc. Alternatively, the source might produce an analog waveform, such as a voice signal from a microphone, the output of a sensor, a video waveform, etc. Or, it might be a sequence of images such as X-rays, photographs, etc.

Whatever the nature of the source, the output from the source will be modeled as a sample function of a random process. It is not obvious why the inputs to communication

[3] Terminology is nonstandard here. A channel coder (including both coding and modulation) is often referred to (both here and elsewhere) as a modulator. It is also often referred to as a modem, although a modem is really a device that contains both modulator for communication in one direction and demodulator for communication in the other.

systems should be modeled as random, and in fact this was not appreciated before Shannon developed information theory in 1948.

The study of communication before 1948 (and much of it well after 1948) was based on Fourier analysis; basically one studied the effect of passing sine waves through various kinds of systems and components and viewed the source signal as a superposition of sine waves. Our study of channels will begin with this kind of analysis (often called Nyquist theory) to develop basic results about sampling, intersymbol interference, and bandwidth.

Shannon's view, however, was that if the recipient knows that a sine wave of a given frequency is to be communicated, why not simply regenerate it at the output rather than send it over a long distance? Or, if the recipient knows that a sine wave of unknown frequency is to be communicated, why not simply send the frequency rather than the entire waveform?

The essence of Shannon's viewpoint is that the set of possible source outputs, rather than any particular output, is of primary interest. The reason is that the communication system must be designed to communicate whichever one of these possible source outputs actually occurs. The objective of the communication system then is to transform each possible source output into a transmitted signal in such a way that these possible transmitted signals can be best distinguished at the channel output. A probability measure is needed on this set of possible source outputs to distinguish the typical from the atypical. This point of view drives the discussion of all components of communication systems throughout this text.

1.2.1 Source coding

The source encoder in Figure 1.1 has the function of converting the input from its original form into a sequence of bits. As discussed before, the major reasons for this almost universal conversion to a bit sequence are as follows: inexpensive digital hardware, standardized interfaces, layering, and the source/channel separation theorem.

The simplest source coding techniques apply to discrete sources and simply involve representing each successive source symbol by a sequence of binary digits. For example, letters from the 27-symbol English alphabet (including a SPACE symbol) may be encoded into 5-bit blocks. Since there are 32 distinct 5-bit blocks, each letter may be mapped into a distinct 5-bit block with a few blocks left over for control or other symbols. Similarly, upper-case letters, lower-case letters, and a great many special symbols may be converted into 8-bit blocks ("bytes") using the standard ASCII code.

Chapter 2 treats coding for discrete sources and generalizes the above techniques in many ways. For example, the input symbols might first be segmented into m-tuples, which are then mapped into blocks of binary digits. More generally, the blocks of binary digits can be generalized into variable-length sequences of binary digits. We shall find that any given discrete source, characterized by its alphabet and probabilistic description, has a quantity called *entropy* associated with it. Shannon showed that this source entropy is equal to the minimum number of binary digits per source symbol required to map the source output into binary digits in such a way that the source symbols may be retrieved from the encoded sequence.

Some discrete sources generate finite segments of symbols, such as email messages, that are statistically unrelated to other finite segments that might be generated at other times. Other discrete sources, such as the output from a digital sensor, generate a virtually unending sequence of symbols with a given statistical characterization. The simpler models of Chapter 2 will correspond to the latter type of source, but the discussion of universal source coding in Section 2.9 is sufficiently general to cover both types of sources and virtually any other kind of source.

The most straightforward approach to analog source coding is called analog to digital (A/D) conversion. The source waveform is first sampled at a sufficiently high rate (called the "Nyquist rate"). Each sample is then quantized sufficiently finely for adequate reproduction. For example, in standard voice telephony, the voice waveform is sampled 8000 times per second; each sample is then quantized into one of 256 levels and represented by an 8-bit byte. This yields a source coding bit rate of 64 kilobits per second (kbps).

Beyond the basic objective of conversion to bits, the source encoder often has the further objective of doing this as efficiently as possible – i.e. transmitting as few bits as possible, subject to the need to reconstruct the input adequately at the output. In this case source encoding is often called data compression. For example, modern speech coders can encode telephone-quality speech at bit rates of the order of 6–16 kbps rather than 64 kbps.

The problems of sampling and quantization are largely separable. Chapter 3 develops the basic principles of quantization. As with discrete source coding, it is possible to quantize each sample separately, but it is frequently preferable to segment the samples into blocks of n and then quantize the resulting n-tuples. As will be shown later, it is also often preferable to view the quantizer output as a discrete source output and then to use the principles of Chapter 2 to encode the quantized symbols. This is another example of layering.

Sampling is one of the topics in Chapter 4. The purpose of sampling is to convert the analog source into a sequence of real-valued numbers, i.e. into a discrete-time, analog-amplitude source. There are many other ways, beyond sampling, of converting an analog source to a discrete-time source. A general approach, which includes sampling as a special case, is to expand the source waveform into an orthonormal expansion and use the coefficients of that expansion to represent the source output. The theory of orthonormal expansions is a major topic of Chapter 4. It forms the basis for the signal space approach to channel encoding/decoding. Thus Chapter 4 provides us with the basis for dealing with waveforms for both sources and channels.

1.3 Communication channels

Next we discuss the channel and channel coding in a generic digital communication system.

In general, a channel is viewed as that part of the communication system between source and destination that is given and not under the control of the designer. Thus, to a source-code designer, the channel might be a digital channel with binary input and output; to a telephone-line modem designer, it might be a 4 kHz voice channel; to a cable modem designer, it might be a physical coaxial cable of up to a certain length, with certain bandwidth restrictions.

When the channel is taken to be the physical medium, the amplifiers, antennas, lasers, etc. that couple the encoded waveform to the physical medium might be regarded as part of the channel or as as part of the channel encoder. It is more common to view these coupling devices as part of the channel, since their design is quite separable from that of the rest of the channel encoder. This, of course, is another example of layering.

Channel encoding and decoding when the channel is the physical medium (either with or without amplifiers, antennas, lasers, etc.) is usually called *(digital) modulation* and *demodulation*, respectively. The terminology comes from the days of analog communication where modulation referred to the process of combining a lowpass signal waveform with a high-frequency sinusoid, thus placing the signal waveform in a frequency band appropriate for transmission and regulatory requirements. The analog signal waveform could modulate the amplitude, frequency, or phase, for example, of the sinusoid, but, in any case, the original waveform (in the absence of noise) could be retrieved by demodulation.

As digital communication has increasingly replaced analog communication, the modulation/demodulation terminology has remained, but now refers to the entire process of digital encoding and decoding. In most cases, the binary sequence is first converted to a baseband waveform and the resulting baseband waveform is converted to bandpass by the same type of procedure used for analog modulation. As will be seen, the challenging part of this problem is the conversion of binary data to baseband waveforms. Nonetheless, this entire process will be referred to as modulation and demodulation, and the conversion of baseband to passband and back will be referred to as frequency conversion.

As in the study of any type of system, a channel is usually viewed in terms of its possible inputs, its possible outputs, and a description of how the input affects the output. This description is usually probabilistic. If a channel were simply a linear time-invariant system (e.g. a filter), it could be completely characterized by its impulse response or frequency response. However, the channels here (and channels in practice) always have an extra ingredient – noise.

Suppose that there were no noise and a single input voltage level could be communicated exactly. Then, representing that voltage level by its infinite binary expansion, it would be possible in principle to transmit an infinite number of binary digits by transmitting a single real number. This is ridiculous in practice, of course, precisely because noise limits the number of bits that can be reliably distinguished. Again, it was Shannon, in 1948, who realized that noise provides the fundamental limitation to performance in communication systems.

The most common channel model involves a waveform input $X(t)$, an added noise waveform $Z(t)$, and a waveform output $Y(t) = X(t) + Z(t)$ that is the sum of the input

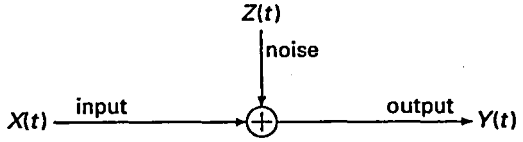


Figure 1.4. Additive white Gaussian noise (AWGN) channel.

and the noise, as shown in Figure 1.4. Each of these waveforms are viewed as random processes. Random processes are studied in Chapter 7, but for now they can be viewed intuitively as waveforms selected in some probabilitistic way. The noise $Z(t)$ is often modeled as white Gaussian noise (also to be studied and explained later). The input is usually constrained in power and bandwidth.

Observe that for any channel with input $X(t)$ and output $Y(t)$, the noise could be defined to be $Z(t) = Y(t) - X(t)$. Thus there must be something more to an additive-noise channel model than what is expressed in Figure 1.4. The additional required ingredient for noise to be called additive is that its probabilistic characterization does not depend on the input.

In a somewhat more general model, called a *linear Gaussian channel*, the input waveform $X(t)$ is first filtered in a linear filter with impulse response $h(t)$, and then independent white Gaussian noise $Z(t)$ is added, as shown in Figure 1.5, so that the channel output is given by

$$Y(t) = X(t) * h(t) + Z(t),$$

where "*" denotes convolution. Note that Y at time t is a function of X over a range of times, i.e.

$$Y(t) = \int_{-\infty}^{\infty} X(t-\tau)h(\tau)\mathrm{d}\tau + Z(t).$$

The linear Gaussian channel is often a good model for wireline communication and for line-of-sight wireless communication. When engineers, journals, or texts fail to describe the channel of interest, this model is a good bet.

The linear Gaussian channel is a rather poor model for non-line-of-sight mobile communication. Here, multiple paths usually exist from source to destination. Mobility of the source, destination, or reflecting bodies can cause these paths to change in time in a way best modeled as random. A better model for mobile communication is to

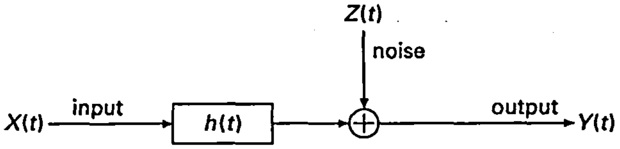


Figure 1.5. Linear Gaussian channel model.

replace the time-invariant filter $h(t)$ in Figure 1.5 by a randomly time varying linear filter, $H(t,\tau)$, that represents the multiple paths as they change in time. Here the output is given by

$$Y(t)=\int_{-\infty}^{\infty} X(t-u)H(u,t)\mathrm{d}u+Z(t).$$

These randomly varying channels will be studied in Chapter 9.

1.3.1 Channel encoding (modulation)

The channel encoder box in Figure 1.1 has the function of mapping the binary sequence at the source/channel interface into a channel waveform. A particularly simple approach to this is called binary pulse amplitude modulation (2-PAM). Let $\{u_1, u_2, \dots ,\}$ denote the incoming binary sequence, and let each $u_n = \pm 1$ (rather than the traditional 0/1). Let $p(t)$ be a given elementary waveform such as a rectangular pulse or a $\sin(\omega t)/\omega t$ function. Assuming that the binary digits enter at R bps, the sequence $u_1, u_2, \dots$ is mapped into the waveform $\sum_n u_n p(t-n/R)$.

Even with this trivially simple modulation scheme, there are a number of interesting questions, such as how to choose the elementary waveform $p(t)$ so as to satisfy frequency constraints and reliably detect the binary digits from the received waveform in the presence of noise and intersymbol interference.

Chapter 6 develops the principles of modulation and demodulation. The simple 2-PAM scheme is generalized in many ways. For example, multilevel modulation first segments the incoming bits into m-tuples. There are $M=2^m$ distinct m-tuples, and in M-PAM, each m-tuple is mapped into a different numerical value (such as $\pm 1, \pm 3, \pm 5, \pm 7$ for $M=8$). The sequence $u_1, u_2, \dots$ of these values is then mapped into the waveform $\sum_n u_n p(t-mn/R)$. Note that the rate at which pulses are sent is now m times smaller than before, but there are 2^m different values to be distinguished at the receiver for each elementary pulse.

The modulated waveform can also be a complex baseband waveform (which is then modulated up to an appropriate passband as a real waveform). In a scheme called quadrature amplitude modulation (QAM), the bit sequence is again segmented into m-tuples, but now there is a mapping from binary m-tuples to a set of $M=2^m$ complex numbers. The sequence $u_1, u_2, \dots$ of outputs from this mapping is then converted to the complex waveform $\sum_n u_n p(t-mn/R)$.

Finally, instead of using a fixed signal pulse $p(t)$ multiplied by a selection from M real or complex values, it is possible to choose M different signal pulses, $p_1(t), \dots, p_M(t)$. This includes frequency shift keying, pulse position modulation, phase modulation, and a host of other strategies.

It is easy to think of many ways to map a sequence of binary digits into a waveform. We shall find that there is a simple geometric "signal-space" approach, based on the results of Chapter 4, for looking at these various combinations in an integrated way.

Because of the noise on the channel, the received waveform is different from the transmitted waveform. A major function of the demodulator is that of detection.

The detector attempts to choose which possible input sequence is most likely to have given rise to the given received waveform. Chapter 7 develops the background in random processes necessary to understand this problem, and Chapter 8 uses the geometric signal-space approach to analyze and understand the detection problem.

1.3.2 Error correction

Frequently the error probability incurred with simple modulation and demodulation techniques is too high. One possible solution is to separate the channel encoder into two layers: first an error-correcting code, then a simple modulator.

As a very simple example, the bit rate into the channel encoder could be reduced by a factor of three, and then each binary input could be repeated three times before entering the modulator. If at most one of the three binary digits coming out of the demodulator were incorrect, it could be corrected by majority rule at the decoder, thus reducing the error probability of the system at a considerable cost in data rate.

The scheme above (repetition encoding followed by majority-rule decoding) is a very simple example of error-correction coding. Unfortunately, with this scheme, small error probabilities are achieved only at the cost of very small transmission rates.

What Shannon showed was the very unintuitive fact that more sophisticated coding schemes can achieve arbitrarily low error probability at any data rate below a value known as the *channel capacity*. The channel capacity is a function of the probabilistic description of the output conditional on each possible input. Conversely, it is not possible to achieve low error probability at rates above the channel capacity. A brief proof of this *channel coding theorem* is given in Chapter 8, but readers should refer to texts on Information Theory such as Gallager (1968) and Cover and Thomas (2006) for detailed coverage.

The channel capacity for a bandlimited additive white Gaussian noise channel is perhaps the most famous result in information theory. If the input power is limited to P, the bandwidth limited to W, and the noise power per unit bandwidth is N_0, then the capacity (in bits per second) is given by

$$C = \mathrm{W}\log_2\left(1 + \frac{P}{N_0\mathrm{W}}\right).$$

Only in the past few years have channel coding schemes been developed that can closely approach this channel capacity.

Early uses of error-correcting codes were usually part of a two-layer system similar to that above, where a digital error-correcting encoder is followed by a modulator. At the receiver, the waveform is first demodulated into a noisy version of the encoded sequence, and then this noisy version is decoded by the error-correcting decoder. Current practice frequently achieves better performance by combining error correction coding and modulation together in coded modulation schemes. Whether the error correction and traditional modulation are separate layers or combined, the combination is generally referred to as a modulator, and a device that does this modulation on data in one direction and demodulation in the other direction is referred to as a modem.

The subject of error correction has grown over the last 50 years to the point where complex and lengthy textbooks are dedicated to this single topic (see, for example, Lin and Costello (2004) and Forney (2005)). This text provides only an introduction to error-correcting codes.

Chapter 9, the final topic of the text, considers channel encoding and decoding for wireless channels. Considerable attention is paid here to modeling physical wireless media. Wireless channels are subject not only to additive noise, but also random fluctuations in the strength of multiple paths between transmitter and receiver. The interaction of these paths causes fading, and we study how this affects coding, signal selection, modulation, and detection. Wireless communication is also used to discuss issues such as channel measurement, and how these measurements can be used at input and output. Finally, there is a brief case study of CDMA (code division multiple access), which ties together many of the topics in the text.

1.4 Digital interface

The interface between the source coding layer and the channel coding layer is a sequence of bits. However, this simple characterization does not tell the whole story. The major complicating factors are as follows.

- Unequal rates: the rate at which bits leave the source encoder is often not perfectly matched to the rate at which bits enter the channel encoder.
- Errors: source decoders are usually designed to decode an exact replica of the encoded sequence, but the channel decoder makes occasional errors.
- Networks: encoded source outputs are often sent over networks, traveling serially over several channels; each channel in the network typically also carries the output from a number of different source encoders.

The first two factors above appear both in point-to-point communication systems and in networks. They are often treated in an *ad hoc* way in point-to-point systems, whereas they must be treated in a standardized way in networks. The third factor, of course, must also be treated in a standardized way in networks.

The usual approach to these problems in networks is to convert the superficially simple binary interface into multiple layers, as illustrated in Figure 1.6

How the layers in Figure 1.6 operate and work together is a central topic in the study of networks and is treated in detail in network texts such as Bertsekas and Gallager (1992). These topics are not considered in detail here, except for the very brief introduction to follow and a few comments as required later.

1.4.1 Network aspects of the digital interface

The output of the source encoder is usually segmented into packets (and in many cases, such as email and data files, is already segmented in this way). Each of the network layers then adds some overhead to these packets, adding a header in the case of TCP

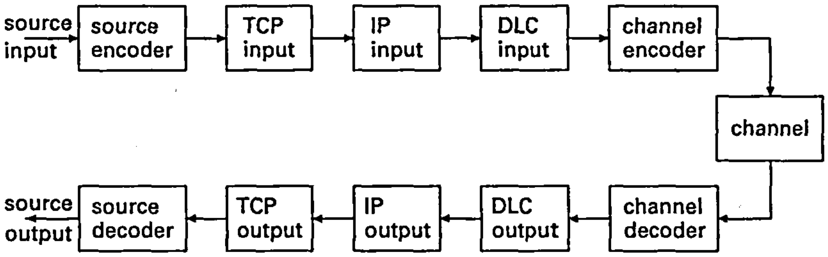


Figure 1.6. The replacement of the binary interface in Figure 1.5 with three layers in an oversimplified view of the internet. There is a TCP (transport control protocol) module associated with each source/destination pair; this is responsible for end-to-end error recovery and for slowing down the source when the network becomes congested. There is an IP (Internet protocol) module associated with each node in the network; these modules work together to route data through the network and to reduce congestion. Finally there is a DLC (data link control) module associated with each channel; this accomplishes rate matching and error recovery on the channel. In network terminology, the channel, with its encoder and decoder, is called the *physical layer.*

(transmission control protocol) and IP (internet protocol) and adding both a header and trailer in the case of DLC (data link control). Thus, what enters the channel encoder is a sequence of frames, where each frame has the structure illustrated in Figure 1.7.

These data frames, interspersed as needed by idle-fill, are strung together, and the resulting bit stream enters the channel encoder at its synchronous bit rate. The header and trailer supplied by the DLC must contain the information needed for the receiving DLC to parse the received bit stream into frames and eliminate the idle-fill.

The DLC also provides protection against decoding errors made by the channel decoder. Typically this is done by using a set of 16 or 32 parity checks in the frame trailer. Each parity check specifies whether a given subset of bits in the frame contains an even or odd number of 1s. Thus if errors occur in transmission, it is highly likely that at least one of these parity checks will fail in the receiving DLC. This type of DLC is used on channels that permit transmission in both directions. Thus, when an erroneous frame is detected, it is rejected and a frame in the opposite direction requests a retransmission of the erroneous frame. Thus the DLC header must contain information about frames traveling in both directions. For details about such protocols, see, for example, Bertsekas and Gallager (1992).

An obvious question at this point is why error correction is typically done both at the physical layer and at the DLC layer. Also, why is feedback (i.e. error detection and retransmission) used at the DLC layer and not at the physical layer? A partial answer is that, if the error correction is omitted at one of the layers, the error probability is increased. At the same time, combining both procedures (with the same

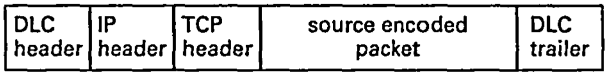


Figure 1.7. Structure of a data frame using the layers of Figure 1.6.

overall overhead) and using feedback at the physical layer can result in much smaller error probabilities. The two-layer approach is typically used in practice because of standardization issues, but, in very difficult communication situations, the combined approach can be preferable. From a tutorial standpoint, however, it is preferable to acquire a good understanding of channel encoding and decoding using transmission in only one direction before considering the added complications of feedback.

When the receiving DLC accepts a frame, it strips off the DLC header and trailer and the resulting packet enters the IP layer. In the IP layer, the address in the IP header is inspected to determine whether the packet is at its destination or must be forwarded through another channel. Thus the IP layer handles routing decisions, and also sometimes the decision to drop a packet if the queues at that node are too long.

When the packet finally reaches its destination, the IP layer strips off the IP header and passes the resulting packet with its TCP header to the TCP layer. The TCP module then goes through another error recovery phase,[4] much like that in the DLC module, and passes the accepted packets, without the TCP header, on to the destination decoder. The TCP and IP layers are also jointly responsible for congestion control, which ultimately requires the ability either to reduce the rate from sources as required or simply to drop sources that cannot be handled (witness dropped cell-phone calls).

In terms of sources and channels, these extra layers simply provide a sharper understanding of the digital interface between source and channel. That is, source encoding still maps the source output into a sequence of bits, and, from the source viewpoint, all these layers can simply be viewed as a channel to send that bit sequence reliably to the destination.

In a similar way, the input to a channel is a sequence of bits at the channel's synchronous input rate. The output is the same sequence, somewhat delayed and with occasional errors.

Thus both source and channel have digital interfaces, and the fact that these are slightly different because of the layering is, in fact, an advantage. The source encoding can focus solely on minimizing the output bit rate (perhaps with distortion and delay constraints) but can ignore the physical channel or channels to be used in transmission. Similarly the channel encoding can ignore the source and focus solely on maximizing the transmission bit rate (perhaps with delay and error rate constraints).

1.5 Supplementary reading

An excellent text that treats much of the material here with more detailed coverage but less depth is Proakis (2000). Another good general text is Wilson (1996). The classic work that introduced the signal space point of view in digital communication is

[4] Even after all these layered attempts to prevent errors, occasional errors are inevitable. Some are caught by human intervention, many do not make any real difference, and a final few have consequences. C'est la vie. The purpose of communication engineers and network engineers is not to eliminate all errors, which is not possible, but rather to reduce their probability as much as practically possible.

Wozencraft and Jacobs (1965). Good undergraduate treatments are provided in Proakis and Salehi (1994), Haykin (2002), and Pursley (2005).

Readers who lack the necessary background in probability should consult Ross (1994) or Bertsekas and Tsitsiklis (2002). More advanced treatments of probability are given in Ross (1996) and Gallager (1996). Feller (1968, 1971) still remains the classic text on probability for the serious student.

Further material on information theory can be found, for example, in Gallager (1968) and Cover and Thomas (2006). The original work by Shannon (1948) is fascinating and surprisingly accessible.

The field of channel coding and decoding has become an important but specialized part of most communication systems. We introduce coding and decoding in Chapter 8, but a separate treatment is required to develop the subject in depth. At MIT, the text here is used for the first of a two-term sequence, and the second term uses a polished set of notes by D. Forney (2005), available on the web. Alternatively, Lin and Costello (2004) is a good choice among many texts on coding and decoding.

Wireless communication is probably the major research topic in current digital communication work. Chapter 9 provides a substantial introduction to this topic, but a number of texts develop wireless communcation in much greater depth. Tse and Viswanath (2005) and Goldsmith (2005) are recommended, and Viterbi (1995) is a good reference for spread spectrum techniques.

4 Source and channel waveforms

4.1 Introduction

This chapter has a dual objective. The first is to understand *analog data compression*, i.e. the compression of sources such as voice for which the output is an arbitrarily varying real- or complex-valued function of time; we denote such functions as *waveforms*. The second is to begin studying the waveforms that are typically transmitted at the input and received at the output of communication channels. The same set of mathematical tools is required for the understanding and representation of both source and channel waveforms; the development of these results is the central topic of this chapter.

These results about waveforms are standard topics in mathematical courses on analysis, real and complex variables, functional analysis, and linear algebra. They are stated here without the precision or generality of a good mathematics text, but with considerably more precision and interpretation than is found in most engineering texts.

4.1.1 Analog sources

The output of many analog sources (voice is the typical example) can be represented as a waveform,[1] $\{u(t) : \mathbb{R} \to \mathbb{R}\}$ or $\{u(t) : \mathbb{R} \to \mathbb{C}\}$. Often, as with voice, we are interested only in real waveforms, but the simple generalization to complex waveforms is essential for Fourier analysis and for baseband modeling of communication channels. Since a real-valued function can be viewed as a special case of a complex-valued function, the results for complex functions are also useful for real functions.

We observed earlier that more complicated analog sources such as video can be viewed as mappings from $\mathbb{R}^n$ to $\mathbb{R}$, e.g. as mappings from horizontal/vertical position and time to real analog values, but for simplicity we consider only waveform sources here.

[1] The notation $\{u(t) : \mathbb{R} \to \mathbb{R}\}$ refers to a function that maps each real number $t \in \mathbb{R}$ into another real number $u(t) \in \mathbb{R}$. Similarly, $\{u(t) : \mathbb{R} \to \mathbb{C}\}$ maps each real number $t \in \mathbb{R}$ into a complex number $u(t) \in \mathbb{C}$. These functions of time, i.e. these waveforms, are usually viewed as dimensionless, thus allowing us to separate physical scale factors in communication problems from the waveform shape.

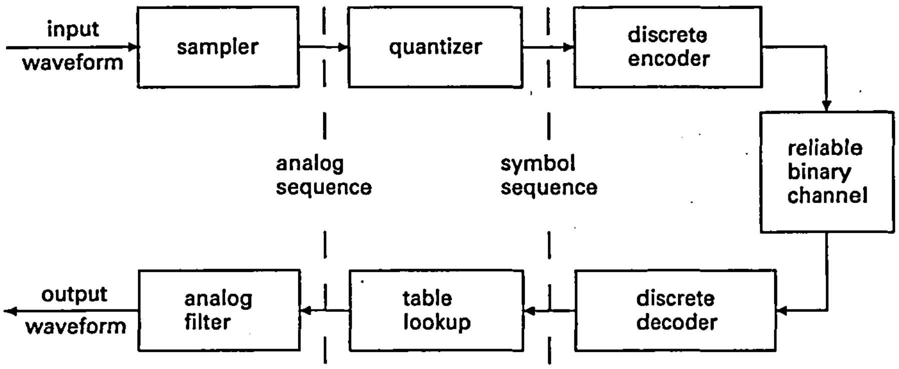


Figure 4.1. Encoding and decoding a waveform source.

We recall in the following why it is desirable to convert analog sources into bits.

- The use of a standard binary interface separates the problem of compressing sources from the problems of channel coding and modulation.
- The outputs from multiple sources can be easily multiplexed together. Multiplexers can work by interleaving bits, 8-bit bytes, or longer packets from different sources.
- When a bit sequence travels serially through multiple links (as in a network), the noisy bit sequence can be cleaned up (regenerated) at each intermediate node, whereas noise tends to accumulate gradually with noisy analog transmission.

A common way of encoding a waveform into a bit sequence is as follows.

(1) Approximate the analog waveform $\{u(t); t \in \mathbb{R}\}$ by its samples[2] $\{u(mT); m \in \mathbb{Z}\}$ at regularly spaced sample times, $\ldots, -T, 0, T, 2T, \ldots$

(2) Quantize each sample (or n-tuple of samples) into a quantization region.

(3) Encode each quantization region (or block of regions) into a string of bits.

These three layers of encoding are illustrated in Figure 4.1, with the three corresponding layers of decoding.

Example 4.1.1 In standard telephony, the voice is filtered to 4000 Hz (4 kHz) and then sampled[3] at 8000 samples/s. Each sample is then quantized to one of 256 possible levels, represented by 8 bits. Thus the voice signal is represented as a 64 kbps sequence. (Modern digital wireless systems use more sophisticated voice coding schemes that reduce the data rate to about 8 kbps with little loss of voice quality.)

The sampling above may be generalized in a variety of ways for converting waveforms into sequences of real or complex numbers. For example, modern voice

[2] $\mathbb{Z}$ denotes the set of integers $-\infty < m < \infty$, so $\{u(mT); m \in \mathbb{Z}\}$ denotes the doubly infinite sequence of samples with $-\infty < m < \infty$.

[3] The sampling theorem, to be discussed in Section 4.6, essentially says that if a waveform is baseband-limited to W Hz, then it can be represented perfectly by 2W samples/s. The highest note on a piano is about 4 kHz, which is considerably higher than most voice frequencies.

compression techniques first segment the voice waveform into 20 ms segments and then use the frequency structure of each segment to generate a vector of numbers. The resulting vector can then be quantized and encoded as previously discussed.

An individual waveform from an analog source should be viewed as a sample waveform from a *random process*. The resulting probabilistic structure on these sample waveforms then determines a probability assignment on the sequences representing these sample waveforms. This random characterization will be studied in Chapter 7; for now, the focus is on ways to map deterministic waveforms to sequences and vice versa. These mappings are crucial both for source coding and channel transmission.

4.1.2 Communication channels

Some examples of communication channels are as follows: a pair of antennas separated by open space; a laser and an optical receiver separated by an optical fiber; a microwave transmitter and receiver separated by a wave guide. For the antenna example, a real waveform at the input in the appropriate frequency band is converted by the input antenna into electromagnetic radiation, part of which is received at the receiving antenna and converted back to a waveform. For many purposes, these physical channels can be viewed as black boxes where the output waveform can be described as a function of the input waveform and noise of various kinds.

Viewing these channels as black boxes is another example of layering. The optical or microwave devices or antennas can be considered as an inner layer around the actual physical channel. This layered view will be adopted here for the most part, since the physics of antennas, optics, and microwaves are largely separable from the digital communication issues developed here. One exception to this is the description of physical channels for wireless communication in Chapter 9. As will be seen, describing a wireless channel as a black box requires some understanding of the underlying physical phenomena.

The function of a channel encoder, i.e. a modulator, is to convert the incoming sequence of binary digits into a waveform in such a way that the noise-corrupted waveform at the receiver can, with high probability, be converted back into the original binary digits. This is typically achieved by first converting the binary sequence into a sequence of analog signals, which are then converted to a waveform. This procession – bit sequence to analog sequence to waveform – is the same procession as performed by a source decoder, and the opposite to that performed by the source encoder. How these functions should be accomplished is very different in the source and channel cases, but both involve converting between waveforms and analog sequences.

The waveforms of interest for channel transmission and reception should be viewed as sample waveforms of random processes (in the same way that source waveforms should be viewed as sample waveforms from a random process). This chapter, however, is concerned only with the relationship between deterministic waveforms and analog sequences; the necessary results about random processes will be postponed until Chapter 7. The reason why so much mathematical precision is necessary here, however, is that these waveforms are a priori unknown. In other words, one cannot use the conventional engineering approach of performing some computation on a function and assuming it is correct if an answer emerges.[4]

4.2 Fourier series

Perhaps the simplest example of an analog sequence that can represent a waveform comes from the Fourier series. The Fourier series is also useful in understanding Fourier transforms and discrete-time Fourier transforms (DTFTs). As will be explained later, our study of these topics will be limited to finite-energy waveforms. Useful models for source and channel waveforms almost invariably fall into the finite-energy class.

The Fourier series represents a waveform, either periodic or time-limited, as a weighted sum of sinusoids. Each weight (coefficient) in the sum is determined by the function, and the function is essentially determined by the sequence of weights. Thus the function and the sequence of weights are essentially equivalent representations.

Our interest here is almost exclusively in time-limited rather than periodic waveforms.[5] Initially the waveforms are assumed to be time-limited to some interval $-T/2 \leq t \leq T/2$ of an arbitrary duration $T > 0$ around 0. This is then generalized to time-limited waveforms centered at some arbitrary time. Finally, an arbitrary waveform is segmented into equal-length segments each of duration T; each such segment is then represented by a Fourier series. This is closely related to modern voice-compression techniques where voice waveforms are segmented into 20 ms intervals, each of which is separately expanded into a Fourier-like series.

Consider a complex function $\{u(t) : \mathbb{R} \to \mathbb{C}\}$ that is nonzero only for $-T/2 \leq t \leq T/2$ (i.e. $u(t) = 0$ for $t < -T/2$ and $t > T/2$). Such a function is frequently indicated by $\{u(t) : [-T/2, T/2] \to \mathbb{C}\}$. The *Fourier series* for such a time-limited function is given by[6]

$$u(t) = \begin{cases} \sum_{k=-\infty}^{\infty} \hat{u}_k \mathrm{e}^{2\pi \mathrm{i} kt/T} & \text{for } -T/2 \leq t \leq T/2; \\ 0 & \text{elsewhere,} \end{cases} \tag{4.1}$$

where i denotes[7] $\sqrt{-1}$. The Fourier series coefficients $\hat{u}_k$ are, in general, complex (even if $u(t)$ is real), and are given by

$$\hat{u}_k = \frac{1}{T} \int_{-T/2}^{T/2} u(t) \mathrm{e}^{-2\pi \mathrm{i} kt/T} \, \mathrm{d}t, \qquad -\infty < k < \infty. \tag{4.2}$$

[4] This is not to disparage the use of computational (either hand or computer) techniques to get a quick answer without worrying about fine points. Such techniques often provide insight and understanding, and the fine points can be addressed later. For a random process, however, one does not know a priori which sample functions can provide computational insight.

[5] Periodic waveforms are not very interesting for carrying information; after one period, the rest of the waveform carries nothing new.

[6] The conditions and the sense in which (4.1) holds are discussed later.

[7] The use of i for $\sqrt{-1}$ is standard in all scientific fields except electrical engineering. Electrical engineers formerly reserved the symbol i for electrical current and thus often use j to denote $\sqrt{-1}$.

The standard rectangular function,

$$\operatorname{rect}(t) = \begin{cases} 1 & \text{for } -1/2 \le t \le 1/2; \\ 0 & \text{elsewhere,} \end{cases}$$

can be used to simplify (4.1) as follows:

$$u(t) = \sum_{k=-\infty}^{\infty} \hat{u}_k \mathrm{e}^{2\pi i k t/T} \operatorname{rect}\left(\frac{t}{T}\right). \tag{4.3}$$

This expresses $u(t)$ as a linear combination of truncated complex sinusoids,

$$u(t) = \sum_{k\in\mathbb{Z}} \hat{u}_k \theta_k(t), \qquad \text{where} \quad \theta_k(t) = \mathrm{e}^{2\pi i k t/T} \operatorname{rect}\left(\frac{t}{T}\right). \tag{4.4}$$

Assuming that (4.4) holds for some set of coefficients $\{\hat{u}_k;\ k \in \mathbb{Z}\}$, the following simple and instructive argument shows why (4.2) is satisfied for that set of coefficients. Two complex waveforms, $\theta_k(t)$ and $\theta_m(t)$, are defined to be *orthogonal* if $\int_{-\infty}^{\infty} \theta_k(t)\theta_m^*(t)\,\mathrm{d}t = 0$. The truncated complex sinusoids in (4.4) are orthogonal since the interval $[-T/2,\ T/2]$ contains an integral number of cycles of each, i.e., for $k \ne m \in \mathbb{Z}$,

$$\int_{-\infty}^{\infty} \theta_k(t)\theta_m^*(t)\,\mathrm{d}t = \int_{-T/2}^{T/2} \mathrm{e}^{2\pi i (k-m)t/T}\,\mathrm{d}t = 0.$$

Thus, the right side of (4.2) can be evaluated as follows:

$$\begin{aligned} \frac{1}{T}\int_{-T/2}^{T/2} u(t)\mathrm{e}^{-2\pi i k t/T}\mathrm{d}t &= \frac{1}{T}\int_{-\infty}^{\infty} \sum_{m=-\infty}^{\infty} \hat{u}_m \theta_m(t)\theta_k^*(t)\,\mathrm{d}t \\ &= \frac{\hat{u}_k}{T}\int_{-\infty}^{\infty} |\theta_k(t)|^2\,\mathrm{d}t \\ &= \frac{\hat{u}_k}{T}\int_{-T/2}^{T/2} \mathrm{d}t = \hat{u}_k. \end{aligned} \tag{4.5}$$

An expansion such as that of (4.4) is called an *orthogonal expansion*. As shown later, the argument in (4.5) can be used to find the coefficients in any orthogonal expansion. At that point, more care will be taken in exchanging the order of integration and summation above.

Example 4.2.1 This and Example 4.2.2 illustrate why (4.4) need not be valid for all values of t. Let $u(t) = \operatorname{rect}(2t)$ (see Figure 4.2). Consider representing $u(t)$ by a Fourier series over the interval $-1/2 \le t \le 1/2$. As illustrated, the series can be shown to converge to $u(t)$ at all $t \in [-1/2, 1/2]$, except for the discontinuities at $t = \pm 1/4$. At $t = \pm 1/4$, the series converges to the midpoint of the discontinuity and (4.4) is not valid[8] at those points. Section 4.3 will show how to state (4.4) precisely so as to avoid these convergence issues.

[8] Most engineers, including the author, would say "So what? Who cares what the Fourier series converges to at a discontinuity of the waveform?" Unfortunately, this example is only the tip of an iceberg, especially when time-sampling of waveforms and sample waveforms of random processes are considered.

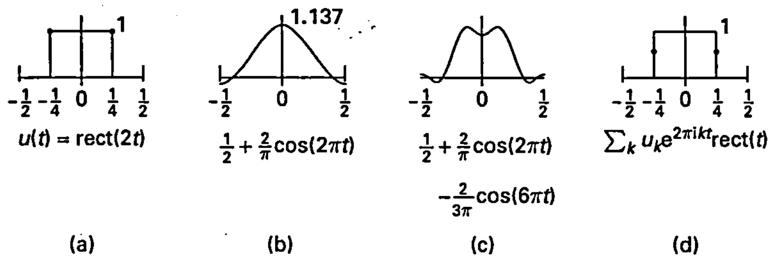


Figure 4.2. The Fourier series (over $[-1/2, 1/2]$) of a rectangular pulse $\mathrm{rect}(2t)$, shown in (a). (b) Partial sum with $k = -1, 0, 1$. (c) Partial sum with $-3 \leq k \leq 3$. Part (d) illustrates that the series converges to $u(t)$ except at the points $t = \pm 1/4$, where it converges to 1/2.

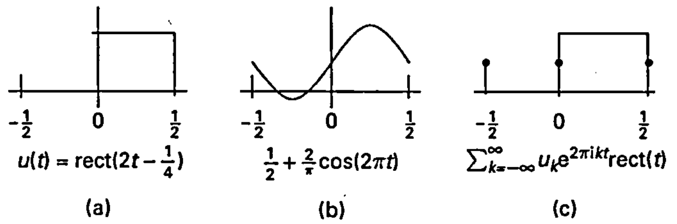


Figure 4.3. The Fourier series over $[-1/2, 1/2]$ of the same rectangular pulse shifted right by 1/4, shown in (a). (b) Partial expansion with $k = -1, 0, 1$. Part (c) depicts that the series converges to $v(t)$ except at the points $t = -1/2, 0$, and $1/2$, at each of which it converges to 1/2.

Example 4.2.2 As a variation of the previous example, let $v(t)$ be 1 for $0 \leq t \leq 1/2$ and 0 elsewhere. Figure 4.3 shows the corresponding Fourier series over the interval $-1/2 \leq t \leq 1/2$.

A peculiar feature of this example is the isolated discontinuity at $t = -1/2$, where the series converges to 1/2. This happens because the untruncated Fourier series, $\sum_{k=-\infty}^{\infty} \hat{v}_k e^{2\pi ikt}$, is periodic with period 1 and thus must have the same value at both $t = -1/2$ and $t = 1/2$. More generally, if an arbitrary function $\{v(t) : [-T/2, T/2] \to \mathbb{C}\}$ has $v(-T/2) \neq v(T/2)$, then its Fourier series over that interval cannot converge to $v(t)$ at both those points.

4.2.1 Finite-energy waveforms

The *energy* in a real or complex waveform $u(t)$ is defined[9] to be $\int_{-\infty}^{\infty} |u(t)|^2 \, dt$. The energy in source waveforms plays a major role in determining how well the waveforms can be compressed for a given level of distortion. As a preliminary explanation, consider the energy in a time-limited waveform $\{u(t) : [-T/2, T/2] \to \mathbb{R}\}$. This energy

[9] Note that $u^2 = |u|^2$ if u is real, but, for complex u, u^2 can be negative or complex and $|u|^2 = uu^* = [\Re(u)]^2 + [\Im(u)]^2$ is required to correspond to the intuitive notion of energy.

is related to the Fourier series coefficients of $u(t)$ by the following *energy equation* which is derived in Exercise 4.2 by the same argument used in (4.5):

$$\int_{t=-T/2}^{T/2} |u(t)|^2 \, \mathrm{d}t = T \sum_{k=-\infty}^{\infty} |\hat{u}_k|^2. \tag{4.6}$$

Suppose that $u(t)$ is compressed by first generating its Fourier series coefficients, $\{\hat{u}_k;\ k \in \mathbb{Z}\}$, and then compressing those coefficients. Let $\{\hat{v}_k;\ k \in \mathbb{Z}\}$ be this sequence of compressed coefficients. Using a squared distortion measure for the coefficients, the overall distortion is $\sum_k |\hat{u}_k - \hat{v}_k|^2$. Suppose these compressed coefficients are now encoded, sent through a channel, reliably decoded, and converted back to a waveform $v(t) = \sum_k \hat{v}_k e^{2\pi i k t/T}$ as in Figure 4.1. The difference between the input waveform $u(t)$ and the output $v(t)$ is then $u(t) - v(t)$, which has the Fourier series $\sum_k (\hat{u}_k - \hat{v}_k) e^{2\pi i k t/T}$. Substituting $u(t) - v(t)$ into (4.6) results in the *difference-energy equation*:

$$\int_{t=-T/2}^{T/2} |u(t) - v(t)|^2 \, \mathrm{d}t = T \sum_k |\hat{u}_k - \hat{v}_k|^2. \tag{4.7}$$

Thus the energy in the difference between $u(t)$ and its reconstruction $v(t)$ is simply T times the sum of the squared differences of the quantized coefficients. This means that reducing the squared difference in the quantization of a coefficient leads directly to reducing the energy in the waveform difference. The energy in the waveform difference is a common and reasonable measure of distortion, but the fact that it is directly related to the mean-squared coefficient distortion provides an important added reason for its widespread use.

There must be at least T units of delay involved in finding the Fourier coefficients for $u(t)$ in $[-T/2, T/2]$ and then reconstituting $v(t)$ from the quantized coefficients at the receiver. There is additional processing and propagation delay in the channel. Thus the output waveform must be a delayed approximation to the input. All of this delay is accounted for by timing recovery processes at the receiver. This timing delay is set so that $v(t)$ at the receiver, according to the receiver timing, is the appropriate approximation to $u(t)$ at the transmitter, according to the transmitter timing. Timing recovery and delay are important problems, but they are largely separable from the problems of current interest. Thus, after recognizing that receiver timing is delayed from transmitter timing, delay can be otherwise ignored for now.

Next, visualize the Fourier coefficients $\hat{u}_k$ as sample values of independent random variables and visualize $u(t)$, as given by (4.3), as a sample value of the corresponding random process (this will be explained carefully in Chapter 7). The expected energy in this random process is equal to T times the sum of the mean-squared values of the coefficients. Similarly the expected energy in the difference between $u(t)$ and $v(t)$ is equal to T times the sum of the mean-squared coefficient distortions. It was seen by scaling in Chapter 3 that the the mean-squared quantization error for an analog random variable is proportional to the variance of that random variable. It is thus not surprising that the expected energy in a random waveform will have a similar relation to the mean-squared distortion after compression.

There is an obvious practical problem with compressing a finite-duration waveform by quantizing an *infinite* set of coefficients. One solution is equally obvious: compress only those coefficients with a significant mean-squared value. Since the expected value of $\sum_k |\hat{u}_k|^2$ is finite for finite-energy functions, the mean-squared distortion from ignoring small coefficients can be made as small as desired by choosing a sufficiently large finite set of coefficients. One then simply chooses $\hat{v}_k = 0$ in (4.7) for each ignored value of k.

The above argument will be explained carefully after developing the required tools. For now, there are two important insights. First, the energy in a source waveform is an important parameter in data compression; second, the source waveforms of interest will have finite energy and can be compressed by compressing a finite number of coefficients.

Next consider the waveforms used for channel transmission. The energy used over any finite interval T is limited both by regulatory agencies and by physical constraints on transmitters and antennas. One could consider waveforms of finite power but infinite duration and energy (such as the lowly sinusoid). On one hand, physical waveforms do not last forever (transmitters wear out or become obsolete), but, on the other hand, *models* of physical waveforms can have infinite duration, modeling physical lifetimes that are much longer than any time scale of communication interest. Nonetheless, for reasons that will gradually unfold, the channel waveforms in this text will almost always be restricted to finite energy.

There is another important reason for concentrating on finite-energy waveforms. Not only are they the appropriate models for source and channel waveforms, but they also have remarkably simple and general properties. These properties rely on an additional constraint called *measurability*, which is explained in Section 4.3. These finite-energy measurable functions are called $\mathcal{L}_2$ functions. When time-constrained, they *always* have Fourier series, and without a time constraint, they *always* have Fourier transforms. Perhaps the most important property, however, is that $\mathcal{L}_2$ functions can be treated essentially as conventional vectors (see Chapter 5).

One might question whether a limitation to finite-energy functions is too constraining. For example, a sinusoid is often used to model the carrier in passband communication, and sinusoids have infinite energy because of their infinite duration. As seen later, however, when a finite-energy baseband waveform is modulated by that sinusoid up to passband, the resulting passband waveform has finite energy.

As another example, the unit impulse (the Dirac delta function $\delta(t)$) is a generalized function used to model waveforms of unit area that are nonzero only in a narrow region around $t = 0$, narrow relative to all other intervals of interest. The impulse response of a linear-time-invariant filter is, of course, the response to a unit impulse; this response approximates the response to a physical waveform that is sufficiently narrow and has unit area. The energy in that physical waveform, however, grows wildly as the waveform narrows. A rectangular pulse of width ε and height $1/\varepsilon$, for example, has unit area for all $\varepsilon > 0$, but has energy $1/\varepsilon$, which approaches ∞ as $\varepsilon \to 0$. One could view the energy in a unit impulse as being either undefined or infinite, but in no way could one view it as being finite.

To summarize, there are many useful waveforms outside the finite-energy class. Although they are not physical waveforms, they are useful models of physical waveforms where energy is not important. Energy is such an important aspect of source and channel waveforms, however, that such waveforms can safely be limited to the finite-energy class.

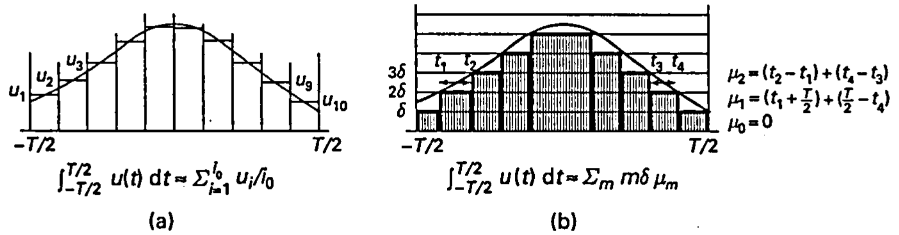
4.3 $\mathcal{L}_2$ functions and Lebesgue integration over $[-T/2, T/2]$

A function $\{u(t) : \mathbb{R} \to \mathbb{C}\}$ is defined to be $\mathcal{L}_2$ if it is Lebesgue measurable and has a finite Lebesgue integral $\int_{-\infty}^{\infty} |u(t)|^2 \, dt$. This section provides a basic and intuitive understanding of what these terms mean. Appendix 4.9 provides proofs of the results, additional examples, and more depth of understanding. Still deeper understanding requires a good mathematics course in real and complex variables. Appendix 4.9 is not required for basic engineering understanding of results in this and subsequent chapters, but it will provide deeper insight.

The basic idea of Lebesgue integration is no more complicated than the more common Riemann integration taught in freshman college courses. Whenever the Riemann integral exists, the Lebesgue integral also exists[10] and has the same value. Thus all the familiar ways of calculating integrals, including tables and numerical procedures, hold without change. The Lebesgue integral is more useful here, partly because it applies to a wider set of functions, but, more importantly, because it greatly simplifies the main results.

This section considers only time-limited functions, $\{u(t) : [-T/2, T/2] \to \mathbb{C}\}$. These are the functions of interest for Fourier series, and the restriction to a finite interval avoids some mathematical details better addressed later.

Figure 4.4 shows intuitively how Lebesgue and Riemann integration differ. Conventional Riemann integration of a nonnegative real-valued function $u(t)$ over an interval $[-T/2, T/2]$ is conceptually performed in Figure 4.4(a) by partitioning $[-T/2, T/2]$ into, say, i_0 intervals each of width T/i_0. The function is then approximated within the


u_1
u_2
u_3
u_9
u_{10}
$-T/2$
$T/2$
$\int_{-T/2}^{T/2} u(t)\,dt \approx \Sigma_{i=1}^{i_0} u_i/i_0$
3δ
2δ
δ
t_1
t_2
t_3
t_4
$\mu_2 = (t_2 - t_1) + (t_4 - t_3)$
$\mu_1 = (t_1 + \frac{T}{2}) + (\frac{T}{2} - t_4)$
$\mu_0 = 0$
$-T/2$
$T/2$
$\int_{-T/2}^{T/2} u(t)\,dt \approx \Sigma_m m\delta\mu_m$

(a)

(b)

Figure 4.4. Example of (a) Riemann and (b) Lebesgue integration.

[10] There is a slight notional qualification to this which is discussed in the sinc function example of Section 4.5.1.

ith such interval by a single value u_i, such as the midpoint of values in the interval. The integral is then approximated as $\sum_{i=1}^{i_0}(T/i_0)u_i$. If the function is sufficiently smooth, then this approximation has a limit, called the Riemann integral, as $i_0 \to \infty$.

To integrate the same function by Lebesgue integration, the vertical axis is partitioned into intervals each of height δ, as shown in Figure 4.4(b). For the mth such interval,[11] $[m\delta, (m+1)\delta)$, let $\mathcal{E}_m$ be the set of values of t such that $m\delta \leq u(t) < (m+1)\delta$. For example, the set $\mathcal{E}_2$ is illustrated by arrows in Figure 4.4(b) and is given by

$$\mathcal{E}_2 = \{t : 2\delta \leq u(t) < 3\delta\} = [t_1, t_2) \cup (t_3, t_4].$$

As explained below, if $\mathcal{E}_m$ is a finite union of separated[12] intervals, its measure, μ_m is the sum of the widths of those intervals; thus μ_2 in the example above is given by

$$\mu_2 = \mu(\mathcal{E}_2) = (t_2 - t_1) + (t_4 - t_3). \tag{4.8}$$

Similarly, $\mathcal{E}_1 = [-T/2, t_1) \cup (t_4, T/2]$ and $\mu_1 = (t_1 + T/2) + (T/2 - t_4)$.

The Lebesque integral is approximated as $\sum_m (m\delta)\mu_m$. This approximation is indicated by the vertically shaded area in Figure 4.4(b). The Lebesgue integral is essentially the limit as $\delta \to 0$.

In short, the Riemann approximation to the area under a curve splits the horizontal axis into uniform segments and sums the corresponding rectangular areas. The Lebesgue approximation splits the vertical axis into uniform segments and sums the height times width measure for each segment. In both cases, a limiting operation is required to find the integral, and Section 4.3.3 gives an example where the limit exists in the Lebesgue but not the Riemann case.

4.3.1 Lebesgue measure for a union of intervals

In order to explain Lebesgue integration further, measure must be defined for a more general class of sets.

The *measure* of an interval I from a to b, $a \leq b$, is defined to be $\mu(I) = b - a \geq 0$. For any finite union of, say, ℓ separated intervals, $\mathcal{E} = \bigcup_{j=1}^{\ell} I_j$, the measure $\mu(\mathcal{E})$ is defined as follows:

$$\mu(\mathcal{E}) = \sum_{j=1}^{\ell} \mu(I_j). \tag{4.9}$$

[11] The notation $[a, b)$ denotes the semiclosed interval $a \leq t < b$. Similarly, $(a, b]$ denotes the semiclosed interval $a < t \leq b$, (a, b) the open interval $a < t < b$, and $[a, b]$ the closed interval $a \leq t \leq b$. In the special case where $a = b$, the interval $[a, a]$ consists of the single point a, whereas $[a, a)$, $(a, a]$, and (a, a) are empty.

[12] Two intervals are *separated* if they are both nonempty and there is at least one point between them that lies in neither interval; i.e., $(0, 1)$ and $(1, 2)$ are separated. In contrast, two sets are *disjoint* if they have no points in common. Thus $(0, 1)$ and $[1, 2]$ are disjoint but not separated.

This definition of $\mu(\mathcal{E})$ was used in (4.8) and is necessary for the approximation in Figure 4.4(b) to correspond to the area under the approximating curve. The fact that the measure of an interval does not depend on the inclusion of the endpoints corresponds to the basic notion of area under a curve. Finally, since these separated intervals are all contained in $[-T/2, T/2]$, it is seen that the sum of their widths is at most T, i.e.

$$0 \le \mu(\mathcal{E}) \le T. \tag{4.10}$$

Any finite union of, say, ℓ arbitrary intervals, $\mathcal{E} = \bigcup_{j=1}^{\ell} I_j$, can also be uniquely expressed as a finite union of at most ℓ separated intervals, say $I'_1, \dots, I'_k$, $k \le \ell$ (see Exercise 4.5), and its measure is then given by

$$\mu(\mathcal{E}) = \sum_{j=1}^{k} \mu(I'_j). \tag{4.11}$$

The union of a countably infinite collection[13] of separated intervals, say $\mathcal{B} = \bigcup_{j=1}^{\infty} I_j$, is also defined to be measurable and has a measure given by

$$\mu(\mathcal{B}) = \lim_{\ell \to \infty} \sum_{j=1}^{\ell} \mu(I_j). \tag{4.12}$$

The summation on the right is bounded between 0 and T for each ℓ. Since $\mu(I_j) \ge 0$, the sum is nondecreasing in ℓ. Thus the limit exists and lies between 0 and T. Also the limit is independent of the ordering of the I_j (see Exercise 4.4).

Example 4.3.1 Let $I_j = (T2^{-2j}, T2^{-2j+1})$ for all integers $j \ge 1$. The jth interval then has measure $\mu(I_j) = 2^{-2j}$. These intervals get smaller and closer to 0 as j increases. They are easily seen to be separated. The union $\mathcal{B} = \bigcup_j I_j$ then has measure $\mu(\mathcal{B}) = \sum_{j=1}^{\infty} T2^{-2j} = T/3$. Visualize replacing the function in Figure 4.4 by one that oscillates faster and faster as $t \to 0$; $\mathcal{B}$ could then represent the set of points on the horizontal axis corresponding to a given vertical slice.

Example 4.3.2 As a variation of Example 4.3.1, suppose $\mathcal{B} = \bigcup_j I_j$, where $I_j = [T2^{-2j}, T2^{-2j}]$ for each j. Then interval I_j consists of the single point $T2^{-2j}$ so $\mu(I_j) = 0$. In this case, $\sum_{j=1}^{\ell} \mu(I_j) = 0$ for each ℓ. The limit of this as $\ell \to \infty$ is also 0, so $\mu(\mathcal{B}) = 0$ in this case. By the same argument, the measure of any countably infinite set of points is 0.

Any countably infinite union of arbitrary (perhaps intersecting) intervals can be uniquely[14] represented as a *countable* (i.e. either a countably infinite or finite) union of separated intervals (see Exercise 4.6); its measure is defined by applying (4.12) to that representation.

[13] An elementary discussion of countability is given in Appendix 4.9.1. Readers unfamiliar with ideas such as the countability of the rational numbers are strongly encouraged to read this appendix.

[14] The collection of separated intervals and the limit in (4.12) is unique, but the ordering of the intervals is not.

4.3.2 Measure for more general sets

It might appear that the class of countable unions of intervals is broad enough to represent any set of interest, but it turns out to be too narrow to allow the general kinds of statements that formed our motivation for discussing Lebesgue integration. One vital generalization is to require that the complement $\overline{\mathcal{B}}$ (relative to $[-T/2, T/2]$) of any measurable set $\mathcal{B}$ also be measurable.[15] Since $\mu([-T/2, T/2]) = T$ and every point of $[-T/2, T/2]$ lies in either $\mathcal{B}$ or $\overline{\mathcal{B}}$ but not both, the measure of $\overline{\mathcal{B}}$ should be $T - \mu(\mathcal{B})$. The reason why this property is necessary in order for the Lebesgue integral to correspond to the area under a curve is illustrated in Figure 4.5.

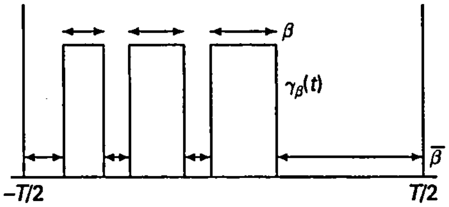


Figure 4.5. Let $f(t)$ have the value 1 on a set $\mathcal{B}$ and the value 0 elsewhere in $[-T/2, T/2]$. Then $\int f(t)\,dt = \mu(\mathcal{B})$. The complement $\overline{\mathcal{B}}$ of $\mathcal{B}$ is also illustrated, and it is seen that $1 - f(t)$ is 1 on the set $\overline{\mathcal{B}}$ and 0 elsewhere. Thus $\int [1 - f(t)]\,dt = \mu(\overline{\mathcal{B}})$, which must equal $T - \mu(\mathcal{B})$ for integration to correspond to the area under a curve.

The *subset inequality* is another property that measure should have: this states that if $\mathcal{A}$ and $\mathcal{B}$ are both measurable and $\mathcal{A} \subseteq \mathcal{B}$, then $\mu(\mathcal{A}) \le \mu(\mathcal{B})$. One can also visualize from Figure 4.5 why this subset inequality is necessary for integration to represent the area under a curve.

Before defining which sets in $[-T/2, T/2]$ are measurable and which are not, a measure-like function called *outer measure* is introduced that exists for *all* sets in $[-T/2, T/2]$. For an arbitrary set $\mathcal{A}$, the set $\mathcal{B}$ is said to *cover* $\mathcal{A}$ if $\mathcal{A} \subseteq \mathcal{B}$ and $\mathcal{B}$ is a countable union of intervals. The outer measure $\mu^{\circ}(\mathcal{A})$ is then essentially the measure of the smallest cover of $\mathcal{A}$. In particular,[16]

$$\mu^{\circ}(\mathcal{A}) = \inf_{\mathcal{B}:\,\mathcal{B}\text{ covers }\mathcal{A}} \mu(\mathcal{B}). \tag{4.13}$$

[15] Appendix 4.9.1 uses the set of rationals in $[-T/2, T/2]$ to illustrate that the complement $\overline{\mathcal{B}}$ of a countable union of intervals $\mathcal{B}$ need not be a countable union of intervals itself. In this case, $\mu(\overline{\mathcal{B}}) = T - \mu(\mathcal{B})$, which is shown to be valid also when $\overline{\mathcal{B}}$ is a countable union of intervals.

[16] The infimum (inf) of a set of real numbers is essentially the minimum of that set. The difference between the minimum and the infimum can be seen in the example of the set of real numbers strictly greater than 1. This set has no minimum, since for each number in the set, there is a smaller number still greater than 1. To avoid this somewhat technical issue, the infimum is defined as the greatest lowerbound of a set. In the example, all numbers less than or equal to 1 are lowerbounds for the set, and 1 is then the greatest lowerbound, i.e. the infimum. Every nonempty set of real numbers has an infinum if one includes $-\infty$ as a choice.

Not surprisingly, the outer measure of a countable union of intervals is equal to its measure as already defined (see Appendix 4.9.3).

Measurable sets and measure over the interval $[-T/2, T/2]$ can now be defined as follows.

Definition 4.3.1 A set $\mathcal{A}$ (over $[-T/2, T/2]$) is *measurable* if $\mu^{\circ}(\mathcal{A}) + \mu^{\circ}(\overline{\mathcal{A}}) = T$. If $\mathcal{A}$ is measurable, then its *measure*, $\mu(\mathcal{A})$, is the outer measure $\mu^{\circ}(\mathcal{A})$.

Intuitively, then, a set is measurable if the set and its complement are sufficiently untangled that each can be covered by countable unions of intervals which have arbitrarily little overlap. The example at the end of Appendix 4.9.4 constructs the simplest nonmeasurable set we are aware of; it should be noted how bizarre it is and how tangled it is with its complement.

The definition of measurability is a "mathematician's definition" in the sense that it is very succinct and elegant, but it does not provide many immediate clues about determining whether a set is measurable and, if so, what its measure is. This is now briefly discussd.

It is shown in Appendix 4.9.3 that countable unions of intervals are measurable according to this definition, and the measure can be found by breaking the set into separated intervals. Also, by definition, the complement of every measurable set is also measurable, so the complements of countable unions of intervals are measurable. Next, if $\mathcal{A} \subseteq \mathcal{A}'$, then any cover of $\mathcal{A}'$ also covers $\mathcal{A}$, so the subset inequality is satisfied. This often makes it possible to find the measure of a set by using a limiting process on a sequence of measurable sets contained in or containing a set of interest. Finally, the following theorem is proven in Appendix 4.9.4.

Theorem 4.3.1 *Let $\mathcal{A}_1, \mathcal{A}_2, \ldots$ be any sequence of measurable sets. Then $\mathcal{S} = \bigcup_{j=1}^{\infty} \mathcal{A}_j$ and $\mathcal{D} = \bigcap_{j=1}^{\infty} \mathcal{A}_j$ are measurable. If $\mathcal{A}_1, \mathcal{A}_2, \ldots$ are also disjoint, then $\mu(\mathcal{S}) = \sum_j \mu(\mathcal{A}_j)$. If $\mu^{\circ}(\mathcal{A}) = 0$, then $\mathcal{A}$ is measurable and has zero measure.*

This theorem and definition say that the collection of measurable sets is closed under countable unions, countable intersections, and complementation. This partly explains why it is so hard to find nonmeasurable sets and also why their existence can usually be ignored – they simply do not arise in the ordinary process of analysis.

Another consequence concerns sets of zero measure. It was shown earlier that any set containing only countably many points has zero measure, but there are many other sets of zero measure. The Cantor set example in Appendix 4.9.4 illustrates a set of zero measure with uncountably many elements. The theorem implies that a set $\mathcal{A}$ has zero measure if, for any $\varepsilon > 0$, $\mathcal{A}$ has a cover $\mathcal{B}$ such that $\mu(\mathcal{B}) \leq \varepsilon$. The definition of measurability shows that the complement of any set of zero measure has measure T, i.e. $[-T/2, T/2]$ is the cover of smallest measure. It will be seen shortly that, for most purposes, including integration, sets of zero measure can be ignored and sets of measure T can be viewed as the entire interval $[-T/2, T/2]$.

This concludes our study of measurable sets on $[-T/2, T/2]$. The bottom line is that not all sets are measurable, but that nonmeasurable sets arise only from bizarre and artificial constructions and can usually be ignored. The definitions of measure and measurability might appear somewhat arbitrary, but in fact they arise simply through the natural requirement that intervals and countable unions of intervals be measurable with the given measure[17] and that the subset inequality and complement property be satisfied. If we wanted additional sets to be measurable, then at least one of the above properties would have to be sacrificed and integration itself would become bizarre. The major result here, beyond basic familiarity and intuition, is Theorem 4.3.1, which is used repeatedly in the following sections. Appendix 4.9 fills in many important details and proves the results here

4.3.3 Measurable functions and integration over $[-T/2, T/2]$

A function $\{u(t) : [-T/2, T/2] \to \mathbb{R}\}$ is said to be *Lebesgue measurable* (or more briefly *measurable*) if the set of points $\{t : u(t) < \beta\}$ is measurable for each $\beta \in \mathbb{R}$. If $u(t)$ is measurable, then, as shown in Exercise 4.11, the sets $\{t : u(t) \leq \beta\}$, $\{t : u(t) \geq \beta\}$, $\{t : u(t) > \beta\}$, and $\{t : \alpha \leq u(t) < \beta\}$ are measurable for all $\alpha < \beta \in \mathbb{R}$. Thus, if a function is measurable, the measure $\mu_m = \mu(\{t : m\delta \leq u(t) < (m+1)\delta\})$ associated with the mth horizontal slice in Figure 4.4 must exist for each $\delta > 0$ and m.

For the Lebesgue integral to exist, it is also necessary that the Figure 4.4 approximation to the Lebesgue integral has a limit as the vertical interval size δ goes to 0. Initially consider only nonnegative functions, $u(t) \geq 0$ for all t. For each integer $n \geq 1$, define the nth-order approximation to the Lebesgue integral as that arising from partitioning the vertical axis into intervals each of height $\delta_n = 2^{-n}$. Thus a unit increase in n corresponds to halving the vertical interval size as illustrated in Figure 4.6.

Let $\mu_{m,n}$ be the measure of $\{t : m2^{-n} \leq u(t) < (m+1)2^{-n}\}$, i.e. the measure of the set of $t \in [-T/2, T/2]$ for which $u(t)$ is in the mth vertical interval for the nth-order

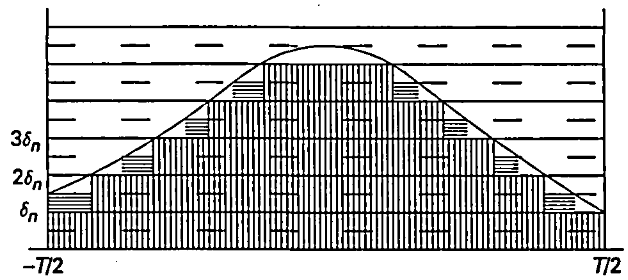


Figure 4.6. Improvement in the approximation to the Lebesgue integral by a unit increase in n is indicated by the horizontal crosshatching.

[17] We have not distinguished between the condition of being measurable and the actual measure assigned a set, which is natural for ordinary integration. The theory can be trivially generalized, however, to random variables restricted to $[-T/2, T/2]$. In this case, the measure of an interval is redefined to be the probability of that interval. Everything else remains the same except that some individual points might have nonzero probability.

approximation. The approximation $\sum_m m2^{-n}\mu_{m,n}$ might be infinite[18] for all n, and in this case the Lebesgue integral is said to be infinite. If the sum is finite for $n = 1$, however, Figure 4.6 shows that the change in going from the approximation of order n to $n+1$ is nonnegative and upperbounded by $T2^{-n-1}$. Thus it is clear that the sequence of approximations has a finite limit which is defined[19] to be the *Lebesgue integral* of $u(t)$. In summary, the Lebesgue integral of an arbitrary measurable nonnegative function $\{u(t) : [-T/2,\ T/2] \to \mathbb{R}\}$ is finite if any approximation is finite and is then given by

$$\int u(t)\,\mathrm{d}t = \lim_{n\to\infty} \sum_{m=0}^{\infty} m2^{-n}\mu_{m,n}, \qquad \text{where} \quad \mu_{m,n} = \mu(t : m2^{-n} \le u(t) < (m+1)2^{-n}). \tag{4.14}$$

Example 4.3.3 Consider a function that has the value 1 for each rational number in $[-T/2, T/2]$ and 0 for all irrational numbers. The set of rationals has zero measure, as shown in Appendix 4.9.1, so that each approximation is zero in Figure 4.6, and thus the Lebesgue integral, as the limit of these approximations, is zero. This is a simple example of a function that has a Lebesgue integral but no Riemann integral.

Next consider two nonnegative measurable functions $u(t)$ and $v(t)$ on $[-T/2,\ T/2]$ and assume $u(t) = v(t)$ except on a set of zero measure. Then each of the approximations in (4.14) are identical for $u(t)$ and $v(t)$, and thus the two integrals are identical (either both infinite or both the same number). This same property will be seen to carry over for functions that also take on negative values and, more generally, for complex-valued functions. This property says that sets of zero measure can be ignored in integration. This is one of the major simplifications afforded by Lebesgue integration. Two functions that are the same except on a set of zero measure are said to be equal *almost everywhere*, abbreviated a.e. For example, the rectangular pulse and its Fourier series representation illustrated in Figure 4.2 are equal a.e.

For functions taking on both positive and negative values, the function $u(t)$ can be separated into a positive part $u^+(t)$ and a negative part $u^-(t)$. These are defined by

$$u^+(t) = \begin{cases} u(t) & \text{for } t : u(t) \ge 0 \\ 0 & \text{for } t : u(t) < 0 \end{cases}; \qquad u^-(t) = \begin{cases} 0 & \text{for } t : u(t) \ge 0 \\ -u(t) & \text{for } t : u(t) < 0. \end{cases}$$

For all $t \in [-T/2,\ T/2]$ then,

$$u(t) = u^+(t) - u^-(t). \tag{4.15}$$

[18] For example, this sum is infinite if $u(t) = 1/|t|$ for $-T/2 \le t \le T/2$. The situation here is essentially the same for Riemann and Lebesgue integration.

[19] This limiting operation can be shown to be independent of how the quantization intervals approach 0.

If $u(t)$ is measurable, then $u^+(t)$ and $u^-(t)$ are also.[20] Since these are nonnegative, they can be integrated as before, and each integral exists with either a finite or infinite value. If at most one of these integrals is infinite, the Lebesgue integral of $u(t)$ is defined as

$$\int u(t) = \int u^+(t) - \int u^-(t)\mathrm{d}t. \tag{4.16}$$

If both $\int u^+(t)\,\mathrm{d}t$ and $\int u^-(t)\,\mathrm{d}t$ are infinite, then the integral is undefined.

Finally, a complex function $\{u(t) : [-T/2\, T/2] \to \mathbb{C}\}$ is defined to be *measurable* if the real and imaginary parts of $u(t)$ are measurable. If the integrals of $\Re(u(t))$ and $\Im(u(t))$ are defined, then the Lebesgue integral $\int u(t)\,\mathrm{d}t$ is defined by

$$\int u(t)\mathrm{d}t = \int \Re(u(t))\mathrm{d}t + \mathrm{i} \int \Im(u(t))\mathrm{d}t. \tag{4.17}$$

The integral is undefined otherwise. Note that this implies that any integration property of complex-valued functions $\{u(t) : [-T/2, T/2] \to \mathbb{C}\}$ is also shared by real-valued functions $\{u(t) : [-T/2, T/2] \to \mathbb{R}\}$.

4.3.4 Measurability of functions defined by other functions

The definitions of measurable functions and Lebesgue integration in Section 4.3.3 were quite simple given the concept of measure. However, functions are often defined in terms of other more elementary functions, so the question arises whether measurability of those elementary functions implies that of the defined function. The bottom-line answer is almost invariably yes. For this reason it is often assumed in the following sections that all functions of interest are measurable. Several results are now given fortifying this bottom-line view.

First, if $\{u(t) : [-T/2, T/2] \to \mathbb{R}\}$ is measurable, then $-u(t)$, $|u(t)|$, $u^2(t)$, $\mathrm{e}^{u(t)}$, and $\ln|u(t)|$ are also measurable. These and similar results follow immediately from the definition of measurable functions and are derived in Exercise 4.12.

Next, if $u(t)$ and $v(t)$ are measurable, then $u(t)+v(t)$ and $u(t)v(t)$ are measurable (see Exercise 4.13).

Finally, if $\{u_k(t) : [-T/2, T/2] \to \mathbb{R}\}$ is a measurable function for each integer $k \geq 1$, then $\inf_k u_k(t)$ is measurable. This can be seen by noting that $\{t : \inf_k[u_k(t)] \leq \alpha\} = \bigcup_k \{t : u_k(t) \leq \alpha\}$, which is measurable for each α. Using this result, Exercise 4.15 shows that $\lim_k u_k(t)$ is measurable if the limit exists for all $t \in [-T/2, T/2]$.

4.3.5 $\mathcal{L}_1$ and $\mathcal{L}_2$ functions over $[-T/2, T/2]$

A function $\{u(t) : [-T/2, T/2] \to \mathbb{C}\}$ is said to be $\mathcal{L}_1$, or in the class $\mathcal{L}_1$, if $u(t)$ is measurable and the Lebesgue integral of $|u(t)|$ is finite.[21]

[20] To see this, note that for $\beta > 0$, $\{t : u^+(t) < \beta\} = \{t : u(t) < \beta\}$. For $\beta \leq 0$, $\{t : u^+(t) < \beta\}$ is the empty set. A similar argument works for $u^-(t)$.

[21] $\mathcal{L}_1$ functions are sometimes called integrable functions.

For the special case of a real function, $\{u(t) : [-T/2, T/2] \to \mathbb{R}\}$, the magnitude $|u(t)|$ can be expressed in terms of the positive and negative parts of $u(t)$ as $|u(t)| = u^+(t) + u^-(t)$. Thus $u(t)$ is $\mathcal{L}_1$ if and only if both $u^+(t)$ and $u^-(t)$ have finite integrals. In other words, $u(t)$ is $\mathcal{L}_1$ if and only if the Lebesgue integral of $u(t)$ is defined and finite.

For a complex function $\{u(t) : [-T/2, T/2] \to \mathbb{C}\}$, it can be seen that $u(t)$ is $\mathcal{L}_1$ if and only if both $\Re[u(t)]$ and $\Im[u(t)]$ are $\mathcal{L}_1$. Thus $u(t)$ is $\mathcal{L}_1$ if and only if $\int u(t)\mathrm{d}t$ is defined and finite.

A function $\{u(t) : [-T/2, T/2] \to \mathbb{R}\}$ or $\{u(t) : [-T/2, T/2] \to \mathbb{C}\}$ is said to be an $\mathcal{L}_2$ function, or a *finite-energy function*, if $u(t)$ is measurable and the Lebesgue integral of $|u(t)|^2$ is finite. All source and channel waveforms discussed in this text will be assumed to be $\mathcal{L}_2$. Although $\mathcal{L}_2$ functions are of primary interest here, the class of $\mathcal{L}_1$ functions is of almost equal importance in understanding Fourier series and Fourier transforms. An important relation between $\mathcal{L}_1$ and $\mathcal{L}_2$ is given in the following simple theorem, illustrated in Figure 4.7.

Theorem 4.3.2 *If* $\{u(t) : [-T/2, T/2] \to \mathbb{C}\}$ *is* $\mathcal{L}_2$*, then it is also* $\mathcal{L}_1$.

Proof: Note that $|u(t)| \le |u(t)|^2$ for all t such that $|u(t)| \ge 1$. Thus $|u(t)| \le |u(t)|^2 + 1$ for all t, so that $\int |u(t)|\,\mathrm{d}t \le \int |u(t)|^2\mathrm{d}t + T$. If the function $u(t)$ is $\mathcal{L}_2$, then the right side of this equation is finite, so the function is also $\mathcal{L}_1$. □

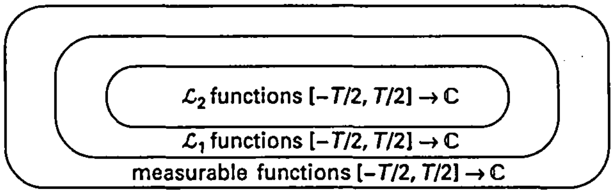

$\mathcal{L}_2$ functions [$-T/2, T/2$] → C
$\mathcal{L}_1$ functions [$-T/2, T/2$] → C
measurable functions [$-T/2, T/2$] → C

Figure 4.7. Illustration showing that for functions from $[-T/2, T/2]$ to $\mathbb{C}$, the class of $\mathcal{L}_2$ functions is contained in the class of $\mathcal{L}_1$ functions, which in turn is contained in the class of measurable functions. The restriction here to a finite domain such as $[-T/2, T/2]$ is necessary, as seen later.

This completes our basic introduction to measure and Lebesgue integration over the finite interval $[-T/2, T/2]$. The fact that the class of measurable sets is closed under complementation, countable unions, and countable intersections underlies the results about the measurability of functions being preserved over countable limits and sums. These in turn underlie the basic results about Fourier series, Fourier integrals, and orthogonal expansions. Some of those results will be stated without proof, but an understanding of measurability will enable us to understand what those results mean. Finally, ignoring sets of zero measure will simplify almost everything involving integration.

4.4 Fourier series for $\mathcal{L}_2$ waveforms

The most important results about Fourier series for $\mathcal{L}_2$ functions are as follows.

Theorem 4.4.1 (Fourier series) *Let $\{u(t) : [-T/2, T/2] \to \mathbb{C}\}$ be an $\mathcal{L}_2$ function. Then for each $k \in \mathbb{Z}$, the Lebesgue integral*

$$\hat{u}_k = \frac{1}{T}\int_{-T/2}^{T/2} u(t)\mathrm{e}^{-2\pi ikt/T}\,\mathrm{d}t \tag{4.18}$$

exists and satisfies $|\hat{u}_k| \le 1/T \int |u(t)|\,\mathrm{d}t < \infty$. Furthermore,

$$\lim_{\ell\to\infty}\int_{-T/2}^{T/2}\left|u(t) - \sum_{k=-\ell}^{\ell}\hat{u}_k\mathrm{e}^{2\pi ikt/T}\right|^2 \mathrm{d}t = 0, \tag{4.19}$$

where the limit is monotonic in ℓ. Also, the energy equation (4.6) is satisfied.

Conversely, if $\{\hat{u}_k; k \in \mathbb{Z}\}$ is a two-sided sequence of complex numbers satisfying $\sum_{k=-\infty}^{\infty}|\hat{u}_k|^2 < \infty$, then an $\mathcal{L}_2$ function $\{u(t) : [-T/2, T/2] \to \mathbb{C}\}$ exists such that (4.6) and (4.19) are satisfied.

The first part of the theorem is simple. Since $u(t)$ is measurable and $\mathrm{e}^{-2\pi ikt/T}$ is measurable for each k, the product $u(t)\mathrm{e}^{-2\pi ikt/T}$ is measurable. Also $|u(t)\mathrm{e}^{-2\pi ikt/T}| = |u(t)|$ so that $u(t)\mathrm{e}^{-2\pi ikt/T}$ is $\mathcal{L}_1$ and the integral exists with the given upperbound (see Exercise 4.17). The rest of the proof is given in, Section 5.3.4.

The integral in (4.19) is the energy in the difference between $u(t)$ and the partial Fourier series using only the terms $-\ell \le k \le \ell$. Thus (4.19) asserts that $u(t)$ can be approximated arbitrarily closely (in terms of difference energy) by finitely many terms in its Fourier series.

A series is defined to *converge in* $\mathcal{L}_2$ if (4.19) holds. The notation l.i.m. (limit in mean-square) is used to denote $\mathcal{L}_2$ convergence, so (4.19) is often abbreviated as follows:

$$u(t) = \text{l.i.m.}\sum_k \hat{u}_k\mathrm{e}^{2\pi ikt/T}\mathrm{rect}\left(\frac{t}{T}\right). \tag{4.20}$$

The notation *does not* indicate that the sum in (4.20) converges pointwise to $u(t)$ at each t; for example, the Fourier series in Figure 4.2 converges to 1/2 rather than 1 at the values $t = \pm 1/4$. In fact, *any* two $\mathcal{L}_2$ functions that are equal a.e. have the same Fourier series coefficients. Thus the best to be hoped for is that $\sum_k \hat{u}_k\mathrm{e}^{2\pi ikt/T}\mathrm{rect}(t/T)$ converges pointwise and yields a "canonical representative" for all the $\mathcal{L}_2$ functions that have the given set of Fourier coefficients, $\{\hat{u}_k; k \in \mathbb{Z}\}$.

Unfortunately, there are some rather bizarre $\mathcal{L}_2$ functions (see the everywhere discontinuous example in Appendix 5.5.1) for which $\sum_k \hat{u}_k\mathrm{e}^{2\pi ikt/T}\mathrm{rect}(t/T)$ diverges for some values of t.

There is an important theorem due to Carleson (1966), however, stating that if $u(t)$ is $\mathcal{L}_2$, then $\sum_k \hat{u}_k\mathrm{e}^{2\pi ikt/T}\mathrm{rect}(t/T)$ converges a.e. on $[-T/2, T/2]$. Thus for any $\mathcal{L}_2$ function $u(t)$, with Fourier coefficients $\{\hat{u}_k : k \in \mathbb{Z}\}$, there is a well defined function,

$$\tilde{u}(t) = \begin{cases} \sum_{k=-\infty}^{\infty}\hat{u}_k\mathrm{e}^{2\pi ikt/T}\mathrm{rect}(t/T) & \text{if the sum converges;} \\ 0 & \text{otherwise.}\end{cases} \tag{4.21}$$

Since the sum above converges a.e., the Fourier coefficients of $\tilde{u}(t)$ given by (4.18) agree with those in (4.21). Thus $\tilde{u}(t)$ can serve as a canonical representative for all the $\mathcal{L}_2$ functions with the same Fourier coefficients $\{\hat{u}_k; k \in \mathbb{Z}\}$. From the difference-energy equation (4.7), it follows that the difference between any two $\mathcal{L}_2$ functions with the same Fourier coefficients has zero energy. Two $\mathcal{L}_2$ functions whose difference has zero energy are said to be *$\mathcal{L}_2$-equivalent*; thus all $\mathcal{L}_2$ functions with the same Fourier coefficients are $\mathcal{L}_2$-equivalent. Exercise 4.18 shows that two $\mathcal{L}_2$ functions are $\mathcal{L}_2$-equivalent if and only if they are equal a.e.

In summary, each $\mathcal{L}_2$ function $\{u(t) : [-T/2, T/2] \to \mathbb{C}\}$ belongs to an equivalence class consisting of all $\mathcal{L}_2$ functions with the same set of Fourier coefficients. Each pair of functions in this equivalence class are $\mathcal{L}_2$-equivalent and equal a.e. The canonical representative in (4.21) is determined solely by the Fourier coefficients and is uniquely defined for any given set of Fourier coefficients satisfying $\sum_k |\hat{u}_k|^2 < \infty$; the corresponding equivalence class consists of the $\mathcal{L}_2$ functions that are equal to $\tilde{u}(t)$ a.e.

From an engineering standpoint, the sequence of ever closer approximations in (4.19) is usually more relevant than the notion of an equivalence class of functions with the same Fourier coefficients. In fact, for physical waveforms, there is no physical test that can distinguish waveforms that are $\mathcal{L}_2$-equivalent, since any such physical test requires an energy difference. At the same time, if functions $\{u(t) : [-T/2, T/2] \to \mathbb{C}\}$ are consistently represented by their Fourier coefficients, then equivalence classes can usually be ignored.

For all but the most bizarre $\mathcal{L}_2$ functions, the Fourier series converges everywhere to some function that is $\mathcal{L}_2$-equivalent to the original function, and thus, as with the points $t = \pm 1/4$ in the example of Figure 4.2, it is usually unimportant how one views the function at those isolated points. Occasionally, however, particularly when discussing sampling and vector spaces, the concept of equivalence classes becomes relevant.

4.4.1 The T-spaced truncated sinusoid expansion

There is nothing special about the choice of 0 as the center point of a time-limited function. For a function $\{v(t) : [\Delta - T/2, \Delta + T/2] \to \mathbb{C}\}$ centered around some arbitrary time Δ, the *shifted Fourier series* over that interval is given by[22]

$$v(t) = \text{l.i.m.} \sum_k \hat{v}_k e^{2\pi i k t/T} \text{rect}\left(\frac{t-\Delta}{T}\right), \tag{4.22}$$

where

$$\hat{v}_k = \frac{1}{T} \int_{\Delta - T/2}^{\Delta + T/2} v(t) e^{-2\pi i k t/T}\, dt, \qquad -\infty < k < \infty. \tag{4.23}$$

[22] Note that the Fourier relationship between the function $v(t)$ and the sequence $\{v_k\}$ depends implicitly on the interval T and the shift Δ.

To see this, let $u(t) = v(t+\Delta)$. Then $u(0) = v(\Delta)$ and $u(t)$ is centered around 0 and has a Fourier series given by (4.20) and (4.18). Letting $\hat{v}_k = \hat{u}_k e^{-2\pi i k\Delta/T}$ yields (4.22) and (4.23). The results about measure and integration are not changed by this shift in the time axis.

Next, suppose that some given function $u(t)$ is either not time-limited or limited to some very large interval. An important method for source coding is first to break such a function into segments, say of duration T, and then to encode each segment[23] separately. A segment can be encoded by expanding it in a Fourier series and then encoding the Fourier series coefficients.

Most voice-compression algorithms use such an approach, usually breaking the voice waveform into 20 ms segments. Voice-compression algorithms typically use the detailed structure of voice rather than simply encoding the Fourier series coefficients, but the frequency structure of voice is certainly important in this process. Thus understanding the Fourier series approach is a good first step in understanding voice compression.

The implementation of voice compression (as well as most signal processing techniques) usually starts with sampling at a much higher rate than the segment duration above. This sampling is followed by high-rate quantization of the samples, which are then processed digitally. Conceptually, however, it is preferable to work directly with the waveform and with expansions such as the Fourier series. The analog parts of the resulting algorithms can then be implemented by the standard techniques of high-rate sampling and digital signal processing.

Suppose that an $\mathcal{L}_2$ waveform $\{u(t) : \mathbb{R} \to \mathbb{C}\}$ is segmented into segments $u_m(t)$ of duration T. Expressing $u(t)$ as the sum of these segments,[24]

$$u(t) = \text{l.i.m.} \sum_m u_m(t), \qquad \text{where } u_m(t) = u(t)\,\text{rect}\Big(\frac{t}{T} - m\Big). \tag{4.24}$$

Expanding each segment $u_m(t)$ by the shifted Fourier series of (4.22) and (4.23) we obtain

$$u_m(t) = \text{l.i.m.} \sum_k \hat{u}_{k,m} e^{2\pi i k t/T}\,\text{rect}\Big(\frac{t}{T} - m\Big), \tag{4.25}$$

where

$$\begin{aligned} \hat{u}_{k,m} &= \frac{1}{T}\int_{mT-T/2}^{mT+T/2} u_m(t) e^{-2\pi i k t/T}\,\mathrm{d}t \\ &= \frac{1}{T}\int_{-\infty}^{\infty} u(t) e^{-2\pi i k t/T}\,\text{rect}\Big(\frac{t}{T} - m\Big)\mathrm{d}t. \end{aligned} \tag{4.26}$$

[23] Any engineer, experienced or not, when asked to analyze a segment of a waveform, will automatically shift the time axis to be centered at 0. The added complication here simply arises from looking at multiple segments together so as to represent the entire waveform.

[24] This sum double-counts the points at the ends of the segments, but this makes no difference in terms of $\mathcal{L}_2$-convergence. Exercise 4.22 treats the convergence in (4.24) and (4.28) more carefully.

Combining (4.24) and (4.25):

$$u(t) = \text{l.i.m.} \sum_m \sum_k \hat{u}_{k,m} e^{2\pi i k t/T} \operatorname{rect}\left(\frac{t}{T} - m\right).$$

This expands $u(t)$ as a weighted sum[25] of the doubly indexed functions:

$$u(t) = \text{l.i.m.} \sum_m \sum_k \hat{u}_{k,m} \theta_{k,m}(t), \qquad \text{where} \quad \theta_{k,m}(t) = e^{2\pi i k t/T} \operatorname{rect}\left(\frac{t}{T} - m\right). \quad (4.27)$$

The functions $\theta_{k,m}(t)$ are orthogonal, since, for $m \neq m'$, the functions $\theta_{k,m}(t)$ and $\theta_{k',m'}(t)$ do not overlap, and, for $m = m'$ and $k \neq k'$, $\theta_{k,m}(t)$ and $\theta_{k',m}(t)$ are orthogonal as before. These functions, $\{\theta_{k,m}(t); k, m \in \mathbb{Z}\}$, are called the *T-spaced truncated sinusoids* and the expansion in (4.27) is called the *T-spaced truncated sinusoid expansion.*

The coefficients $\hat{u}_{k,m}$ are indexed by $k, m \in \mathbb{Z}$ and thus form a countable set.[26] This permits the conversion of an arbitrary $\mathcal{L}_2$ waveform into a countably infinite sequence of complex numbers, in the sense that the numbers can be found from the waveform, and the waveform can be reconstructed from the sequence, at least up to $\mathcal{L}_2$-equivalence.

The l.i.m. notation in (4.27) denotes $\mathcal{L}_2$-convergence; i.e.,

$$\lim_{n,\ell \to \infty} \int_{-\infty}^{\infty} \left| u(t) - \sum_{m=-n}^{n} \sum_{k=-\ell}^{\ell} \hat{u}_{k,m} \theta_{k,m}(t) \right|^2 dt = 0. \quad (4.28)$$

This shows that any given $u(t)$ can be approximated arbitrarily closely by a finite set of coefficients. In particular, each segment can be approximated by a finite set of coefficients, and a finite set of segments approximates the entire waveform (although the required number of segments and coefficients per segment clearly depend on the particular waveform).

For data compression, a waveform $u(t)$ represented by the coefficients $\{\hat{u}_{k,m}; k, m \in \mathbb{Z}\}$ can be compressed by quantizing each $\hat{u}_{k,m}$ into a representative $\hat{v}_{k,m}$. The energy equation (4.6) and the difference-energy equation (4.7) generalize easily to the T-spaced truncated sinusoid expansion as follows:

$$\int_{-\infty}^{\infty} |u(t)|^2 \, dt = T \sum_{m=-\infty}^{\infty} \sum_{k=-\infty}^{\infty} |\hat{u}_{k,m}|^2, \quad (4.29)$$

$$\int_{-\infty}^{\infty} |u(t) - v(t)|^2 \, dt = T \sum_{k=-\infty}^{\infty} \sum_{m=-\infty}^{\infty} |\hat{u}_{k,m} - \hat{v}_{k,m}|^2. \quad (4.30)$$

[25] Exercise 4.21 shows why (4.27) (and similar later expressions) are independent of the order of the limits.

[26] Example 4.9.2 in Section 4.9.1 explains why the doubly indexed set above is countable.

As in Section 4.2.1, a finite set of coefficients should be chosen for compression and the remaining coefficients should be set to 0. The problem of compression (given this expansion) is then to decide how many coefficients to compress, and how many bits to use for each selected coefficient. This of course requires a probabilistic model for the coefficients; this issue is discussed later.

There is a practical problem with the use of T-spaced truncated sinusoids as an expansion to be used in data compression. The boundaries of the segments usually act like step discontinuities (as in Figure 4.3), and this leads to slow convergence over the Fourier coefficients for each segment. These discontinuities could be removed prior to taking a Fourier series, but the current objective is simply to illustrate one general approach for converting arbitrary $\mathcal{L}_2$ waveforms to sequences of numbers. Before considering other expansions, it is important to look at Fourier transforms.

4.5 Fourier transforms and $\mathcal{L}_2$ waveforms

The T-spaced truncated sinusoid expansion corresponds closely to our physical notion of frequency. For example, musical notes correspond to particular frequencies (and their harmonics), but these notes persist for finite durations and then change to notes at other frequencies. However, the parameter T in the T-spaced expansion is arbitrary, and quantizing frequencies in increments of $1/T$ is awkward.

The Fourier transform avoids the need for segmentation into T-spaced intervals, but also removes the capability of looking at frequencies that change in time. It maps a function of time, $\{u(t) : \mathbb{R} \to \mathbb{C}\}$, into a function of frequency,[27] $\{\hat{u}(f) : \mathbb{R} \to \mathbb{C}\}$. The inverse Fourier transform maps $\hat{u}(f)$ back into $u(t)$, essentially making $\hat{u}(f)$ an alternative representation of $u(t)$.

The Fourier transform and its inverse are defined as follows:

$$\hat{u}(f) = \int_{-\infty}^{\infty} u(t) e^{-2\pi i f t}\, dt; \tag{4.31}$$

$$u(t) = \int_{-\infty}^{\infty} \hat{u}(f) e^{2\pi i f t}\, df. \tag{4.32}$$

The time units are seconds and the frequency units hertz (Hz), i.e. cycles per second.

For now we take the conventional engineering viewpoint that any respectable function $u(t)$ has a Fourier transform $\hat{u}(f)$ given by (4.31), and that $u(t)$ can be retrieved from $\hat{u}(f)$ by (4.32). This will shortly be done more carefully for $\mathcal{L}_2$ waveforms.

The following list of equations reviews a few standard Fourier transform relations. In the list, $u(t)$ and $\hat{u}(f)$ denote a Fourier transform pair, written $u(t) \leftrightarrow \hat{u}(f)$, and similarly $v(t) \leftrightarrow \hat{v}(f)$:

[27] The notation $\hat{u}(f)$, rather than the more usual $U(f)$, is used here since capitalization is used to distinguish random variables from sample values. Later, $\{U(t) : \mathbb{R} \to \mathbb{C}\}$ will be used to denote a random process, where, for each t, $U(t)$ is a random variable.

$$au(t)+bv(t) \leftrightarrow a\hat{u}(f)+b\hat{v}(f) \qquad \text{linearity;} \tag{4.33}$$

$$u^*(-t) \leftrightarrow \hat{u}^*(f) \qquad \text{conjugation;} \tag{4.34}$$

$$\hat{u}(t) \leftrightarrow u(-f) \qquad \text{time–frequency duality;} \tag{4.35}$$

$$u(t-\tau) \leftrightarrow \mathrm{e}^{-2\pi i f\tau}\hat{u}(f) \qquad \text{time shift;} \tag{4.36}$$

$$u(t)\,\mathrm{e}^{2\pi i f_0 t} \leftrightarrow \hat{u}(f-f_0) \qquad \text{frequency shift;} \tag{4.37}$$

$$u(t/T) \leftrightarrow T\,\hat{u}(fT) \qquad \text{scaling (for } T>0\text{);} \tag{4.38}$$

$$\mathrm{d}u(t)/\mathrm{d}t \leftrightarrow 2\pi i f\hat{u}(f) \qquad \text{differentiation;} \tag{4.39}$$

$$\int_{-\infty}^{\infty} u(\tau)v(t-\tau)\mathrm{d}\tau \leftrightarrow \hat{u}(f)\hat{v}(f) \qquad \text{convolution;} \tag{4.40}$$

$$\int_{-\infty}^{\infty} u(\tau)v^*(\tau-t)\mathrm{d}\tau \leftrightarrow \hat{u}(f)\hat{v}^*(f) \qquad \text{correlation.} \tag{4.41}$$

These relations will be used extensively in what follows. Time–frequency duality is particularly important, since it permits the translation of results about Fourier transforms to inverse Fourier transforms and vice versa.

Exercise 4.23 reviews the convolution relation (4.40). Equation (4.41) results from conjugating $\hat{v}(f)$ in (4.40).

Two useful special cases of any Fourier transform pair are as follows:

$$u(0) = \int_{-\infty}^{\infty} \hat{u}(f)\mathrm{d}f; \tag{4.42}$$

$$\hat{u}(0) = \int_{-\infty}^{\infty} u(t)\mathrm{d}t. \tag{4.43}$$

These are useful in checking multiplicative constants. Also *Parseval's theorem* results from applying (4.42) to (4.41):

$$\int_{-\infty}^{\infty} u(t)v^*(t)\mathrm{d}t = \int_{-\infty}^{\infty} \hat{u}(f)\hat{v}^*(f)\mathrm{d}f. \tag{4.44}$$

As a corollary, replacing $v(t)$ by $u(t)$ in (4.44) results in the *energy equation* for Fourier transforms, namely

$$\int_{-\infty}^{\infty} |u(t)|^2\,\mathrm{d}t = \int_{-\infty}^{\infty} |\hat{u}(f)|^2\,\mathrm{d}f. \tag{4.45}$$

The magnitude squared of the frequency function, $|\hat{u}(f)|^2$, is called the *spectral density* of $u(t)$. It is the energy per unit frequency (for positive and negative frequencies) in the waveform. The energy equation then says that energy can be calculated by integrating over either time or frequency.

As another corollary of (4.44), note that if $u(t)$ and $v(t)$ are orthogonal, then $\hat{u}(f)$ and $\hat{v}(f)$ are orthogonal, i.e.

$$\int_{-\infty}^{\infty} u(t)v^*(t)\,\mathrm{d}t = 0 \quad \text{if and only if} \quad \int_{-\infty}^{\infty} \hat{u}(f)\hat{v}^*(f)\,\mathrm{d}f = 0. \tag{4.46}$$

The following gives a short set of useful and familiar transform pairs:

$$\mathrm{sinc}(t) = \frac{\sin(\pi t)}{\pi t} \leftrightarrow \mathrm{rect}(f) = \begin{cases} 1 & \text{for} \quad |f| \le 1/2; \\ 0 & \text{for} \quad |f| > 1/2, \end{cases} \tag{4.47}$$

$$\mathrm{e}^{-\pi t^2} \leftrightarrow \mathrm{e}^{-\pi f^2}, \tag{4.48}$$

$$\mathrm{e}^{-at};\ t \ge 0 \leftrightarrow \frac{1}{a + 2\pi \mathrm{i} f} \qquad \text{for} \quad a > 0, \tag{4.49}$$

$$\mathrm{e}^{-a|t|} \leftrightarrow \frac{2a}{a^2 + (2\pi \mathrm{i} f)^2} \qquad \text{for} \quad a > 0. \tag{4.50}$$

Equations (4.47)–(4.50), in conjunction with the Fourier relations (4.33)–(4.41), yield a large set of transform pairs. Much more extensive tables of relations are widely available.

4.5.1 Measure and integration over $\mathbb{R}$

A set $\mathcal{A} \subseteq \mathbb{R}$ is defined to be *measurable* if $\mathcal{A} \cap [-T/2, T/2]$ is measurable for all $T > 0$. The definitions of measurability and measure in Section 4.3.2 were given in terms of an overall interval $[-T/2, T/2]$, but Exercise 4.14 verifies that those definitions are in fact independent of T. That is, if $\mathcal{D} \subseteq [-T/2, T/2]$ is measurable relative to $[-T/2, T/2]$, then $\mathcal{D}$ is measurable relative to $[-T_1/2, T_1/2]$, for each $T_1 > T$, and $\mu(\mathcal{D})$ is the same relative to each of those intervals. Thus measure is defined unambiguously for all sets of bounded duration.

For an arbitrary measurable set $\mathcal{A} \in \mathbb{R}$, the measure of $\mathcal{A}$ is defined to be

$$\mu(\mathcal{A}) = \lim_{T\to\infty} \mu(\mathcal{A} \cap [-T/2, T/2]). \tag{4.51}$$

Since $\mathcal{A} \cap [-T/2, T/2]$ is increasing in T, the subset inequality says that $\mu(\mathcal{A} \cap [-T/2, T/2])$ is also increasing, so the limit in (4.51) must exist as either a finite or infinite value. For example, if $\mathcal{A}$ is taken to be $\mathbb{R}$ itself, then $\mu(\mathbb{R} \cap [-T/2, T/2]) = T$ and $\mu(\mathbb{R}) = \infty$. The possibility for measurable sets to have infinite measure is the primary difference between measure over $[-T/2, T/2]$ and $\mathbb{R}$.[28]

Theorem 4.3.1 carries over without change to sets defined over $\mathbb{R}$. Thus the collection of measurable sets over $\mathbb{R}$ is closed under countable unions and intersections. The measure of a measurable set might be infinite in this case, and if a set has finite measure, then its complement (over $\mathbb{R}$) must have infinite measure.

A real function $\{u(t) : \mathbb{R} \to \mathbb{R}\}$ is *measurable* if the set $\{t : u(t) \le \beta\}$ is measurable for each $\beta \in \mathbb{R}$. Equivalently, $\{u(t) : \mathbb{R} \to \mathbb{R}\}$ is measurable if and only if $u(t)\mathrm{rect}(t/T)$ is measurable for all $T > 0$. A complex function $\{u(t) : \mathbb{R} \to \mathbb{C}\}$ is measurable if the real and imaginary parts of $u(t)$ are measurable.

[28] In fact, it was the restriction to finite measure that permitted the simple definition of measurability in terms of sets and their complements in Section 4.3.2.

If $\{u(t) : \mathbb{R} \to \mathbb{R}\}$ is measurable and nonnegative, there are two approaches to its Lebesgue integral. The first is to use (4.14) directly and the other is to evaluate first the integral over $[-T/2, T/2]$ and then go to the limit $T \to \infty$. Both approaches give the same result.[29]

For measurable real functions $\{u(t) : \mathbb{R} \to \mathbb{R}\}$ that take on both positive and negative values, the same approach as in the finite duration case is successful. That is, let $u^+(t)$ and $u^-(t)$ be the positive and negative parts of $u(t)$, respectively. If at most one of these has an infinite integral, the integral of $u(t)$ is defined and has the value

$$\int u(t)\mathrm{d}t = \int u^+(t)\mathrm{d}t - \int u^-(t)\mathrm{d}t.$$

Finally, a complex function $\{u(t) : \mathbb{R} \to \mathbb{C}\}$ is defined to be *measurable* if the real and imaginary parts of $u(t)$ are measurable. If the integral of $\Re(u(t))$ and that of $\Im(u(t))$ are defined, then

$$\int u(t)\mathrm{d}t = \int \Re(u(t))\mathrm{d}t + \mathrm{i}\int \Im(u(t))\mathrm{d}t. \tag{4.52}$$

A function $\{u(t) : \mathbb{R} \to \mathbb{C}\}$ is said to be in the class $\mathcal{L}_1$ if $u(t)$ is measurable and the Lebesgue integral of $|u(t)|$ is finite. As with integration over a finite interval, an $\mathcal{L}_1$ function has real and imaginary parts that are both $\mathcal{L}_1$. Also the positive and negative parts of those real and imaginary parts have finite integrals.

Example 4.5.1 The sinc function, $\mathrm{sinc}(t) = \sin(\pi t)/\pi t$ is sketched in Figure 4.8 and provides an interesting example of these definitions. Since $\mathrm{sinc}(t)$ approaches 0 with increasing t only as $1/t$, the Riemann integral of $|\mathrm{sinc}(t)|$ is infinite, and with a little thought it can be seen that the Lebesgue integral is also infinite. Thus $\mathrm{sinc}(t)$ is not an $\mathcal{L}_1$ function. In a similar way, $\mathrm{sinc}^+(t)$ and $\mathrm{sinc}^-(t)$ have infinite integrals, and thus the Lebesgue integral of $\mathrm{sinc}(t)$ over $(-\infty, \infty)$ is undefined.

The Riemann integral in this case is said to be improper, but can still be calculated by integrating from $-A$ to $+A$ and then taking the limit $A \to \infty$. The result of this integration is 1, which is most easily found through the Fourier relationship (4.47) combined with (4.43). Thus, in a sense, the sinc function is an example where the Riemann integral exists but the Lebesgue integral does not. In a deeper sense, however,

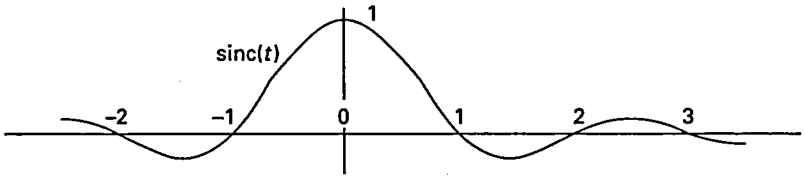


Figure 4.8. The function $\mathrm{sinc}(t)$ goes to 0 as $1/t$ with increasing t.

[29] As explained shortly in the sinc function example, this is not necessarily true for functions taking on positive and negative values.

the issue is simply one of definitions; one can always use Lebesgue integration over $[-A, A]$ and go to the limit $A \to \infty$, getting the same answer as the Riemann integral provides.

A function $\{u(t) : \mathbb{R} \to \mathbb{C}\}$ is said to be in the class $\mathcal{L}_2$ if $u(t)$ is measurable and the Lebesgue integral of $|u(t)|^2$ is finite. All source and channel waveforms will be assumed to be $\mathcal{L}_2$. As pointed out earlier, any $\mathcal{L}_2$ function of finite duration is also $\mathcal{L}_1$. However, $\mathcal{L}_2$ functions of infinite duration need not be $\mathcal{L}_1$; the sinc function is a good example. Since sinc(t) decays as $1/t$, it is not $\mathcal{L}_1$. However, $|\mathrm{sinc}(t)|^2$ decays as $1/t^2$ as $t \to \infty$, so the integral is finite and sinc(t) is an $\mathcal{L}_2$ function.

In summary, measure and integration over $\mathbb{R}$ can be treated in essentially the same way as over $[-T/2, T/2]$. The point sets and functions of interest can be truncated to $[-T/2, T/2]$ with a subsequent passage to the limit $T \to \infty$. As will be seen, however, this requires some care with functions that are not $\mathcal{L}_1$.

4.5.2 Fourier transforms of $\mathcal{L}_2$ functions

The Fourier transform does not exist for all functions, and when the Fourier transform does exist, there is not necessarily an inverse Fourier transform. This section first discusses $\mathcal{L}_1$ functions and then $\mathcal{L}_2$ functions. A major result is that $\mathcal{L}_1$ functions always have well defined Fourier transforms, but the inverse transform does not always have very nice properties; $\mathcal{L}_2$ functions also always have Fourier transforms, but only in the sense of $\mathcal{L}_2$-equivalence. Here however, the inverse transform also exists in the sense of $\mathcal{L}_2$-equivalence. We are primarily interested in $\mathcal{L}_2$ functions, but the results about $\mathcal{L}_1$ functions will help in understanding the $\mathcal{L}_2$ transform.

Lemma 4.5.1 *Let $\{u(t) : \mathbb{R} \to \mathbb{C}\}$ be $\mathcal{L}_1$. Then $\hat{u}(f) = \int_{-\infty}^{\infty} u(t)e^{-2\pi i f t}\,dt$ both exists and satisfies $|\hat{u}(f)| \leq \int |u(t)|\,dt$ for each $f \in \mathbb{R}$. Furthermore, $\{\hat{u}(f) : \mathbb{R} \to \mathbb{C}\}$ is a continuous function of f.*

Proof Note that $|u(t)e^{-2\pi i f t}| = |u(t)|$ for all real t and f. Thus $u(t)e^{-2\pi i f t}$ is $\mathcal{L}_1$ for each f and the integral exists and satisfies the given bound. This is the same as the argument about Fourier series coefficients in Theorem 4.4.1. The continuity follows from asimple ϵ/δ argument (see Exercise 4.24). □

As an example, the function $u(t) = \mathrm{rect}(t)$ is $\mathcal{L}_1$, and its Fourier transform, defined at each f, is the continuous function sinc(f). As discussed before, sinc(f) is not $\mathcal{L}_1$. The inverse transform of sinc(f) exists at all t, equaling rect(t) except at $t = \pm 1/2$, where it has the value 1/2. Lemma 4.5.1 also applies to inverse transforms and verifies that sinc(f) cannot be $\mathcal{L}_1$, since its inverse transform is discontinuous.

Next consider $\mathcal{L}_2$ functions. It will be seen that the pointwise Fourier transform $\int u(t)e^{-2\pi i f t}\,dt$ does not necessarily exist at each f, but that it does exist as an $\mathcal{L}_2$ limit. In exchange for this added complexity, however, the inverse transform exists in exactly the same sense. This result is called Plancherel's theorem and has a nice interpretation in terms of approximations over finite time and frequency intervals.

For any $\mathcal{L}_2$ function $\{u(t) : \mathbb{R} \to \mathbb{C}\}$ and any positive number A, define $\hat{u}_A(f)$ as the Fourier transform of the truncation of $u(t)$ to $[-A, A]$; i.e.,

$$\hat{u}_A(f) = \int_{-A}^{A} u(t) e^{-2\pi i f t} \, dt. \tag{4.53}$$

The function $u(t)\text{rect}(t/2A)$ has finite duration and is thus $\mathcal{L}_1$. It follows that $\hat{u}_A(f)$ is continuous and exists for all f by Lemma 4.5.1. One would normally expect to take the limit in (4.53) as $A \to \infty$ to get the Fourier transform $\hat{u}(f)$, but this limit does not necessarily exist for each f. Plancherel's theorem, however, asserts that this limit exists in the $\mathcal{L}_2$ sense. This theorem is proved in Appendix 5.5.1.

Theorem 4.5.1 (Plancherel, part 1) *For any $\mathcal{L}_2$ function $\{u(t) : \mathbb{R} \to \mathbb{C}\}$, an $\mathcal{L}_2$ function $\{\hat{u}(f) : \mathbb{R} \to \mathbb{C}\}$ exists satisfying both*

$$\lim_{A \to \infty} \int_{-\infty}^{\infty} |\hat{u}(f) - \hat{u}_A(f)|^2 \, df = 0 \tag{4.54}$$

and the energy equation, (4.45).

This not only guarantees the existence of a Fourier transform (up to $\mathcal{L}_2$-equivalence), but also guarantees that it is arbitrarily closely approximated (in difference energy) by the continuous Fourier transforms of the truncated versions of $u(t)$. Intuitively what is happening here is that $\mathcal{L}_2$ functions must have an arbitrarily large fraction of their energy within sufficiently large truncated limits; the part of the function outside of these limits cannot significantly affect the $\mathcal{L}_2$-convergence of the Fourier transform.

The inverse transform is treated very similarly. For any $\mathcal{L}_2$ function $\{\hat{u}(f) : \mathbb{R} \to \mathbb{C}\}$ and any B, $0 < B < \infty$, define

$$u_B(t) = \int_{-B}^{B} \hat{u}(f) e^{2\pi i f t} \, df. \tag{4.55}$$

As before, $u_B(t)$ is a continuous $\mathcal{L}_2$ function for all B, $0<B<\infty$. The final part of Plancherel's theorem is then given by Theorem 4.5.2.

Theorem 4.5.2 (Plancherel, part 2) *For any $\mathcal{L}_2$ function $\{u(t) : \mathbb{R} \to \mathbb{C}\}$, let $\{\hat{u}(f) : \mathbb{R} \to \mathbb{C}\}$ be the Fourier transform of Theorem 4.5.1 and let $u_B(t)$ satisfy (4.55). Then*

$$\lim_{B \to \infty} \int_{-\infty}^{\infty} |u(t) - u_B(t)|^2 \, dt = 0. \tag{4.56}$$

The interpretation is similar to the first part of the theorem. Specifically the inverse transforms of finite frequency truncations of the transform are continuous and converge to an $\mathcal{L}_2$ limit as $B \to \infty$. It also says that this $\mathcal{L}_2$ limit is equivalent to the original function $u(t)$.

Using the limit in mean-square notation, both parts of the Plancherel theorem can be expressed by stating that every $\mathcal{L}_2$ function $u(t)$ has a Fourier transform $\hat{u}(f)$ satisfying

$$\hat{u}(f) = \underset{A \to \infty}{\text{l.i.m.}} \int_{-A}^{A} u(t) e^{-2\pi i f t} \, dt; \qquad u(t) = \underset{B \to \infty}{\text{l.i.m.}} \int_{-B}^{B} \hat{u}(f) e^{2\pi i f t} \, df;$$

i.e., the inverse Fourier transform of $\hat{u}(f)$ is $\mathcal{L}_2$-equivalent to $u(t)$. The first integral above converges pointwise if $u(t)$ is also $\mathcal{L}_1$, and in this case converges pointwise to a continuous function $\hat{u}(f)$. If $u(t)$ is not $\mathcal{L}_1$, then the first integral need not converge pointwise. The second integral behaves in the analogous way.

It may help in understanding the Plancherel theorem to interpret it in terms of finding Fourier transforms using Riemann integration. Riemann integration over an infinite region is defined as a limit over finite regions. Thus, the Riemann version of the Fourier transform is shorthand for

$$\hat{u}(f) = \lim_{A\to\infty} \int_{-A}^{A} u(t)e^{-2\pi i f t}\,dt = \lim_{A\to\infty} \hat{u}_A(f). \tag{4.57}$$

Thus, the Plancherel theorem can be viewed as replacing the Riemann integral with a Lebesgue integral and replacing the pointwise limit (if it exists) in (4.57) with $\mathcal{L}_2$-convergence. The Fourier transform over the finite limits $-A$ to A is continuous and well behaved, so the major difference comes in using $\mathcal{L}_2$-convergence as $A \to \infty$.

As an example of the Plancherel theorem, let $u(t) = \text{rect}(t)$. Then $\hat{u}_A(f) = \text{sinc}(f)$ for all $A \geq 1/2$, so $\hat{u}(f) = \text{sinc}(f)$. For the inverse transform, $u_B(t) = \int_{-B}^{B} \text{sinc}(f)df$ is messy to compute but can be seen to approach rect(t) as $B \to \infty$ except at $t = \pm 1/2$, where it equals 1/2. At $t = \pm 1/2$, the inverse transform is 1/2, whereas $u(t) = 1$.

As another example, consider the function $u(t)$, where $u(t) = 1$ for rational values of $t \in [0, 1]$ and $u(t) = 0$ otherwise. Since this is 0 a.e, the Fourier transform $\hat{u}(f)$ is 0 for all f and the inverse transform is 0, which is $\mathcal{L}_2$-equivalent to $u(t)$. Finally, Example 5.5.1 in Appendix 5.5.1 illustrates a bizarre $\mathcal{L}_1$ function $g(t)$ that is everywhere discontinuous. Its transform $\hat{g}(f)$ is bounded and continuous by Lemma 4.5.1, but is not $\mathcal{L}_1$. The inverse transform is again discontinuous everywhere in $(0, 1)$ and unbounded over every subinterval. This example makes clear why the inverse transform of a continuous function of frequency might be bizarre, thus reinforcing our focus on $\mathcal{L}_2$ functions rather than a more conventional focus on notions such as continuity.

In what follows, $\mathcal{L}_2$-convergence, as in the Plancherel theorem, will be seen as increasingly friendly and natural. Regarding two functions whose difference has zero energy as being the same (formally, as $\mathcal{L}_2$-equivalent) allows us to avoid many trivialities, such as how to define a discontinuous function at its discontinuities. In this case, engineering commonsense and sophisticated mathematics arrive at the same conclusion.

Finally, it can be shown that all the Fourier transform relations in (4.33)–(4.41) except differentiation hold for all $\mathcal{L}_2$ functions (see Exercises 4.26 and 5.15). The derivative of an $\mathcal{L}_2$ function need not be $\mathcal{L}_2$, and need not have a well defined Fourier transform.

4.6 The DTFT and the sampling theorem

The discrete-time Fourier transform (DTFT) is the time–frequency dual of the Fourier series. It will be shown that the DTFT leads immediately to the sampling theorem.

4.6.1 The discrete-time Fourier transform

Let $\hat{u}(f)$ be an $\mathcal{L}_2$ function of frequency, nonzero only for $-\mathrm{W} \le f \le \mathrm{W}$. The DTFT of $\hat{u}(f)$ over $[-\mathrm{W}, \mathrm{W}]$ is then defined by

$$\hat{u}(f) = \mathrm{l.i.m.} \sum_k u_k \mathrm{e}^{-2\pi i k f/2\mathrm{W}} \,\mathrm{rect}\left(\frac{f}{2}\mathrm{W}\right), \tag{4.58}$$

where the DTFT coefficients $\{u_k;\, k \in \mathbb{Z}\}$ are given by

$$u_k = \frac{1}{2\mathrm{W}} \int_{-\mathrm{W}}^{\mathrm{W}} \hat{u}(f) \mathrm{e}^{2\pi i k f/2\mathrm{W}} \,\mathrm{d}f. \tag{4.59}$$

These are the same as the Fourier series equations, replacing t by f, T by 2W, and $\mathrm{e}^{2\pi i \cdots}$ by $\mathrm{e}^{-2\pi i \cdots}$. Note that $\hat{u}(f)$ has an inverse Fourier transform $u(t)$ which is thus baseband-limited to $[-\mathrm{W}, \mathrm{W}]$. As will be shown shortly, the sampling theorem relates the samples of this baseband waveform to the coefficients in (4.59).

The Fourier series theorem (Theorem 4.4.1) clearly applies to (4.58) and (4.59) with the above notational changes; it is repeated here for convenience.

Theorem 4.6.1 (DTFT) *Let $\{\hat{u}(f) : [-\mathrm{W}, \mathrm{W}] \to \mathbb{C}\}$ be an $\mathcal{L}_2$ function. Then for each $k \in \mathbb{Z}$, the Lebesgue integral (4.59) exists and satisfies $|u_k| \le (1/2\mathrm{W}) \int |\hat{u}(f)| \,\mathrm{d}f < \infty$. Furthermore,*

$$\lim_{\ell \to \infty} \int_{-\mathrm{W}}^{\mathrm{W}} \left| \hat{u}(f) - \sum_{k=-\ell}^{\ell} u_k \mathrm{e}^{-2\pi i k f/2\mathrm{W}} \right|^2 \mathrm{d}f = 0 \tag{4.60}$$

and

$$\int_{-\mathrm{W}}^{\mathrm{W}} |\hat{u}(f)|^2 \,\mathrm{d}f = 2\mathrm{W} \sum_{k=-\infty}^{\infty} |u_k|^2. \tag{4.61}$$

Finally, if $\{u_k,\, k \in \mathbb{Z}\}$ is a sequence of complex numbers satisfying $\sum_k |u_k|^2 < \infty$, then an $\mathcal{L}_2$ function $\{\hat{u}(f) : [-\mathrm{W}, \mathrm{W}] \to \mathbb{C}\}$ exists satisfying (4.60) and (4.61).

As before, (4.58) is shorthand for (4.60). Again, this says that any desired approximation accuracy, in terms of energy, can be achieved by using enough terms in the series.

Both the Fourier series and the DTFT provide a one-to-one transformation (in the sense of $\mathcal{L}_2$-convergence) between a function and a sequence of complex numbers. In the case of the Fourier series, one usually starts with a function $u(t)$ and uses the sequence of coefficients to represent the function (up to $\mathcal{L}_2$-equivalence). In the case of the DTFT, one often starts with the sequence and uses the frequency function to represent the sequence. Since the transformation goes both ways, however, one can view the function and the sequence as equally fundamental.

4.6.2 The sampling theorem

The DTFT is next used to establish the sampling theorem, which in turn will help interpret the DTFT. The DTFT (4.58) expresses $\hat{u}(f)$ as a weighted sum of truncated sinusoids in frequency,

$$\hat{u}(f) = \text{l.i.m.} \sum_k u_k \hat{\phi}_k(f), \quad \text{where} \quad \hat{\phi}_k(f) = \mathrm{e}^{-2\pi ikf/2\mathrm{W}} \operatorname{rect}\left(\frac{f}{2\mathrm{W}}\right). \tag{4.62}$$

Ignoring any questions of convergence for the time being, the inverse Fourier transform of $\hat{u}(f)$ is then given by $u(t) = \sum_k u_k \phi_k(t)$, where $\phi_k(t)$ is the inverse transform of $\hat{\phi}_k(f)$. Since the inverse transform[30] of $\operatorname{rect}(f/2\mathrm{W})$ is $2\mathrm{W}\operatorname{sinc}(2\mathrm{W}t)$, the time-shift relation implies that the inverse transform of $\hat{\phi}_k(f)$ is given by

$$\phi_k(t) = 2\mathrm{W}\operatorname{sinc}(2\mathrm{W}t - k) \quad \leftrightarrow \quad \hat{\phi}_k(f) = \mathrm{e}^{-2\pi ikf/2\mathrm{W}} \operatorname{rect}\left(\frac{f}{2\mathrm{W}}\right). \tag{4.63}$$

Thus $u(t)$, the inverse transform of $\hat{u}(f)$, is given by

$$u(t) = \sum_{k=-\infty}^{\infty} u_k \phi_k(t) = \sum_{k=-\infty}^{\infty} 2\mathrm{W}u_k \operatorname{sinc}(2\mathrm{W}t - k). \tag{4.64}$$

Since the set of truncated sinusoids $\{\hat{\phi}_k; k \in \mathbb{Z}\}$ is orthogonal, the sinc functions $\{\phi_k; k \in \mathbb{Z}\}$ are also orthogonal, from (4.46). Figure 4.9 illustrates ϕ_0 and ϕ_1 for the normalized case where $\mathrm{W} = 1/2$.

Note that $\operatorname{sinc}(t)$ equals 1 for $t = 0$ and 0 for all other integer t. Thus if (4.64) is evaluated for $t = k/2\mathrm{W}$, the result is that $u(k/2\mathrm{W}) = 2\mathrm{W}u_k$ for all integer k. Substituting this into (4.64) results in the equation known as the sampling equation:

$$u(t) = \sum_{k=-\infty}^{\infty} u\left(\frac{k}{2\mathrm{W}}\right) \operatorname{sinc}(2\mathrm{W}t - k).$$

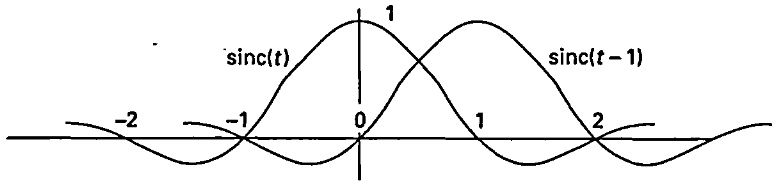


Figure 4.9. Sketch of $\operatorname{sinc}(t) = \sin(\pi t)/\pi t$ and $\operatorname{sinc}(t-1)$. Note that these spaced sinc functions are orthogonal to each other.

[30] This is the time/frequency dual of (4.47). $\hat{u}(f) = \operatorname{rect}(f/2W)$ is both $\mathcal{L}_1$ and $\mathcal{L}_2$; $u(t)$ is continuous and $\mathcal{L}_2$ but not $\mathcal{L}_1$. From the Plancherel theorem, the transform of $u(t)$, in the $\mathcal{L}_2$ sense, is $\hat{u}(f)$.

This says that a baseband-limited function is specified by its samples at intervals $T = 1/2\mathsf{W}$. In terms of this sample interval, the sampling equation is given by

$$u(t) = \sum_{k=-\infty}^{\infty} u(kT)\,\mathrm{sinc}\left(\frac{t}{T} - k\right). \tag{4.65}$$

The following theorem makes this precise. See Appendix 5.5.2 for an insightful proof.

Theorem 4.6.2 (Sampling theorem) *Let $\{u(t) : \mathbb{R} \to \mathbb{C}\}$ be a continuous $\mathcal{L}_2$ function baseband-limited to* W. *Then (4.65) specifies $u(t)$ in terms of its T-spaced samples with $T = 1/2\mathsf{W}$. The sum in (4.65) converges to $u(t)$ for each $t \in \mathbb{R}$, and $u(t)$ is bounded at each t by $|u(t)| \le \int_{-\mathsf{W}}^{\mathsf{W}} |\hat{u}(f)|\,\mathrm{d}f < \infty$.*

The following example illustrates why $u(t)$ is assumed to be continuous above.

Example 4.6.1 (A discontinuous baseband function) Let $u(t)$ be a continuous $\mathcal{L}_2$ baseband function limited to $|f| \le 1/2$. Let $v(t)$ satisfy $v(t) = u(t)$ for all noninteger t and $v(t) = u(t) + 1$ for all integer t. Then $u(t)$ and $v(t)$ are $\mathcal{L}_2$-equivalent, but their samples at each integer time differ by 1. Their Fourier transforms are the same, say $\hat{u}(f)$, since the differences at isolated points have no effect on the transform. Since $\hat{u}(f)$ is nonzero only in $[-\mathsf{W}, \mathsf{W}]$, it is $\mathcal{L}_1$. According to the time–frequency dual of Lemma 4.5.1, the pointwise inverse Fourier transform of $\hat{u}(f)$ is a continuous function, say $u(t)$. Out of all the $\mathcal{L}_2$-equivalent waveforms that have the transform $\hat{u}(f)$, only $u(t)$ is continuous, and it is that $u(t)$ that satisfies the sampling theorem.

The function $v(t)$ is equal to $u(t)$ except for the isolated discontinuities at each integer point. One could view $u(t)$ as baseband-limited also, but this is clearly not physically meaningful and is not the continuous function of the theorem.

Example 4.6.1 illustrates an ambiguity about the meaning of baseband-limited functions. One reasonable definition is that an $\mathcal{L}_2$ function $v(t)$ is baseband-limited to W if $\hat{u}(f)$ is 0 for $|f| > \mathsf{W}$. Another reasonable definition is that $u(t)$ is baseband-limited to W if $u(t)$ is the pointwise inverse Fourier transform of a function $\hat{u}(f)$ that is 0 for $|f| > \mathsf{W}$. For a given $\hat{u}(f)$, there is a unique $u(t)$ according to the second definition and it is continuous; all the functions that are $\mathcal{L}_2$-equivalent to $u(t)$ are bandlimited by the first definition, and all but $u(t)$ are discontinuous and potentially violate the sampling equation. Clearly the second definition is preferable on both engineering and mathematical grounds.

Definition 4.6.1 An $\mathcal{L}_2$ function is *baseband-limited* to W if it is the pointwise inverse transform of an $\mathcal{L}_2$ function $\hat{u}(f)$ that is 0 for $|f| > \mathsf{W}$. Equivalently, it is baseband-limited to W if it is continuous and its Fourier transform is 0 for $|f| > 0$.

The DTFT can now be further interpreted. Any baseband-limited $\mathcal{L}_2$ function $\{\hat{u}(f) : [-\mathsf{W}, \mathsf{W}] \to \mathbb{C}\}$ has both an inverse Fourier transform $u(t) = \int \hat{u}(f) e^{2\pi i f t}\,\mathrm{d}f$ and a DTFT sequence given by (4.58). The coefficients u_k of the DTFT are the scaled samples, $Tu(kT)$, of $u(t)$, where $T = 1/2\mathsf{W}$. Put in a slightly different way, the DTFT in (4.58) is the Fourier transform of the sampling equation (4.65) with $u(kT) = u_k/T$.[31]

It is somewhat surprising that the sampling theorem holds with pointwise convergence, whereas its transform, the DTFT, holds only in the $\mathcal{L}_2$-equivalence sense. The reason is that the function $\hat{u}(f)$ in the DTFT is $\mathcal{L}_1$ but not necessarily continuous, whereas its inverse transform $u(t)$ is necessarily continuous but not necessarily $\mathcal{L}_1$.

The set of functions $\{\hat{\phi}_k(f);\, k \in \mathbb{Z}\}$ in (4.63) is an orthogonal set, since the interval $[-\mathsf{W}, \mathsf{W}]$ contains an integer number of cycles from each sinusoid. Thus, from (4.46), the set of sinc functions in the sampling equation is also orthogonal. Thus both the DTFT and the sampling theorem expansion are orthogonal expansions. It follows (as will be shown carefully later) that the energy equation,

$$\int_{-\infty}^{\infty} |u(t)|^2 \, \mathrm{d}t = T \sum_{k=-\infty}^{\infty} |u(kT)|^2, \tag{4.66}$$

holds for any continuous $\mathcal{L}_2$ function $u(t)$ baseband-limited to $[-\mathsf{W}, \mathsf{W}]$ with $T = 1/2\mathsf{W}$.

In terms of source coding, the sampling theorem says that any $\mathcal{L}_2$ function $u(t)$ that is baseband-limited to W can be sampled at rate 2W (i.e. at intervals $T = 1/2\mathsf{W}$) and the samples can later be used to reconstruct the function perfectly. This is slightly different from the channel coding situation where a sequence of signal values are mapped into a function from which the signals can later be reconstructed. The sampling theorem shows that any $\mathcal{L}_2$ baseband-limited function can be represented by its samples. The following theorem, proved in Appendix 5.5.2, covers the channel coding variation.

Theorem 4.6.3 (Sampling theorem for transmission) *Let $\{a_k;\, k\in\mathbb{Z}\}$ be an arbitrary sequence of complex numbers satisfying $\sum_k |a_k|^2 < \infty$. Then $\sum_k a_k \operatorname{sinc}(2\mathsf{W}t - k)$ converges pointwise to a continuous bounded $\mathcal{L}_2$ function $\{u(t) : \mathbb{R} \to \mathbb{C}\}$ that is baseband-limited to* W *and satisfies $a_k = u(k/2\mathsf{W})$ for each k.*

4.6.3 Source coding using sampled waveforms

Section 4.1 and Figure 4.1 discuss the sampling of an analog waveform $u(t)$ and quantizing the samples as the first two steps in analog source coding. Section 4.2 discusses an alternative in which successive segments $\{u_m(t)\}$ of the source are each expanded in a Fourier series, and then the Fourier series coefficients are quantized. In this latter case, the received segments $\{v_m(t)\}$ are reconstructed from the quantized coefficients. The energy in $u_m(t) - v_m(t)$ is given in (4.7) as a scaled version of the sum of the squared coefficient differences. This section treats the analogous relationship when quantizing the samples of a baseband-limited waveform.

For a continuous function $u(t)$, baseband-limited to W, the samples $\{u(kT);\, k \in \mathbb{Z}\}$ at intervals $T = 1/2\mathsf{W}$ specify the function. If $u(kT)$ is quantized to $v(kT)$ for each k, and

[31] Note that the DTFT is the time-frequency *dual* of the Fourier series but is the *Fourier transform* of the sampling equation.

$u(t)$ is reconstructed as $v(t) = \sum_k v(kT)\,\mathrm{sinc}(t/T - k)$, then, from (4.66), the mean-squared error (MSE) is given by

$$\int_{-\infty}^{\infty} |u(t) - v(t)|^2 \,\mathrm{d}t = T \sum_{k=-\infty}^{\infty} |u(kT) - v(kT)|^2. \tag{4.67}$$

Thus, whatever quantization scheme is used to minimize the MSE between a sequence of samples, that same strategy serves to minimize the MSE between the corresponding waveforms.

The results in Chapter 3 regarding mean-squared distortion for uniform vector quantizers give the distortion at any given bit rate per sample as a linear function of the mean-squared value of the source samples. If any sample has an infinite mean-squared value, then either the quantization rate is infinite or the mean-squared distortion is infinite. This same result then carries over to waveforms. This starts to show why the restriction to $\mathcal{L}_2$ source waveforms is important. It also starts to show why general results about $\mathcal{L}_2$ waveforms are important.

The sampling theorem tells the story for sampling baseband-limited waveforms. However, physical source waveforms are not perfectly limited to some frequency W; rather, their spectra usually drop off rapidly above some nominal frequency W. For example, audio spectra start dropping off well before the nominal cutoff frequency of 4 kHz, but often have small amounts of energy up to 20 kHz. Then the samples at rate 2W do not quite specify the waveform, which leads to an additional source of error, called aliasing. Aliasing will be discussed more fully in Section 4.7.

There is another unfortunate issue with the sampling theorem. The sinc function is nonzero over all noninteger times. Recreating the waveform at the receiver[32] from a set of samples thus requires infinite delay. Practically, of course, sinc functions can be truncated, but the sinc waveform decays to zero as $1/t$, which is impractically slow. Thus the clean result of the sampling theorem is not quite as practical as it first appears.

4.6.4 The sampling theorem for $[\Delta - \mathsf{W}, \Delta + \mathsf{W}]$

Just as the Fourier series generalizes to time intervals centered at some arbitrary time Δ, the DTFT generalizes to frequency intervals centered at some arbitrary frequency Δ.

Consider an $\mathcal{L}_2$ frequency function $\{\hat{v}(f) : [\Delta - \mathsf{W}, \Delta + \mathsf{W}] \to \mathbb{C}\}$. The *shifted DTFT* for $\hat{v}(f)$ is then given by

$$\hat{v}(f) = l.i.m. \sum_k v_k e^{-2\pi i k f/2\mathsf{W}} \,\mathrm{rect}\left(\frac{f-\Delta}{2\mathsf{W}}\right), \tag{4.68}$$

[32] Recall that the receiver time reference is delayed from that at the source by some constant τ. Thus $v(t)$, the receiver estimate of the source waveform $u(t)$ at source time t, is recreated at source time $t+\tau$. With the sampling equation, even if the sinc function is approximated, τ is impractically large.

where

$$v_k = \frac{1}{2\mathrm{W}} \int_{\Delta-\mathrm{W}}^{\Delta+\mathrm{W}} \hat{v}(f) e^{2\pi i k f/2\mathrm{W}}\, df. \tag{4.69}$$

Equation (4.68) is an orthogonal expansion,

$$\hat{v}(f) = l.i.m. \sum_k v_k \hat{\theta}_k(f), \quad \text{where} \quad \hat{\theta}_k(f) = e^{-2\pi i k f/2\mathrm{W}} \mathrm{rect}\left(\frac{f-\Delta}{2\mathrm{W}}\right).$$

The inverse Fourier transform of $\hat{\theta}_k(f)$ can be calculated by shifting and scaling as follows:

$$\theta_k(t) = 2\mathrm{W}\,\mathrm{sinc}(2\mathrm{W}t - k)\, e^{2\pi i \Delta(t-k/2\mathrm{W})} \leftrightarrow \hat{\theta}_k(f) = e^{-2\pi i k f/2\mathrm{W}} \mathrm{rect}\left(\frac{f-\Delta}{2\mathrm{W}}\right). \tag{4.70}$$

Let $v(t)$ be the inverse Fourier transform of $\hat{v}(f)$:

$$v(t) = \sum_k v_k \theta_k(t) = \sum_k 2\mathrm{W} v_k\, \mathrm{sinc}(2\mathrm{W}t - k) e^{2\pi i \Delta(t-k/2\mathrm{W})}.$$

For $t = k/2\mathrm{W}$, only the kth term is nonzero, and $v(k/2\mathrm{W}) = 2\mathrm{W}v_k$. This generalizes the sampling equation to the frequency band $[\Delta - \mathrm{W}, \Delta + \mathrm{W}]$:

$$v(t) = \sum_k v\left(\frac{k}{2\mathrm{W}}\right) \mathrm{sinc}(2\mathrm{W}t - k) e^{2\pi i \Delta(t-k/2\mathrm{W})}.$$

Defining the sampling interval $T = 1/2\mathrm{W}$ as before, this becomes

$$v(t) = \sum_k v(kT)\, \mathrm{sinc}\left(\frac{t}{T} - k\right) e^{2\pi i \Delta(t-kT)}. \tag{4.71}$$

Theorems 4.6.2 and 4.6.3 apply to this more general case. That is, with $v(t) = \int_{\Delta-\mathrm{W}}^{\Delta+\mathrm{W}} \hat{v}(f) e^{2\pi i f t}\, df$, the function $v(t)$ is bounded and continuous and the series in (4.71) converges for all t. Similarly, if $\sum_k |v(kT)|^2 < \infty$, there is a unique continuous $\mathcal{L}_2$ function $\{v(t) : [\Delta - \mathrm{W}, \Delta + \mathrm{W}] \to \mathbb{C}\}$, $\mathrm{W} = 1/2T$, with those sample values.

4.7 Aliasing and the sinc-weighted sinusoid expansion

In this section an orthogonal expansion for arbitrary $\mathcal{L}_2$ functions called the *T-spaced sinc-weighted sinusoid expansion* is developed. This expansion is very similar to the T-spaced truncated sinusoid expansion discussed earlier, except that its set of orthogonal waveforms consists of time and frequency shifts of a sinc function rather than a rectangular function. This expansion is then used to discuss the important concept of degrees of freedom. Finally this same expansion is used to develop the concept of aliasing. This will help in understanding sampling for functions that are only approximately frequency-limited.

4.7.1 The T-spaced sinc-weighted sinusoid expansion

Let $u(t) \leftrightarrow \hat{u}(f)$ be an arbitrary $\mathcal{L}_2$ transform pair, and segment $\hat{u}(f)$ into intervals[33] of width 2W. Thus,

$$\hat{u}(f) = \text{l.i.m.}\sum_m \hat{v}_m(f), \qquad \text{where} \quad \hat{v}_m(f) = \hat{u}(f)\,\text{rect}\left(\frac{f}{2\text{W}} - m\right).$$

Note that $\hat{v}_0(f)$ is nonzero only in $[-\text{W}, \text{W}]$ and thus corresponds to an $\mathcal{L}_2$ function $v_0(t)$ baseband-limited to W. More generally, for arbitrary integer m, $\hat{v}_m(f)$ is nonzero only in $[\Delta - \text{W}, \Delta + \text{W}]$ for $\Delta = 2\text{W}m$. From (4.71), the inverse transform with $T = 1/2\text{W}$ satisfies the following:

$$\begin{aligned} v_m(t) &= \sum_k v_m(kT)\,\text{sinc}\left(\frac{t}{T} - k\right) e^{2\pi i(m/T)(t-kT)} \\ &= \sum_k v_m(kT)\,\text{sinc}\left(\frac{t}{T} - k\right) e^{2\pi i m t/T}. \end{aligned} \tag{4.72}$$

Combining all of these frequency segments,

$$u(t) = \text{l.i.m.}\sum_m v_m(t) = \text{l.i.m.}\sum_{m,k} v_m(kT)\,\text{sinc}\left(\frac{t}{T} - k\right) e^{2\pi i m t/T}. \tag{4.73}$$

This converges in $\mathcal{L}_2$, but does not not necessarily converge pointwise because of the infinite summation over m. It expresses an arbitrary $\mathcal{L}_2$ function $u(t)$ in terms of the samples of each frequency slice, $v_m(t)$, of $u(t)$.

This is an orthogonal expansion in the doubly indexed set of functions

$$\left\{\psi_{m,k}(t) = \text{sinc}\left(\frac{t}{T} - k\right) e^{2\pi i m t/T};\ m, k \in \mathbb{Z}\right\}. \tag{4.74}$$

These are the time and frequency shifts of the basic function $\psi_{0,0}(t) = \text{sinc}(t/T)$. The time shifts are in multiples of T and the frequency shifts are in multiples of $1/T$. This set of orthogonal functions is called the set of *T-spaced sinc-weighted sinusoids*.

The T-spaced sinc-weighted sinusoids and the T-spaced truncated sinusoids are quite similar. Each function in the first set is a time and frequency translate of $\text{sinc}(t/T)$. Each function in the second set is a time and frequency translate of $\text{rect}(t/T)$. Both sets are made up of functions separated by multiples of T in time and $1/T$ in frequency.

[33] The boundary points between frequency segments can be ignored, as in the case for time segments.

4.7.2 Degrees of freedom

An important rule of thumb used by communication engineers is that the class of real functions that are approximately baseband-limited to W_0 and approximately time-limited to $[-T_0/2, T_0/2]$ have about $2T_0W_0$ real degrees of freedom if $T_0W_0 \gg 1$. This means that any function within that class can be specified approximately by specifying about $2T_0W_0$ real numbers as coefficients in an orthogonal expansion. The same rule is valid for complex functions in terms of complex degrees of freedom.

This somewhat vague statement is difficult to state precisely, since time-limited functions cannot be frequency-limited and vice versa. However, the concept is too important to ignore simply because of lack of precision. Thus several examples are given.

First, consider applying the sampling theorem to real (complex) functions $u(t)$ that are strictly baseband-limited to W_0. Then $u(t)$ is specified by its real (complex) samples at rate $2W_0$. If the samples are nonzero only within the interval $[-T_0/2, T_0/2]$, then there are about $2T_0W_0$ nonzero samples, and these specify $u(t)$ within this class. Here a precise class of functions have been specified, but *functions* that are zero outside of an interval have been replaced with functions whose *samples* are zero outside of the interval.

Second, consider complex functions $u(t)$ that are again strictly baseband-limited to W_0, but now apply the sinc-weighted sinusoid expansion with $W = W_0/(2n+1)$ for some positive integer n. That is, the band $[-W_0, W_0]$ is split into $2n+1$ slices and each slice is expanded in a sampling-theorem expansion. Each slice is specified by samples at rate $2W$, so all slices are specified collectively by samples at an aggregate rate $2W_0$ as before. If the samples are nonzero only within $[-T_0/2, T_0/2]$, then there are about[34] $2T_0W_0$ nonzero complex samples that specify any $u(t)$ in this class.

If the functions in this class are further constrained to be real, then the coefficients for the central frequency slice are real and the negative slices are specified by the positive slices. Thus each real function in this class is specified by about $2T_0W_0$ real numbers.

This class of functions is slightly different for each choice of n, since the detailed interpretation of what "approximately time-limited" means is changing. From a more practical perspective, however, all of these expansions express an approximately baseband-limited waveform by samples at rate $2W_0$. As the overall duration T_0 of the class of waveforms increases, the initial transient due to the samples centered close to $-T_0/2$ and the final transient due to samples centered close to $T_0/2$ should become unimportant relative to the rest of the waveform.

The same conclusion can be reached for functions that are strictly time-limited to $[-T_0/2, T_0/2]$ by using the truncated sinusoid expansion with coefficients outside of $[-W_0, W_0]$ set to 0.

[34] Calculating this number of samples carefully yields $(2n+1)\left(1+\left\lfloor \frac{T_0W_0}{2n+1}\right\rfloor\right)$.

In summary, all the above expansions require roughly $2W_0T_0$ numbers for the approximate specification of a waveform essentially limited to time T_0 and frequency W_0 for T_0W_0 large.

It is possible to be more precise about the number of degrees of freedom in a given time and frequency band by looking at the prolate spheroidal waveform expansion (see Appendix 5.5.3). The orthogonal waveforms in this expansion maximize the energy in the given time/frequency region in a certain sense. It is perhaps simpler and better, however, to live with the very approximate nature of the arguments based on the sinc-weighted sinusoid expansion and the truncated sinusoid expansion.

4.7.3 Aliasing – a time-domain approach

Both the truncated sinusoid and the sinc-weighted sinusoid expansions are conceptually useful for understanding waveforms that are approximately time- and bandwidth-limited, but in practice waveforms are usually sampled, perhaps at a rate much higher than twice the nominal bandwidth, before digitally processing the waveforms. Thus it is important to understand the error involved in such sampling.

Suppose an $\mathcal{L}_2$ function $u(t)$ is sampled with T-spaced samples, $\{u(kT); k \in \mathbb{Z}\}$. Let $s(t)$ denote the approximation to $u(t)$ that results from the sampling theorem expansion:

$$s(t) = \sum_k u(kT)\,\mathrm{sinc}\left(\frac{t}{T} - k\right). \tag{4.75}$$

If $u(t)$ is baseband-limited to $W = 1/2T$, then $s(t) = u(t)$, but here it is no longer assumed that $u(t)$ is baseband-limited. The expansion of $u(t)$ into individual frequency slices, repeated below from (4.73), helps in understanding the difference between $u(t)$ and $s(t)$:

$$u(t) = \mathrm{l.i.m.} \sum_{m,k} v_m(kT)\,\mathrm{sinc}\left(\frac{t}{T} - k\right) e^{2\pi i m t/T}, \tag{4.76}$$

where

$$v_m(t) = \int \hat{u}(f)\,\mathrm{rect}(fT - m) e^{2\pi i f t}\,\mathrm{d}f. \tag{4.77}$$

For an arbitrary $\mathcal{L}_2$ function $u(t)$, the sample points $u(kT)$ might be at points of discontinuity and thus be ill defined. Also (4.75) need not converge, and (4.76) might not converge pointwise. To avoid these problems, $\hat{u}(f)$ will later be restricted beyond simply being $\mathcal{L}_2$. First, however, questions of convergence are disregarded and the relevant equations are derived without questioning when they are correct.

From (4.75), the samples of $s(t)$ are given by $s(kT) = u(kT)$, and combining with (4.76) we obtain

$$s(kT) = u(kT) = \sum_m v_m(kT). \tag{4.78}$$

Thus the samples from different frequency slices are summed together in the samples of $u(t)$. This phenomenon is called *aliasing*. There is no way to tell, from the samples $\{u(kT); k \in \mathbb{Z}\}$ alone, how much contribution comes from each frequency slice and thus, as far as the samples are concerned, every frequency band is an "alias" for every other.

Although $u(t)$ and $s(t)$ agree at the sample times, they differ elsewhere (assuming that $u(t)$ is not strictly baseband-limited to $1/2T$). Combining (4.78) and (4.75) we obtain

$$s(t)=\sum_k\sum_m v_m(kT)\,\mathrm{sinc}\left(\frac{t}{T}-k\right). \tag{4.79}$$

The expresssions in (4.79) and (4.76) agree at $m=0$, so the difference between $u(t)$ and $s(t)$ is given by

$$u(t)-s(t)=\sum_k\sum_{m\neq 0}-v_m(kT)\,\mathrm{sinc}\left(\frac{t}{T}-k\right)+\sum_k\sum_{m\neq 0}v_m(kT)\mathrm{e}^{2\pi imt/T}\mathrm{sinc}\left(\frac{t}{T}-k\right).$$

The first term above is $v_0(t)-s(t)$, i.e. the difference in the nominal baseband [−W, W]. This is the error caused by the aliased terms in $s(t)$. The second term is the energy in the nonbaseband portion of $u(t)$, which is orthogonal to the first error term. Since each term is an orthogonal expansion in the sinc-weighted sinusoids of (4.74), the energy in the error is given by[35]

$$\int\left|u(t)-s(t)\right|^2\mathrm{d}t=T\sum_k\left|\sum_{m\neq 0}v_m(kT)\right|^2+T\sum_k\sum_{m\neq 0}\left|v_m(kT)\right|^2. \tag{4.80}$$

Later, when the source waveform $u(t)$ is viewed as a sample function of a random process $U(t)$, it will be seen that under reasonable conditions the expected values of these two error terms are approximately equal. Thus, if $u(t)$ is filtered by an ideal lowpass filter before sampling, then $s(t)$ becomes equal to $v_0(t)$ and only the second error term in (4.80) remains; this reduces the expected MSE roughly by a factor of 2. It is often easier, however, simply to sample a little faster.

4.7.4 Aliasing – a frequency-domain approach

Aliasing can be, and usually is, analyzed from a frequency-domain standpoint. From (4.79), $s(t)$ can be separated into the contribution from each frequency band as follows:

$$s(t)=\sum_m s_m(t),\qquad \text{where}\quad s_m(t)=\sum_k v_m(kT)\,\mathrm{sinc}\left(\frac{t}{T}-k\right). \tag{4.81}$$

Comparing $s_m(t)$ to $v_m(t)=\sum_k v_m(kT)\,\mathrm{sinc}(t/T-k)\mathrm{e}^{2\pi imt/T}$, it is seen that

$$v_m(t)=s_m(t)\mathrm{e}^{2\pi imt/T}.$$

From the Fourier frequency shift relation, $\hat{v}_m(f)=\hat{s}_m(f-m/T)$, so

$$\hat{s}_m(f)=\hat{v}_m\left(f+\frac{m}{T}\right). \tag{4.82}$$

[35] As shown by example in Exercise 4.38, $s(t)$ need not be $\mathcal{L}_2$ unless the additional restrictions of Theorem 5.5.2 are applied to $\hat{u}(f)$. In these bizarre situations, the first sum in (4.80) is infinite and $s(t)$ is a complete failure as an approximation to $u(t)$.

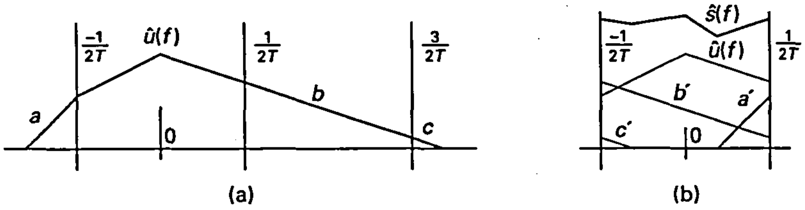


Figure 4.10. The transform $\hat{s}(f)$ of the baseband-sampled approximation $s(t)$ to $u(t)$ is constructed by folding the transform $\hat{u}(f)$ into $[-1/2T, 1/2T]$. For example, using real functions for pictorial clarity, the component a is mapped into a', b into b', and c into c'. These folded components are added to obtain $\hat{s}(f)$. If $\hat{u}(f)$ is complex, then both the real and imaginary parts of $\hat{u}(f)$ must be folded in this way to get the real and imaginary parts, respectively, of $\hat{s}(f)$. The figure further clarifies the two terms on the right of (4.80). The first term is the energy of $\hat{u}(f) - \hat{s}(f)$ caused by the folded components in part (b). The final term is the energy in part (a) outside of $[-T/2, T/2]$.

Finally, since $\hat{v}_m(f) = \hat{u}(f)\,\mathrm{rect}(fT - m)$, one sees that $\hat{v}_m(f + m/T) = \hat{u}(f + m/T)$ $\mathrm{rect}(fT)$. Thus, summing (4.82) over m, we obtain

$$\hat{s}(f) = \sum_m \hat{u}\left(f + \frac{m}{T}\right) \mathrm{rect}(fT). \tag{4.83}$$

Each frequency slice $\hat{v}_m(f)$ is shifted down to baseband in this equation, and then all these shifted frequency slices are summed together, as illustrated in Figure 4.10. This establishes the essence of the following aliasing theorem, which is proved in Appendix 5.5.2.

Theorem 4.7.1 (Aliasing theorem) *Let $\hat{u}(f)$ be $\mathcal{L}_2$, and let $\hat{u}(f)$ satisfy the condition $\lim_{|f|\to\infty} \hat{u}(f)|f|^{1+\varepsilon} = 0$ for some $\varepsilon > 0$. Then $\hat{u}(f)$ is $\mathcal{L}_1$, and the inverse Fourier transform $u(t) = \int \hat{u}(f)e^{2\pi i f t}\,\mathrm{d}f$ converges pointwise to a continuous bounded function. For any given $T > 0$, the sampling approximation $\sum_k u(kT)\,\mathrm{sinc}(t/T - k)$ converges pointwise to a continuous bounded $\mathcal{L}_2$ function $s(t)$. The Fourier transform of $s(t)$ satisfies*

$$\hat{s}(f) = \text{l.i.m.} \sum_m \hat{u}\left(f + \frac{m}{T}\right) \mathrm{rect}(fT). \tag{4.84}$$

The condition that $\lim \hat{u}(f)f^{1+\varepsilon} = 0$ implies that $\hat{u}(f)$ goes to 0 with increasing f at a faster rate than $1/f$. Exercise 4.37 gives an example in which the theorem fails in the absence of this condition.

Without the mathematical convergence details, what the aliasing theorem says is that, corresponding to a Fourier transform pair $u(t) \leftrightarrow \hat{u}(f)$, there is another Fourier transform pair $s(t) \leftrightarrow \hat{s}(f)$; $s(t)$ is a baseband sampling expansion using the T-spaced samples of $u(t)$, and $\hat{s}(f)$ is the result of folding the transform $\hat{u}(f)$ into the band $[-\mathrm{W}, \mathrm{W}]$ with $\mathrm{W} = 1/2T$.

4.8 Summary

The theory of $\mathcal{L}_2$ (finite-energy) functions has been developed in this chapter. These are, in many ways, the ideal waveforms to study, both because of the simplicity and generality of their mathematical properties and because of their appropriateness for modeling both source waveforms and channel waveforms.

For encoding source waveforms, the general approach is as follows:

- expand the waveform into an orthogonal expansion;
- quantize the coefficients in that expansion;
- use discrete source coding on the quantizer output.

The distortion, measured as the energy in the difference between the source waveform and the reconstructed waveform, is proportional to the squared quantization error in the quantized coefficients.

For encoding waveforms to be transmitted over communication channels, the approach is as follows:

- map the incoming sequence of binary digits into a sequence of real or complex symbols;
- use the symbols as coefficients in an orthogonal expansion.

Orthogonal expansions have been discussed in this chapter and will be further discussed in Chapter 5. Chapter 6 will discuss the choice of symbol set, the mapping from binary digits, and the choice of orthogonal expansion.

This chapter showed that every $\mathcal{L}_2$ time-limited waveform has a Fourier series, where each Fourier coefficient is given as a Lebesgue integral and the Fourier series converges in $\mathcal{L}_2$, i.e. as more and more Fourier terms are used in approximating the function, the energy difference between the waveform and the approximation gets smaller and approaches 0 in the limit.

Also, by the Plancherel theorem, every $\mathcal{L}_2$ waveform $u(t)$ (time-limited or not) has a Fourier integral $\hat{u}(f)$. For each truncated approximation, $u_A(t) = u(t)\ \mathrm{rect}(t/2A)$, the Fourier integral $\hat{u}_A(f)$ exists with pointwise convergence and is continuous. The Fourier integral $\hat{u}(f)$ is then the $\mathcal{L}_2$ limit of these approximation waveforms. The inverse transform exists in the same way.

These powerful $\mathcal{L}_2$-convergence results for Fourier series and integrals are not needed for computing the Fourier transforms and series for the conventional waveforms appearing in exercises. They become important both when the waveforms are sample functions of random processes and when one wants to find limits on possible performance. In both of these situations, one is dealing with a large class of potential waveforms, rather than a single waveform, and these general results become important.

The DTFT is the frequency–time dual of the Fourier series, and the sampling theorem is simply the Fourier transform of the DTFT, combined with a little care about convergence.

The T-spaced truncated sinusoid expansion and the T-spaced sinc-weighted sinusoid expansion are two orthogonal expansions of an arbitrary $\mathcal{L}_2$ waveform. The first is formed by segmenting the waveform into T-length segments and expanding each segment in a Fourier series. The second is formed by segmenting the waveform in frequency and sampling each frequency band. The orthogonal waveforms in each are the time–frequency translates of $\mathrm{rect}(t/T)$ for the first case and $\mathrm{sinc}(t/T)$ for the second. Each expansion leads to the notion that waveforms roughly limited to a time interval T_0 and a baseband frequency interval W_0 have approximately $2T_0\mathrm{W}_0$ degrees of freedom when $T_0\mathrm{W}_0$ is large.

Aliasing is the ambiguity in a waveform that is represented by its T-spaced samples. If an $\mathcal{L}_2$ waveform is baseband-limited to $1/2T$, then its samples specify the waveform, but if the waveform has components in other bands, these components are aliased with the baseband components in the samples. The aliasing theorem says that the Fourier transform of the baseband reconstruction from the samples is equal to the original Fourier transform folded into that baseband.

4.9 Appendix: Supplementary material and proofs

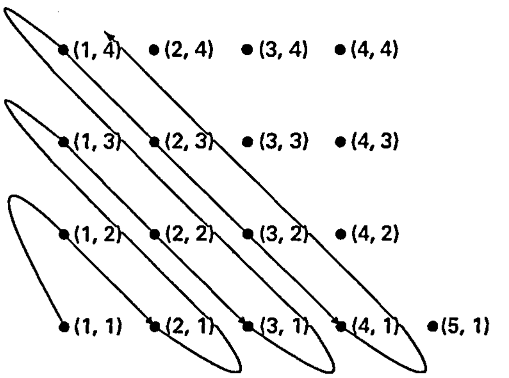
The first part of the appendix is an introduction to countable sets. These results are used throughout the chapter, and the material here can serve either as a first exposure or a review. The following three parts of the appendix provide added insight and proofs about the results on measurable sets.

4.9.1 Countable sets

A collection of distinguishable objects is *countably-infinite* if the objects can be put into one-to-one correspondence with the positive integers. Stated more intuitively, the collection is countably infinite if the set of elements can be arranged as a sequence $a_1, a_2, \ldots$ A set is countable if it contains either a finite or countably infinite set of elements.

Example 4.9.1 (The set of all integers) The integers can be arranged as the sequence $0, -1, +1, -2, +2, -3, \ldots$, and thus the set is countably infinite. Note that each integer appears once and only once in this sequence, and the one-to-one correspondence is $(0 \leftrightarrow 1)$, $(-1 \leftrightarrow 2)$, $(+1 \leftrightarrow 3)$, $(-2 \leftrightarrow 4)$, etc. There are many other ways to list the integers as a sequence, such as $0, -1, +1, +2, -2, +3, +4, -3, +5, \ldots$, but, for example, listing all the nonnegative integers first followed by all the negative integers is not a valid one-to-one correspondence since there are no positive integers left over for the negative integers to map into.

Example 4.9.2 (The set of 2-tuples of positive integers) Figure 4.11 shows that this set is countably infinite by showing one way to list the elements in a sequence. Note that every 2-tuple is eventually reached in this list. In a weird sense, this means that there are as many positive integers as there are pairs of positive integers, but what is happening is that the integers in the 2-tuple advance much more slowly than the position in the list. For example, it can be verified that (n, n) appears in position $2n(n-1)+1$ of the list.



1 ↔ (1, 1)
2 ↔ (1, 2)
3 ↔ (2, 1)
4 ↔ (1, 3)
5 ↔ (2, 2)
6 ↔ (3, 1)
7 ↔ (1, 4)
and so forth

Figure 4.11. One-to-one correspondence between positive integers and 2-tuples of positive integers.

By combining the ideas in the previous two examples, it can be seen that the collection of all integer 2-tuples is countably infinite. With a little more ingenuity, it can be seen that the set of integer n-tuples is countably infinite for all positive integer n. Finally, it is straightforward to verify that any subset of a countable set is also countable. Also a finite union of countable sets is countable, and in fact a countable union of countable sets must be countable.

Example 4.9.3 (The set of rational numbers) Each rational number can be represented by an integer numerator and denominator, and can be uniquely represented by its irreducible numerator and denominator. Thus the rational numbers can be put into one-to-one correspondence with a subset of the collection of 2-tuples of integers, and are thus countable. The rational numbers in the interval $[-T/2, T/2]$ for any given $T > 0$ form a subset of all rational numbers, and therefore are countable also.

As seen in Section 4.3.1, any countable set of numbers $a_1, a_2, \ldots$ can be expressed as a disjoint countable union of zero-measure sets, $[a_1, a_1]$, $[a_2, a_2], \ldots$, so the measure of any countable set is zero. Consider a function that has the value 1 at each rational argument and 0 elsewhere.

The Lebesgue integral of that function is 0. Since rational numbers exist in every positive-sized interval of the real line, no matter how small, the Riemann integral of this function is undefined. This function is not of great practical interest, but provides insight into why Lebesgue integration is so general.

Example 4.9.4 (The set of binary sequences) An example of an *uncountable* set of elements is the set of (unending) sequences of binary digits. It will be shown that this set contains uncountably many elements by assuming the contrary and constructing a contradiction. Thus, suppose we can list all binary sequences, $a_1, a_2, a_3, \ldots$ Each sequence, a_n, can be expressed as $a_n = (a_{n,1}, a_{n,2}, \ldots)$, resulting in a doubly infinite array of binary digits. We now construct a new binary sequence $b = b_1, b_2, \ldots$ in the following way. For each integer $n > 0$, choose $b_n \neq a_{n,n}$; since b_n is binary, this specifies b_n for each n and thus specifies b. Now b differs from each of the listed sequences in at least one binary digit, so that b is a binary sequence not on the list. This is a contradiction, since by assumption the list contains each binary sequence.

This example clearly extends to ternary sequences and sequences from any alphabet with more than one member.

Example 4.9.5 (The set of real numbers in $[0, 1)$) This is another uncountable set, and the proof is very similar to that of Example 4.9.4. Any real number $r \in [0, 1)$ can be represented as a binary expansion $0.r_1r_2, \ldots$ whose elements r_k are chosen to satisfy $r = \sum_{k=1}^{\infty} r_k 2^{-k}$ and where each $r_k \in \{0, 1\}$. For example, 1/2 can be represented as 0.1, 3/8 as 0.011, etc. This expansion is unique except in the special cases where r can be represented by a finite binary expansion, $r = \sum_{k=1}^{m} r_k$; for example, 1/2 can also be represented as $0.0111\cdots$. By convention, for each such r (other than $r = 0$) choose m as small as possible; thus in the infinite expansion, $r_m = 1$ and $r_k = 0$ for all $k > m$. Each such number can be alternatively represented with $r_m = 0$ and $r_k = 1$ for all $k > m$.

By convention, map each such r into the expansion terminating with an infinite sequence of 0s. The set of binary sequences is then the union of the representations of the reals in $[0, 1)$ and the set of binary sequences terminating in an infinite sequence of 1s. This latter set is countable because it is in one-to-one correspondence with the rational numbers of the form $\sum_{k=1}^{m} r_k 2^{-k}$ with binary r_k and finite m. Thus if the reals were countable, their union with this latter set would be countable, contrary to the known uncountability of the binary sequences.

By scaling the interval [0,1), it can be seen that the set of real numbers in any interval of nonzero size is uncountably infinite. Since the set of rational numbers in such an interval is countable, the irrational numbers must be uncountable (otherwise the union of rational and irrational numbers, i.e. the reals, would be countable).

The set of irrationals in $[-T/2, T/2]$ is the complement of the rationals and thus has measure T. Each pair of distinct irrationals is separated by rational numbers. Thus the irrationals can be represented as a union of intervals only by using an uncountable union[36] of intervals, each containing a single element. The class of uncountable unions of intervals is not very interesting since it includes all subsets of $\mathbb{R}$.

4.9.2 Finite unions of intervals over $[-T/2, T/2]$

Let $\mathcal{M}_f$ be the class of finite unions of intervals, i.e. the class of sets whose elements can each be expressed as $\mathcal{E} = \bigcup_{j=1}^{\ell} I_j$, where $\{I_1, \ldots, I_\ell\}$ are intervals and $\ell \geq 1$ is an integer. Exercise 4.5 shows that each such $\mathcal{E} \in \mathcal{M}_f$ can be uniquely expressed as a finite union of $k \leq \ell$ *separated* intervals, say $\mathcal{E} = \bigcup_{j=1}^{k} I'_j$. The measure of $\mathcal{E}$ was defined as $\mu(\mathcal{E}) = \sum_{j=1}^{k} \mu(I'_j)$. Exercise 4.7 shows that $\mu(\mathcal{E}) \leq \sum_{j=1}^{\ell} \mu(I_j)$ for the original

[36] This might be a shock to one's intuition. Each partial union $\bigcup_{j=1}^{k} [a_j, a_j]$ of rationals has a complement which is the union of $k+1$ intervals of nonzero width; each unit increase in k simply causes one interval in the complement to split into two smaller intervals (although maintaining the measure at T). In the limit, however, this becomes an uncountable set of separated points.

intervals making up $\mathcal{E}$ and shows that this holds with equality whenever $I_1, \dots, I_\ell$ are disjoint.[37]

The class $\mathcal{M}_f$ is closed under the union operation, since if $\mathcal{E}_1$ and $\mathcal{E}_2$ are each finite unions of intervals, then $\mathcal{E}_1 \cup \mathcal{E}_2$ is the union of both sets of intervals. It also follows from this that if $\mathcal{E}_1$ and $\mathcal{E}_2$ are disjoint then

$$\mu(\mathcal{E}_1 \cup \mathcal{E}_2) = \mu(\mathcal{E}_1) + \mu(\mathcal{E}_2). \tag{4.85}$$

The class $\mathcal{M}_f$ is also closed under the intersection operation, since, if $\mathcal{E}_1 = \bigcup_j I_{1,j}$ and $\mathcal{E}_2 = \bigcup_\ell I_{2,\ell}$, then $\mathcal{E}_1 \cap \mathcal{E}_2 = \bigcup_{j,\ell}(I_{1,j} \cap I_{2,\ell})$. Finally, $\mathcal{M}_f$ is closed under complementation. In fact, as illustrated in Figure 4.5, the complement $\overline{\mathcal{E}}$ of a finite union of separated intervals $\mathcal{E}$ is simply the union of separated intervals lying between the intervals of $\mathcal{E}$. Since $\mathcal{E}$ and its complement $\overline{\mathcal{E}}$ are disjoint and fill all of $[-T/2, T/2]$, each $\mathcal{E} \in \mathcal{M}_f$ satisfies the complement property,

$$T = \mu(\mathcal{E}) + \mu(\overline{\mathcal{E}}). \tag{4.86}$$

An important generalization of (4.85) is the following: for any $\mathcal{E}_1, \mathcal{E}_2 \in \mathcal{M}_f$,

$$\mu(\mathcal{E}_1 \cup \mathcal{E}_2) + \mu(\mathcal{E}_1 \cap \mathcal{E}_2) = \mu(\mathcal{E}_1) + \mu(\mathcal{E}_2). \tag{4.87}$$

To see this intuitively, note that each interval in $\mathcal{E}_1 \cap \mathcal{E}_2$ is counted twice on each side of (4.87), whereas each interval in only $\mathcal{E}_1$ or only $\mathcal{E}_2$ is counted once on each side. More formally, $\mathcal{E}_1 \cup \mathcal{E}_2 = \mathcal{E}_1 \cup (\mathcal{E}_2 \cap \overline{\mathcal{E}_1})$. Since this is a disjoint union, (4.85) shows that $\mu(\mathcal{E}_1 \cup \mathcal{E}_2) = \mu(\mathcal{E}_1) + \mu(\mathcal{E}_2 \cap \overline{\mathcal{E}_1})$. Similarly, $\mu(\mathcal{E}_2) = \mu(\mathcal{E}_2 \cap \mathcal{E}_1) + \mu(\mathcal{E}_2 \cap \overline{\mathcal{E}_1})$. Combining these equations results in (4.87).

4.9.3 Countable unions and outer measure over $[-T/2, T/2]$

Let $\mathcal{M}_c$ be the class of countable unions of intervals, i.e. each set $\mathcal{B} \in \mathcal{M}_c$ can be expressed as $\mathcal{B} = \bigcup_j I_j$, where $\{I_1, I_2, \dots\}$ is either a finite or countably infinite collection of intervals. The class $\mathcal{M}_c$ is closed under both the union operation and the intersection operation by the same argument as used for $\mathcal{M}_f$. Note that $\mathcal{M}_c$ is also closed under countable unions (see Exercise 4.8) but not closed under complements or countable intersections.[38]

Each $\mathcal{B} \in \mathcal{M}_c$ can be uniquely[39] expressed as a countable union of separated intervals, say $\mathcal{B} = \bigcup_j I'_j$, where $\{I'_1, I'_2, \dots\}$ are separated (see Exercise 4.6). The measure of $\mathcal{B}$ is defined as

[37] Recall that intervals such as (0,1], (1,2] are disjoint but not separated. A set $\mathcal{E} \in \mathcal{M}_f$ has many representations as disjoint intervals but only one as separated intervals, which is why the definition refers to separated intervals.

[38] Appendix 4.9.1 shows that the complement of the rationals, i.e. the set of irrationals, does not belong to $\mathcal{M}_c$. The irrationals can also be viewed as the intersection of the complements of the rationals, giving an example where $\mathcal{M}_c$ is not closed under countable intersections.

[39] What is unique here is the *collection* of intervals, not the *particular ordering*; this does not affect the infinite sum in (4.88) (see Exercise 4.4).

$$\mu(\mathcal{B}) = \sum_j \mu(I'_j). \tag{4.88}$$

As shown in Section 4.3.1, the right side of (4.88) always converges to a number between 0 and T. For $B = \bigcup_j I_j$, where $I_1, I_2, \ldots$ are arbitrary intervals, Exercise 4.7 establishes the following union bound:

$$\mu(\mathcal{B}) \leq \sum_j \mu(I_j) \qquad \text{with equality if } I_1, I_2, \ldots \text{ are disjoint.} \tag{4.89}$$

The *outer measure* $\mu^{\circ}(\mathcal{A})$ of an arbitary set $\mathcal{A}$ was defined in (4.13) as

$$\mu^{\circ}(\mathcal{A}) = \inf_{\mathcal{B} \in \mathcal{M}_c,\, \mathcal{A} \subseteq \mathcal{B}} \mu(\mathcal{B}). \tag{4.90}$$

Note that $[-T/2, T/2]$ is a cover of $\mathcal{A}$ for all $\mathcal{A}$ (recall that only sets in $[-T/2, T/2]$ are being considered). Thus $\mu^{\circ}(\mathcal{A})$ must lie between 0 and T for all $\mathcal{A}$. Also, for any two sets $\mathcal{A} \subseteq \mathcal{A}'$, any cover of $\mathcal{A}'$ also covers $\mathcal{A}$. This implies the *subset inequality* for outer measure:

$$\mu^{\circ}(\mathcal{A}) \leq \mu^{\circ}(\mathcal{A}') \qquad \text{for } \mathcal{A} \subseteq \mathcal{A}'. \tag{4.91}$$

The following lemma develops the *union bound* for outer measure called the *union bound.* Its proof illustrates several techniques that will be used frequently.

Lemma 4.9.1 *Let $\mathcal{S} = \bigcup_k \mathcal{A}_k$ be a countable union of arbitrary sets in* $[-T/2, T/2]$. *Then*

$$\mu^{\circ}(\mathcal{S}) \leq \sum_k \mu^{\circ}(\mathcal{A}_k). \tag{4.92}$$

Proof The approach is first to establish an arbitrarily tight cover to each $\mathcal{A}_k$ and then show that the union of these covers is a cover for $\mathcal{S}$. Specifically, let ε be an arbitrarily small positive number. For each $k \geq 1$, the infimum in (4.90) implies that covers exist with measures arbitrarily little greater than that infimum. Thus a cover $\mathcal{B}_k$ to $\mathcal{A}_k$ exists with

$$\mu(\mathcal{B}_k) \leq \varepsilon 2^{-k} + \mu^{\circ}(\mathcal{A}_k).$$

For each k, let $\mathcal{B}_k = \bigcup_j I'_{j,k}$, where $I'_{1,k}, I'_{2,k}, \ldots$ represents $\mathcal{B}_k$ by separated intervals. Then $\mathcal{B} = \bigcup_k \mathcal{B}_k = \bigcup_k \bigcup_j I'_{j,k}$ is a countable union of intervals, so, from (4.89) and Exercise 4.4, we have

$$\mu(\mathcal{B}) \leq \sum_k \sum_j \mu(I'_{j,k}) = \sum_k \mu(\mathcal{B}_k).$$

Since $\mathcal{B}_k$ covers $\mathcal{A}_k$ for each k, it follows that $\mathcal{B}$ covers $\mathcal{S}$. Since $\mu^{\circ}(S)$ is the infimum of its covers,

$$\mu^{\circ}(S) \leq \mu(\mathcal{B}) \leq \sum_k \mu(\mathcal{B}_k) \leq \sum_k \left(\varepsilon 2^{-k} + \mu^{\circ}(\mathcal{A}_k)\right) = \varepsilon + \sum_k \mu^{\circ}(\mathcal{A}_k).$$

Since $\varepsilon > 0$ is arbitrary, (4.92) follows. □

An important special case is the union of any set $\mathcal{A}$ and its complement $\overline{\mathcal{A}}$. Since $[-T/2, T/2] = \mathcal{A} \cup \overline{\mathcal{A}}$,

$$T \leq \mu^{\text{o}}(\mathcal{A}) + \mu^{\text{o}}(\overline{\mathcal{A}}). \tag{4.93}$$

Section 4.9.4 will define measurability and measure for arbitrary sets. Before that, the following theorem shows both that countable unions of intervals are measurable and that their measure, as defined in (4.88), is consistent with the general definition to be given later.

Theorem 4.9.1 *Let $\mathcal{B} = \bigcup_j I_j$, where $\{I_1, I_2, \dots\}$ is a countable collection of intervals in $[-T/2, T/2]$ (i.e., $\mathcal{B} \in \mathcal{M}_c$). Then*

$$\mu^{\text{o}}(\mathcal{B}) + \mu^{\text{o}}(\overline{\mathcal{B}}) = T \tag{4.94}$$

and

$$\mu^{\text{o}}(\mathcal{B}) = \mu(\mathcal{B}). \tag{4.95}$$

Proof Let $\{I'_j;\, j \geq 1\}$ be the collection of separated intervals representing $\mathcal{B}$ and let

$$\mathcal{E}^k = \bigcup_{j=1}^{k} I'_j;$$

then

$$\mu(\mathcal{E}^1) \leq \mu(\mathcal{E}^2) \leq \mu(\mathcal{E}^3) \leq \cdots \leq \lim_{k\to\infty} \mu(\mathcal{E}^k) = \mu(\mathcal{B}).$$

For any $\varepsilon > 0$, choose k large enough that

$$\mu(\mathcal{E}^k) \geq \mu(\mathcal{B}) - \varepsilon. \tag{4.96}$$

The idea of the proof is to approximate $\mathcal{B}$ by $\mathcal{E}^k$, which, being in $\mathcal{M}_f$, satisfies $T = \mu(\mathcal{E}^k) + \mu(\overline{\mathcal{E}^k})$. Thus,

$$\mu(\overline{\mathcal{B}}) \leq \mu(\mathcal{E}^k) + \varepsilon = T - \mu(\overline{\mathcal{E}^k}) + \varepsilon \leq T - \mu^{\text{o}}(\overline{\mathcal{B}}) + \varepsilon, \tag{4.97}$$

where the final inequality follows because $\mathcal{E}^k \subseteq \mathcal{B}$, and thus $\overline{\mathcal{B}} \subseteq \overline{\mathcal{E}^k}$ and $\mu^{\text{o}}(\overline{\mathcal{B}}) \leq \mu(\overline{\mathcal{E}^k})$.

Next, since $\mathcal{B} \in \mathcal{M}_c$ and $\mathcal{B} \subseteq \mathcal{B}$, $\mathcal{B}$ is a cover of itself and is a choice in the infimum defining $\mu^{\text{o}}(\mathcal{B})$; thus, $\mu^{\text{o}}(\overline{\mathcal{B}}) \leq \mu(\overline{\mathcal{B}})$. Combining this with (4.97), $\mu^{\text{o}}(\mathcal{B}) + \mu^{\text{o}}(\overline{\mathcal{B}}) \leq T + \varepsilon$. Since $\varepsilon > 0$ is arbitrary, this implies

$$\mu^{\text{o}}(\mathcal{B}) + \mu^{\text{o}}(\overline{\mathcal{B}}) \leq T. \tag{4.98}$$

This combined with (4.93) establishes (4.94). Finally, substituting $T \leq \mu^{\text{o}}(\mathcal{B}) + \mu^{\text{o}}(\overline{\mathcal{B}})$ into (4.97), $\mu(\mathcal{B}) \leq \mu^{\text{o}}(\mathcal{B}) + \varepsilon$. Since $\mu^{\text{o}}(\mathcal{B}) \leq \mu(\mathcal{B})$ and $\varepsilon > 0$ is arbitrary, this establishes (4.95). □

Finally, before proceeding to arbitrary measurable sets, the joint union and intersection property, (4.87), is extended to $\mathcal{M}_c$.

Lemma 4.9.2 *Let $\mathcal{B}_1$ and $\mathcal{B}_2$ be arbitrary sets in $\mathcal{M}_c$. Then*

$$\mu(\mathcal{B}_1 \cup \mathcal{B}_2) + \mu(\mathcal{B}_1 \cap \mathcal{B}_2) = \mu(\mathcal{B}_1) + \mu(\mathcal{B}_2). \tag{4.99}$$

Proof Let $\mathcal{B}_1$ and $\mathcal{B}_2$ be represented, respectively, by separated intervals, $\mathcal{B}_1 = \bigcup_j I_{1,j}$ and $\mathcal{B}_2 = \bigcup_j I_{2,j}$. For $\ell = 1, 2$, let $\mathcal{E}_\ell^k = \bigcup_{j=1}^k I_{\ell,j}$ and $\mathcal{D}_\ell^k = \bigcup_{j=k+1}^\infty I_{\ell,j}$. Thus $\mathcal{B}_\ell = \mathcal{E}_\ell^k \cup \mathcal{D}_\ell^k$ for each integer $k \geq 1$ and $\ell = 1, 2$. The proof is based on using $\mathcal{E}_\ell^k$, which is in $\mathcal{M}_f$ and satisfies the joint union and intersection property, as an approximation to $\mathcal{B}_\ell$. To see how this goes, note that

$$\mathcal{B}_1 \cap \mathcal{B}_2 = (\mathcal{E}_1^k \cup \mathcal{D}_1^k) \cap (\mathcal{E}_2^k \cup \mathcal{D}_2^k) = (\mathcal{E}_1^k \cap \mathcal{E}_2^k) \cup (\mathcal{E}_1^k \cap \mathcal{D}_2^k) \cup (\mathcal{D}_1^k \cap \mathcal{B}_2).$$

For any $\varepsilon > 0$ we can choose k large enough that $\mu(\mathcal{E}_\ell^k) \geq \mu(\mathcal{B}_\ell) - \varepsilon$ and $\mu(\mathcal{D}_\ell^k) \leq \varepsilon$ for $\ell = 1, 2$. Using the subset inequality and the union bound, we then have

$$\begin{aligned} \mu(\mathcal{B}_1 \cap \mathcal{B}_2) &\leq \mu(\mathcal{E}_1^k \cap \mathcal{E}_2^k) + \mu(\mathcal{D}_2^k) + \mu(\mathcal{D}_1^k) \\ &\leq \mu(\mathcal{E}_1^k \cap \mathcal{E}_2^k) + 2\varepsilon. \end{aligned}$$

By a similar but simpler argument,

$$\begin{aligned} \mu(\mathcal{B}_1 \cup \mathcal{B}_2) &\leq \mu(\mathcal{E}_1^k \cup \mathcal{E}_2^k) + \mu(\mathcal{D}_1^k) + \mu(\mathcal{D}_2^k) \\ &\leq \mu(\mathcal{E}_1^k \cup \mathcal{E}_2^k) + 2\varepsilon. \end{aligned}$$

Combining these inequalities and using (4.87) on $\mathcal{E}_1^k \subseteq \mathcal{M}_f$ and $\mathcal{E}_2^k \subseteq \mathcal{M}_f$, we have

$$\begin{aligned} \mu(\mathcal{B}_1 \cap \mathcal{B}_2) + \mu(\mathcal{B}_1 \cup \mathcal{B}_2) &\leq \mu(\mathcal{E}_1^k \cap \mathcal{E}_2^k) + \mu(\mathcal{E}_1^k \cup \mathcal{E}_2^k) + 4\varepsilon \\ &= \mu(\mathcal{E}_1^k) + \mu(\mathcal{E}_2^k) + 4\varepsilon \\ &\leq \mu(\mathcal{B}_1) + \mu(\mathcal{B}_2) + 4\varepsilon, \end{aligned}$$

where we have used the subset inequality in the final inequality.

For a bound in the opposite direction, we start with the subset inequality:

$$\begin{aligned} \mu(\mathcal{B}_1 \cup \mathcal{B}_2) + \mu(\mathcal{B}_1 \cap \mathcal{B}_2) &\geq \mu(\mathcal{E}_1^k \cup \mathcal{E}_2^k) + \mu(\mathcal{E}_1^k \cap \mathcal{E}_2^k) \\ &= \mu(\mathcal{E}_1^k) + \mu(\mathcal{E}_2^k) \\ &\geq \mu(\mathcal{B}_1) + \mu(\mathcal{B}_2) - 2\varepsilon. \end{aligned}$$

Since ε is arbitrary, these two bounds establish (4.99). □

4.9.4 Arbitrary measurable sets over $[-T/2, T/2]$

An arbitrary set $\mathcal{A} \in [-T/2, T/2]$ was defined to be *measurable* if

$$T = \mu^\circ(\mathcal{A}) + \mu^\circ(\overline{\mathcal{A}}). \tag{4.100}$$

The *measure* of a measurable set was defined to be $\mu(\mathcal{A}) = \mu^{\text{o}}(\mathcal{A})$. The class of measurable sets is denoted as $\mathcal{M}$. Theorem 4.9.1 shows that each set $\mathcal{B} \in \mathcal{M}_{\text{c}}$ is measurable, i.e. $\mathcal{B} \in \mathcal{M}$ and thus $\mathcal{M}_{\text{f}} \subseteq \mathcal{M}_{\text{c}} \subseteq \mathcal{M}$. The measure of $\mathcal{B} \in \mathcal{M}_{\text{c}}$ is $\mu(\mathcal{B}) = \sum_j \mu(I_j)$ for any disjoint sequence of intervals, $I_1, I_2, \dots$, whose union is $\mathcal{B}$.

Although the complements of sets in $\mathcal{M}_{\text{c}}$ are not necessarily in $\mathcal{M}_{\text{c}}$ (as seen from the rational number example), they must be in $\mathcal{M}$; in fact, from (4.100), all sets in $\mathcal{M}$ have complements in $\mathcal{M}$, i.e. $\mathcal{M}$ is closed under complements. We next show that $\mathcal{M}$ is closed under finite, and then countable, unions and intersections. The key to these results is to show first that the joint union and intersection property is valid for outer measure.

Lemma 4.9.3 *For any measurable sets $\mathcal{A}_1$ and $\mathcal{A}_2$,*

$$\mu^{\text{o}}(\mathcal{A}_1 \cup \mathcal{A}_2) + \mu^{\text{o}}(\mathcal{A}_1 \cap \mathcal{A}_2) = \mu^{\text{o}}(\mathcal{A}_1) + \mu^{\text{o}}(\mathcal{A}_2). \tag{4.101}$$

Proof The proof is very similar to that of Lemma 4.9.2, but here we use sets in $\mathcal{M}_{\text{c}}$ to approximate those in $\mathcal{M}$. For any $\varepsilon > 0$, let $\mathcal{B}_1$ and $\mathcal{B}_2$ be covers of $\mathcal{A}_1$ and $\mathcal{A}_2$, respectively, such that $\mu(\mathcal{B}_\ell) \leq \mu^{\text{o}}(\mathcal{A}_\ell) + \varepsilon$ for $\ell = 1, 2$. Let $\mathcal{D}_\ell = \mathcal{B}_\ell \cap \overline{\mathcal{A}_\ell}$ for $\ell = 1, 2$. Note that $\mathcal{A}_\ell$ and $\mathcal{D}_\ell$ are disjoint and $\mathcal{B}_\ell = \mathcal{A}_\ell \cup \mathcal{D}_\ell$:

$$\mathcal{B}_1 \cap \mathcal{B}_2 = (\mathcal{A}_1 \cup \mathcal{D}_1) \cap (\mathcal{A}_2 \cup \mathcal{D}_2) = (\mathcal{A}_1 \cap \mathcal{A}_2) \cup (\mathcal{D}_1 \cap \mathcal{A}_2) \cup (\mathcal{B}_1 \cap \mathcal{D}_2).$$

Using the union bound and subset inequality for outer measure on this and the corresponding expansion of $\mathcal{B}_1 \cup \mathcal{B}_2$, we obtain

$$\begin{aligned}
\mu(\mathcal{B}_1 \cap \mathcal{B}_2) &\leq \mu^{\text{o}}(\mathcal{A}_1 \cap \mathcal{A}_2) + \mu^{\text{o}}(\mathcal{D}_1) + \mu^{\text{o}}(\mathcal{D}_2) \leq \mu^{\text{o}}(\mathcal{A}_1 \cap \mathcal{A}_2) + 2\varepsilon, \\
\mu(\mathcal{B}_1 \cup \mathcal{B}_2) &\leq \mu^{\text{o}}(\mathcal{A}_1 \cup \mathcal{A}_2) + \mu^{\text{o}}(\mathcal{D}_1) + \mu^{\text{o}}(\mathcal{D}_2) \leq \mu^{\text{o}}(\mathcal{A}_1 \cup \mathcal{A}_2) + 2\varepsilon,
\end{aligned}$$

where we have also used the fact (see Exercise 4.9) that $\mu^{\text{o}}(\mathcal{D}_\ell) \leq \varepsilon$ for $\ell = 1, 2$. Summing these inequalities and rearranging terms, we obtain

$$\begin{aligned}
\mu^{\text{o}}(\mathcal{A}_1 \cup \mathcal{A}_2) + \mu^{\text{o}}(\mathcal{A}_1 \cap \mathcal{A}_2) &\geq \mu(\mathcal{B}_1 \cap \mathcal{B}_2) + \mu(\mathcal{B}_1 \cup \mathcal{B}_2) - 4\varepsilon \\
&= \mu(\mathcal{B}_1) + \mu(\mathcal{B}_2) - 4\varepsilon \\
&\geq \mu^{\text{o}}(\mathcal{A}_1) + \mu^{\text{o}}(\mathcal{A}_2) - 4\varepsilon,
\end{aligned}$$

where we have used (4.99) and then used $\mathcal{A}_\ell \subseteq \mathcal{B}_\ell$ for $\ell = 1, 2$. Using the subset inequality and (4.99) to bound in the opposite direction,

$$\mu(\mathcal{B}_1) + \mu(\mathcal{B}_2) = \mu(\mathcal{B}_1 \cup \mathcal{B}_2) + \mu(\mathcal{B}_1 \cap \mathcal{B}_2) \geq \mu^{\text{o}}(\mathcal{A}_1 \cup \mathcal{A}_2) + \mu^{\text{o}}(\mathcal{A}_1 \cap \mathcal{A}_2).$$

Rearranging and using $\mu(\mathcal{B}_\ell) \leq \mu^{\text{o}}(\mathcal{A}_\ell) + \varepsilon$, we obtain

$$\mu^{\text{o}}(\mathcal{A}_1 \cup \mathcal{A}_2) + \mu^{\text{o}}(\mathcal{A}_1 \cap \mathcal{A}_2) \leq \mu^{\text{o}}(\mathcal{A}_1) + \mu^{\text{o}}(\mathcal{A}_2) + 2\varepsilon.$$

Since ε is arbitrary, these bounds establish (4.101). □

Theorem 4.9.2 *Assume $\mathcal{A}_1, \mathcal{A}_2 \in \mathcal{M}$. Then $\mathcal{A}_1 \cup \mathcal{A}_2 \in \mathcal{M}$ and $\mathcal{A}_1 \cap \mathcal{A}_2 \in \mathcal{M}$.*

Proof Apply (4.101) to $\overline{\mathcal{A}_1}$ and $\overline{\mathcal{A}_2}$, to obtain

$$\mu^{\circ}(\overline{\mathcal{A}_1} \cup \overline{\mathcal{A}_2}) + \mu^{\circ}(\overline{\mathcal{A}_1} \cap \overline{\mathcal{A}_2}) = \mu^{\circ}(\overline{\mathcal{A}_1}) + \mu^{\circ}(\overline{\mathcal{A}_2}).$$

Rewriting $\overline{\mathcal{A}_1} \cup \overline{\mathcal{A}_2}$ as $\overline{\mathcal{A}_1 \cap \mathcal{A}_2}$ and $\overline{\mathcal{A}_1} \cap \overline{\mathcal{A}_2}$ as $\overline{\mathcal{A}_1 \cup \mathcal{A}_2}$ and adding this to (4.101) yields

$$\begin{aligned}\left[\mu^{\circ}(\mathcal{A}_1 \cup \mathcal{A}_2) + \mu^{\circ}(\overline{\mathcal{A}_1 \cup \mathcal{A}_2}\right] + \left[\mu^{\circ}(\mathcal{A}_1 \cap \mathcal{A}_2) + \mu^{\circ}(\overline{\mathcal{A}_1 \cap \mathcal{A}_2})\right] \\ = \mu^{\circ}(\mathcal{A}_1) + \mu^{\circ}(\mathcal{A}_2) + \mu^{\circ}(\overline{\mathcal{A}_1}) + \mu^{\circ}(\overline{\mathcal{A}_2}) \;=\; 2T,\end{aligned} \tag{4.102}$$

where we have used (4.100). Each of the bracketed terms above is at least T from (4.93), so each term must be exactly T. Thus $\mathcal{A}_1 \cup \mathcal{A}_2$ and $\mathcal{A}_1 \cap \mathcal{A}_2$ are measurable. □

Since $\mathcal{A}_1 \cup \mathcal{A}_2$ and $\mathcal{A}_1 \cap \mathcal{A}_2$ are measurable if $\mathcal{A}_1$ and $\mathcal{A}_2$ are, the joint union and intersection property holds for measure as well as outer measure for all measurable functions, i.e.

$$\mu(\mathcal{A}_1 \cup \mathcal{A}_2) + \mu(\mathcal{A}_1 \cap \mathcal{A}_2) = \mu(\mathcal{A}_1) + \mu(\mathcal{A}_2). \tag{4.103}$$

If $\mathcal{A}_1$ and $\mathcal{A}_2$ are disjoint, then (4.103) simplifies to the additivity property:

$$\mu(\mathcal{A}_1 \cup \mathcal{A}_2) = \mu(\mathcal{A}_1) + \mu(\mathcal{A}_2). \tag{4.104}$$

Actually, (4.103) shows that (4.104) holds whenever $\mu(\mathcal{A}_1 \cap \mathcal{A}_2) = 0$. That is, $\mathcal{A}_1$ and $\mathcal{A}_2$ need not be disjoint, but need only have an intersection of zero measure. This is another example in which sets of zero measure can be ignored.

The following theorem shows that $\mathcal{M}$ is closed over disjoint countable unions and that $\mathcal{M}$ is countably additive.

Theorem 4.9.3 *Assume that $\mathcal{A}_j \in \mathcal{M}$ for each integer $j \geq 1$ and that $\mu(\mathcal{A}_j \cap \mathcal{A}_\ell) = 0$ for all $j \neq \ell$. Let $\mathcal{A} = \bigcup_j \mathcal{A}_j$. Then $\mathcal{A} \in \mathcal{M}$ and*

$$\mu(\mathcal{A}) = \sum_j \mu(\mathcal{A}_j). \tag{4.105}$$

Proof Let $\mathcal{A}^k = \bigcup_{j=1}^k \mathcal{A}_j$ for each integer $k \geq 1$. Then $\mathcal{A}^{k+1} = \mathcal{A}^k \cup \mathcal{A}_{k+1}$ and, by induction on Theorem 4.9.2, $\mathcal{A}^k \in \mathcal{M}$ for all $k \geq 1$. It also follows that

$$\mu(\mathcal{A}^k) = \sum_{j=1}^k \mu(\mathcal{A}_j).$$

The sum on the right is nondecreasing in k and bounded by T, so the limit as $k \to \infty$ exists. Applying the union bound for outer measure to $\mathcal{A}$,

$$\mu^{\circ}(\mathcal{A}) \leq \sum_j \mu^{\circ}(\mathcal{A}_j) = \lim_{k\to\infty} \mu^{\circ}(\mathcal{A}^k) = \lim_{k\to\infty} \mu(\mathcal{A}^k). \tag{4.106}$$

Since $\mathcal{A}^k \subseteq \mathcal{A}$, we see that $\overline{\mathcal{A}} \subseteq \overline{\mathcal{A}^k}$ and $\mu^\circ(\overline{\mathcal{A}}) \le \mu(\overline{\mathcal{A}^k}) = T - \mu(\mathcal{A}^k)$. Thus

$$\mu^\circ(\overline{\mathcal{A}}) \le T - \lim_{k\to\infty} \mu(\mathcal{A}^k). \tag{4.107}$$

Adding (4.106) and (4.107) shows that $\mu^\circ(\mathcal{A}) + \mu^\circ(\overline{\mathcal{A}}) \le T$. Combining with (4.93), $\mu^\circ(\mathcal{A}) + \mu^\circ(\overline{\mathcal{A}}) = T$ and (4.106) and (4.107) are satisfied with equality. Thus $\mathcal{A} \in \mathcal{M}$ and countable additivity, (4.105), is satisfied. □

Next it is shown that $\mathcal{M}$ is closed under arbitrary countable unions and intersections.

Theorem 4.9.4 *Assume that $\mathcal{A}_j \in \mathcal{M}$ for each integer $j \ge 1$. Then $\mathcal{A} = \bigcup_j \mathcal{A}_j$ and $\mathcal{D} = \bigcap_j \mathcal{A}_j$ are both in $\mathcal{M}$.*

Proof Let $\mathcal{A}'_1 = \mathcal{A}_1$ and, for each $k \ge 1$, let $\mathcal{A}^k = \bigcup_{j=1}^k \mathcal{A}_j$ and let $\mathcal{A}'_{k+1} = \mathcal{A}_{k+1} \cap \overline{\mathcal{A}^k}$. By induction, the sets $\mathcal{A}'_1, \mathcal{A}'_2, \ldots$ are disjoint and measurable and $\mathcal{A} = \bigcup_j \mathcal{A}'_j$. Thus, from Theorem 4.9.3, $\mathcal{A}$ is measurable. Next suppose $\mathcal{D} = \cap \mathcal{A}_j$. Then $\overline{\mathcal{D}} = \cup \overline{\mathcal{A}_j}$. Thus, $\overline{\mathcal{D}} \in \mathcal{M}$, so $\mathcal{D} \in \mathcal{M}$ also. □

Proof of Theorem 4.3.1 The first two parts of Theorem 4.3.1 are Theorems 4.9.4 and 4.9.3. The third part, that $\mathcal{A}$ is measurable with zero measure if $\mu^\circ(\mathcal{A}) = 0$, follows from $T \le \mu^\circ(\mathcal{A}) + \mu^\circ(\overline{\mathcal{A}}) = \mu^\circ(\overline{\mathcal{A}})$ and $\mu^\circ(\overline{\mathcal{A}}) \le T$, i.e. that $\mu^\circ(\overline{\mathcal{A}}) = T$. □

Sets of zero measure are quite important in understanding Lebesgue integration, so it is important to know whether there are also uncountable sets of points that have zero measure. The answer is yes; a simple example follows.

Example 4.9.6 (The Cantor set) Express each point in the interval (0,1) by a ternary expansion. Let $\mathcal{B}$ be the set of points in (0,1) for which that expansion contains only 0s and 2s and is also nonterminating. Thus $\mathcal{B}$ excludes the interval $[1/3, 2/3)$, since all these expansions start with 1. Similarly, $\mathcal{B}$ excludes $[1/9, 2/9)$ and $[7/9, 8/9)$, since the second digit is 1 in these expansions. The right endpoint for each of these intervals is also excluded since it has a terminating expansion. Let $\mathcal{B}_n$ be the set of points with no 1 in the first n digits of the ternary expansion. Then $\mu(\mathcal{B}_n) = (2/3)^n$. Since $\mathcal{B}$ is contained in $\mathcal{B}_n$ for each $n \ge 1$, $\mathcal{B}$ is measurable and $\mu(\mathcal{B}) = 0$.

The expansion for each point in $\mathcal{B}$ is a binary sequence (viewing 0 and 2 as the binary digits here). There are uncountably many binary sequences (see Section 4.9.1), and this remains true when the countable number of terminating sequences are removed. Thus we have demonstrated an uncountably infinite set of numbers with zero measure.

Not all point sets are Lebesgue measurable, and an example follows.

Example 4.9.7 (A non-measurable set) Consider the interval $[0, 1)$. We define a collection of equivalence classes where two points in $[0, 1)$ are in the same equivalence class if the difference between them is rational. Thus one equivalence class consists of the rationals in [0,1). Each other equivalence class consists of a countably infinite set of irrationals whose differences are rational. This partitions $[0, 1)$ into an uncountably infinite set of equivalence classes. Now consider a set $\mathcal{A}$ that contains exactly one number chosen from each equivalence class. We will assume that $\mathcal{A}$ is measurable and show that this leads to a contradiction.

For the given set $\mathcal{A}$, let $\mathcal{A}+r$, for r rational in $(0, 1)$, denote the set that results from mapping each $t \in \mathcal{A}$ into either $t+r$ or $t+r-1$, whichever lies in $[0, 1)$. The set $\mathcal{A}+r$ is thus the set $\mathcal{A}$, shifted by r, and then rotated to lie in $[0, 1)$. By looking at outer measures, it is easy to see that $\mathcal{A}+r$ is measurable if $\mathcal{A}$ is and that both then have the same measure. Finally, each $t \in [0, 1)$ lies in exactly one equivalence class, and if τ is the element of $\mathcal{A}$ in that equivalence class, then t lies in $\mathcal{A}+r$, where $r = t-\tau$ or $t-\tau+1$. In other words, $[0, 1) = \bigcup_r(\mathcal{A}+r)$ and the sets $\mathcal{A}+r$ are disjoint. Assuming that $\mathcal{A}$ is measurable, Theorem 4.9.3 asserts that $1 = \sum_r \mu(\mathcal{A}+r)$. However, the sum on the right is 0 if $\mu(\mathcal{A}) = 0$ and infinite if $\mu(\mathcal{A}) > 0$, establishing the contradiction.

4.10 Exercises

4.1 (Fourier series)

(a) Consider the function $u(t) = \text{rect}(2t)$ of Figure 4.2. Give a general expression for the Fourier series coefficients for the Fourier series over $[-1/2, 1/2]$ and show that the series converges to $1/2$ at each of the endpoints, $-1/4$ and $1/4$. [Hint. You do not need to know anything about convergence here.]

(b) Represent the same function as a Fourier series over the interval $[-1/4, 1/4]$. What does this series converge to at $-1/4$ and $1/4$? Note from this exercise that the Fourier series depends on the interval over which it is taken.

4.2 (Energy equation) Derive (4.6), the energy equation for Fourier series. [Hint. Substitute the Fourier series for $u(t)$ into $\int u(t)u^*(t)\,dt$. Don't worry about convergence or interchange of limits here.]

4.3 (Countability) As shown in Appendix 4.9.1, many subsets of the real numbers, including the integers and the rationals, are countable. Sometimes, however, it is necessary to give up the ordinary numerical ordering in listing the elements of these subsets. This exercise shows that this is sometimes inevitable.

(a) Show that every listing of the integers (such as $0, -1, 1, -2, \ldots$) fails to preserve the numerical ordering of the integers. [Hint. Assume such a numerically ordered listing exists and show that it can have no first element (i.e., no smallest element).]

(b) Show that the rational numbers in the interval (0, 1) cannot be listed in a way that preserves their numerical ordering.

(c) Show that the rationals in [0,1] cannot be listed with a preservation of numerical ordering. (The first element is no problem, but what about the second?)

4.4 (Countable sums) Let $a_1, a_2, \ldots$ be a countable set of nonnegative numbers and assume that $s_a(k) = \sum_{j=1}^{k} a_j \leq A$ for all k and some given $A > 0$.

(a) Show that the limit $\lim_{k\to\infty} s_a(k)$ exists with some value S_a between 0 and A. (Use any level of mathematical care that you feel comfortable with.)

(b) Now let $b_1, b_2, \dots$ be another ordering of the numbers $a_1, a_2, \dots$ That is, let $b_1 = a_{j(1)}, b_2 = a_{j(2)}, \dots, b_\ell = a_{j(\ell)}, \dots,$ where $j(\ell)$ is a permutation of the positive integers, i.e. a one-to-one function from $\mathbb{Z}^+$ to $\mathbb{Z}^+$. Let $s_b(k) = \sum_{\ell=1}^{k} b_\ell$. Show that $\lim_{k\to\infty} s_b(k) \le S_a$. Note that

$$\sum_{\ell=1}^{k} b_\ell = \sum_{\ell=1}^{k} a_{j(\ell)}.$$

(c) Define $S_b = \lim_{k\to\infty} s_b(k)$ and show that $S_b \ge S_a$. [Hint. Consider the inverse permuation, say $\ell^{-1}(j)$, which for given j' is that ℓ for which $j(\ell) = j'$.] Note that you have shown that a countable sum of nonnegative elements does not depend on the order of summation.

(d) Show that the above result is not necessarily true for a countable sum of numbers that can be positive or negative. [Hint. Consider alternating series.]

4.5 (Finite unions of intervals) Let $\mathcal{E} = \bigcup_{j=1}^{\ell} I_j$ be the union of $\ell \ge 2$ arbitrary nonempty intervals. Let a_j and b_j denote the left and right endpoints, respectively, of I_j; each endpoint can be included or not. Assume the intervals are ordered so that $a_1 \le a_2 \le \cdots \le a_\ell$.

(a) For $\ell = 2$, show that either I_1 and I_2 are separated or that $\mathcal{E}$ is a single interval whose left endpoint is a_1.

(b) For $\ell > 2$ and $2 \le k < \ell$, let $\mathcal{E}^k = \bigcup_{j=1}^{k} I_j$. Give an algorithm for constructing a union of separated intervals for $\mathcal{E}^{k+1}$ given a union of separated intervals for $\mathcal{E}^k$.

(c) Note that using part (b) inductively yields a representation of $\mathcal{E}$ as a union of separated intervals. Show that the left endpoint for each separated interval is drawn from $a_1, \dots, a_\ell$ and the right endpoint is drawn from $b_1, \dots, b_\ell$.

(d) Show that this representation is unique, i.e. that $\mathcal{E}$ cannot be represented as the union of any other set of separated intervals. Note that this means that $\mu(\mathcal{E})$ is defined unambiguously in (4.9).

4.6 (Countable unions of intervals) Let $\mathcal{B} = \bigcup_j I_j$ be a countable union of arbitrary (perhaps intersecting) intervals. For each $k \ge 1$, let $\mathcal{B}^k = \bigcup_{j=1}^{k} I_j$, and for each $k \ge j$ let $I_{j,k}$ be the separated interval in $\mathcal{B}^k$ containing I_j (see Exercise 4.5).

(a) For each $k \ge j \ge 1$, show that $I_{j,k} \subseteq I_{j,k+1}$.

(b) Let $\bigcup_{k=j}^{\infty} I_{j,k} = I'_j$. Explain why I'_j is an interval and show that $I'_j \subseteq \mathcal{B}$.

(c) For any i, j, show that either $I'_j = I'_i$ or I'_j and I'_i are separated intervals.

(d) Show that the sequence $\{I'_j; 1 \le j < \infty\}$ with repetitions removed is a countable separated-interval representation of $\mathcal{B}$.

(e) Show that the collection $\{I'_j; j \ge 1\}$ with repetitions removed is unique; i.e., show that if an arbitrary interval I is contained in $\mathcal{B}$, then it is contained in one of the I'_j. Note, however, that the ordering of the I'_j is not unique.

4.7 (Union bound for intervals) Prove the validity of the union bound for a countable collection of intervals in (4.89). The following steps are suggested.

(a) Show that if $\mathcal{B} = I_1 \bigcup I_2$ for arbitrary intervals I_1, I_2, then $\mu(\mathcal{B}) \leq \mu(I_1) + \mu(I_2)$ with equality if I_1 and I_2 are disjoint. Note: this is true by definition if I_1 and I_2 are separated, so you need only treat the cases where I_1 and I_2 intersect or are disjoint but not separated.

(b) Let $\mathcal{B}^k = \bigcup_{j=1}^{k} I_j$ be represented as the union of, say, m_k separated intervals $(m_k \leq k)$, so $\mathcal{B}^k = \bigcup_{j=1}^{m_k} I_j$. Show that $\mu(\mathcal{B}^k \bigcup I_{k+1}) \leq \mu(\mathcal{B}^k) + \mu(I_{k+1})$ with equality if $\mathcal{B}^k$ and I_{k+1} are disjoint.

(c) Use finite induction to show that if $\mathcal{B} = \bigcup_{j=1}^{k} I_j$ is a finite union of arbitrary intervals, then $\mu(\mathcal{B}) \leq \sum_{j=1}^{k} \mu(I_j)$ with equality if the intervals are disjoint.

(d) Extend part (c) to a countably infinite union of intervals.

4.8 For each positive integer n, let $\mathcal{B}_n$ be a countable union of intervals. Show that $\mathcal{B} = \bigcup_{n=1}^{\infty} \mathcal{B}_n$ is also a countable union of intervals. [Hint. Look at Example 4.9.2 in Section 4.9.1.]

4.9 (Measure and covers) Let $\mathcal{A}$ be an arbitrary measurable set in $[-T/2, T/2]$ and let $\mathcal{B}$ be a cover of $\mathcal{A}$. Using only results derived prior to Lemma 4.9.3, show that $\mu^{\circ}(\mathcal{B} \cap \overline{\mathcal{A}}) = \mu(\mathcal{B}) - \mu(\mathcal{A})$. You may use the following steps if you wish.

(a) Show that $\mu^{\circ}(\mathcal{B} \cap \overline{\mathcal{A}}) \geq \mu(\mathcal{B}) - \mu(\mathcal{A})$.

(b) For any $\delta > 0$, let $\mathcal{B}'$ be a cover of $\overline{\mathcal{A}}$ with $\mu(\mathcal{B}') \leq \mu(\overline{\mathcal{A}}) + \delta$. Use Lemma 4.9.2 to show that $\mu(\mathcal{B} \cap \mathcal{B}') = \mu(\mathcal{B}) + \mu(\mathcal{B}') - T$.

(c) Show that $\mu^{\circ}(\mathcal{B} \cap \overline{\mathcal{A}}) \leq \mu(\mathcal{B} \cap \mathcal{B}') \leq \mu(\mathcal{B}) - \mu(\mathcal{A}) + \delta$.

(d) Show that $\mu^{\circ}(\mathcal{B} \cap \overline{\mathcal{A}}) = \mu(\mathcal{B}) - \mu(\mathcal{A})$.

4.10 (Intersection of covers) Let $\mathcal{A}$ be an arbitrary set in $[-T/2, T/2]$.

(a) Show that $\mathcal{A}$ has a sequence of covers, $\mathcal{B}_1, \mathcal{B}_2, \ldots$, such that $\mu^{\circ}(\mathcal{A}) = \mu(\mathcal{D})$, where $\mathcal{D} = \bigcap_n \mathcal{B}_n$.

(b) Show that $\mathcal{A} \subseteq \mathcal{D}$.

(c) Show that if $\mathcal{A}$ is measurable, then $\mu(\mathcal{D} \cap \overline{\mathcal{A}}) = 0$. Note that you have shown that an arbitrary measurable set can be represented as a countable intersection of countable unions of intervals, less a set of zero measure. Argue by example that if $\mathcal{A}$ is not measurable, then $\mu^{\circ}(\mathcal{D} \cap \overline{A})$ need not be 0.

4.11 (Measurable functions)

(a) For $\{u(t) : [-T/2, T/2] \to R\}$, show that if $\{t : u(t) < \beta\}$ is measurable, then $\{t : u(t) \geq \beta\}$ is measurable.

(b) Show that if $\{t : u(t) < \beta\}$ and $\{t : u(t) < \alpha\}$ are measurable, $\alpha < \beta$, then $\{t : \alpha \leq u(t) < \beta\}$ is measurable.

(c) Show that if $\{t : u(t) < \beta\}$ is measurable for all β, then $\{t : u(t) \leq \beta\}$ is also measurable. [Hint. Express $\{t : u(t) \leq \beta\}$ as a countable intersection of measurable sets.]

(d) Show that if $\{t : u(t) \leq \beta\}$ is measurable for all β, then $\{t : u(t) < \beta\}$ is also measurable, i.e. the definition of measurable function can use either strict or nonstrict inequality.

4.12 (Measurable functions) Assume throughout that $\{u(t) : [-T/2, T/2] \to \mathbb{R}\}$ is measurable.

(a) Show that $-u(t)$ and $|u(t)|$ are measurable.

(b) Assume that $\{g(x) : \mathbb{R} \to \mathbb{R}\}$ is an increasing function (i.e. $x_1 < x_2 \implies g(x_1) < g(x_2)$). Prove that $v(t) = g(u(t))$ is measurable. [Hint. This is a one liner. If the abstraction confuses you, first show that $\exp(u(t))$ is measurable and then prove the more general result.]

(c) Show that $\exp[u(t)]$, $u^2(t)$, and $\ln|u(t)|$ are all measurable.

4.13 (Measurable functions)

(a) Show that if $\{u(t) : [-T/2, T/2] \to \mathbb{R}\}$ and $\{v(t) : [-T/2, T/2] \to \mathbb{R}\}$ are measurable, then $u(t) + v(t)$ is also measurable. [Hint. Use a discrete approximation to the sum and then go to the limit.]

(b) Show that $u(t)v(t)$ is also measurable.

4.14 (Measurable sets) Suppose $\mathcal{A}$ is a subset of $[-T/2,\ T/2]$ and is measurable over $[-T/2,\ T/2]$. Show that $\mathcal{A}$ is also measurable, with the same measure, over $[-T'/2,\ T'/2]$ for any T' satisfying $T' > T$. [Hint. Let $\mu'(\mathcal{A})$ be the outer measure of $\mathcal{A}$ over $[-T'/2,\ T'/2]$ and show that $\mu'(\mathcal{A}) = \mu^{\circ}(\mathcal{A})$, where μ° is the outer measure over $[-T/2,\ T/2]$. Then let $\overline{\mathcal{A}}'$ be the complement of $\mathcal{A}$ over $[-T'/2,\ T'/2]$ and show that $\mu'(\overline{\mathcal{A}}') = \mu^{\circ}(\overline{\mathcal{A}}) + T' - T$.]

4.15 (Measurable limits)

(a) Assume that $\{u_n(t) : [-T/2,\ T/2] \to \mathbb{R}\}$ is measurable for each $n \geq 1$. Show that $\liminf_n u_n(t)$ is measurable ($\liminf_n u_n(t)$ means $\lim_m v_m(t)$, where $v_m(t) = \inf_{n=m}^{\infty} u_n(t)$ and infinite values are allowed).

(b) Show that $\lim_n u_n(t)$ exists for a given t if and only if $\liminf_n u_n(t) = \limsup_n u_n(t)$.

(c) Show that the set of t for which $\lim_n u_n(t)$ exists is measurable. Show that a function $u(t)$ that is $\lim_n u_n(t)$ when the limit exists and is 0 otherwise is measurable.

4.16 (Lebesgue integration) For each integer $n \geq 1$, define $u_n(t) = 2^n \operatorname{rect}(2^n t - 1)$. Sketch the first few of these waveforms. Show that $\lim_{n\to\infty} u_n(t) = 0$ for all t. Show that $\int \lim_n u_n(t)\,dt \neq \lim_n \int u_n(t)dt$.

4.17 ($\mathcal{L}_1$ integrals)

(a) Assume that $\{u(t) : [-T/2,\ T/2] \to \mathbb{R}\}$ is $\mathcal{L}_1$. Show that

$$\left|\int u(t)\,dt\right| = \left|\int u^+(t)dt - \int u^-(t)dt\right| \leq \int |u(t)|\,dt.$$

(b) Assume that $\{u(t) : [-T/2,\ T/2] \to \mathbb{C}\}$ is $\mathcal{L}_1$. Show that

$$\left|\int u(t)dt\right| \leq \int |u(t)|\,dt.$$

[Hint. Choose α such that $\alpha \int u(t)\,dt$ is real and nonnegative and $|\alpha| = 1$. Use part (a) on $\alpha u(t)$.]

4.18 ($\mathcal{L}_2$-equivalence) Assume that $\{u(t) : [-T/2, T/2] \to \mathbb{C}\}$ and $\{v(t) : [-T/2, T/2] \to \mathbb{C}\}$ are $\mathcal{L}_2$ functions.

(a) Show that if $u(t)$ and $v(t)$ are equal a.e., then they are $\mathcal{L}_2$-equivalent.

(b) Show that if $u(t)$ and $v(t)$ are $\mathcal{L}_2$-equivalent, then for any $\varepsilon > 0$ the set $\{t : |u(t) - v(t)|^2 \geq \varepsilon\}$ has zero measure.

(c) Using (b), show that $\mu\{t : |u(t) - v(t)| > 0\} = 0$, i.e. that $u(t) = v(t)$ a.e.

4.19 (Orthogonal expansions) Assume that $\{u(t) : \mathbb{R} \to \mathbb{C}\}$ is $\mathcal{L}_2$. Let $\{\theta_k(t);\ 1 \leq k < \infty\}$ be a set of orthogonal waveforms and assume that $u(t)$ has the following orthogonal expansion:

$$u(t) = \sum_{k=1}^{\infty} u_k \theta_k(t).$$

Assume the set of orthogonal waveforms satisfy

$$\int_{-\infty}^{\infty} \theta_k(t)\theta_j^*(t)\mathrm{d}t = \begin{cases} 0 & \text{for } k \neq j; \\ A_j & \text{for } k = j, \end{cases}$$

where $\{A_j;\ j \in \mathbb{Z}^+\}$ is an arbitrary set of positive numbers. Do not concern yourself with convergence issues in this exercise.

(a) Show that each u_k can be expressed in terms of $\int_{-\infty}^{\infty} u(t)\theta_k^*(t)\mathrm{d}t$ and A_k.

(b) Find the energy $\int_{-\infty}^{\infty} |u(t)|^2\, \mathrm{d}t$ in terms of $\{u_k\}$ and $\{A_k\}$.

(c) Suppose that $v(t) = \sum_k v_k \theta_k(t)$, where $v(t)$ also has finite energy. Express $\int_{-\infty}^{\infty} u(t)v^*(t)\,\mathrm{d}t$ as a function of $\{u_k, v_k, A_k;\ k \in \mathbb{Z}\}$.

4.20 (Fourier series)

(a) Verify that (4.22) and (4.23) follow from (4.20) and (4.18) using the transformation $u(t) = v(t + \Delta)$.

(b) Consider the Fourier series in periodic form, $w(t) = \sum_k \hat{w}_k \mathrm{e}^{2\pi ikt/T}$, where $\hat{w}_k = (1/T)\int_{-T/2}^{T/2} w(t)\mathrm{e}^{-2\pi ikt/T}\,\mathrm{d}t$. Show that for any real Δ, $(1/T)\int_{-T/2+\Delta}^{T/2+\Delta} w(t)\mathrm{e}^{-2\pi ikt/T}\,\mathrm{d}t$ is also equal to $\hat{w}_k$, providing an alternative derivation of (4.22) and (4.23).

4.21 Equation (4.27) claims that

$$\lim_{n\to\infty, \ell\to\infty} \int \left| u(t) - \sum_{m=-n}^{n} \sum_{k=-\ell}^{\ell} \hat{u}_{k,m}\theta_{k,m}(t) \right|^2 \mathrm{d}t = 0.$$

(a) Show that the integral above is nonincreasing in both ℓ and n.

(b) Show that the limit is independent of how n and ℓ approach ∞. [Hint. See Exercise 4.4.]

(c) More generally, show that the limit is the same if the pair (k, m), $k \in \mathbb{Z}$, $m \in \mathbb{Z}$, is ordered in an arbitrary way and the limit above is replaced by a limit on the partial sums according to that ordering.

4.22 (Truncated sinusoids)

(a) Verify (4.24) for $\mathcal{L}_2$ waveforms, i.e. show that

$$\lim_{n\to\infty}\int \left|u(t)-\sum_{m=-n}^{n} u_m(t)\right|^2 \mathrm{d}t = 0.$$

(b) Break the integral in (4.28) into separate integrals for $|t| > (n+1/2)T$ and $|t| \le (n+1/2)T$. Show that the first integral goes to 0 with increasing n.

(c) For given n, show that the second integral above goes to 0 with increasing ℓ.

4.23 (Convolution) The left side of (4.40) is a function of t. Express the Fourier transform of this as a double integral over t and τ. For each t, make the substitution $r = t - \tau$ and integrate over r. Then integrate over τ to get the right side of (4.40). Do not concern yourself with convergence issues here.

4.24 (Continuity of $\mathcal{L}_1$ transform) Assume that $\{u(t) : \mathbb{R} \to \mathbb{C}\}$ is $\mathcal{L}_1$ and let $\hat{u}(f)$ be its Fourier transform. Let ε be any given positive number.

(a) Show that for sufficiently large T, $\int_{|t|>T} |u(t)\mathrm{e}^{-2\pi \mathrm{i}ft} - u(t)\mathrm{e}^{-2\pi\mathrm{i}(f-\delta)t}|\,\mathrm{d}t < \varepsilon/2$ for all f and all $\delta > 0$.

(b) For the ε and T selected above, show that $\int_{|t|\le T} |u(t)\mathrm{e}^{-2\pi\mathrm{i}ft} - u(t)\mathrm{e}^{-2\pi\mathrm{i}(f-\delta)t}|\,\mathrm{d}t < \varepsilon/2$ for all f and sufficiently small $\delta > 0$. This shows that $\hat{u}(f)$ is continuous.

4.25 (Plancherel) - The purpose of this exercise is to get some understanding of the Plancherel theorem. Assume that $u(t)$ is $\mathcal{L}_2$ and has a Fourier transform $\hat{u}(f)$.

(a) Show that $\hat{u}(f) - \hat{u}_A(f)$ is the Fourier transform of the function $x_A(t)$ that is 0 from $-A$ to A and equal to $u(t)$ elsewhere.

(b) Argue that since $\int_{-\infty}^{\infty} |u(t)|^2\,\mathrm{d}t$ is finite, the integral $\int_{-\infty}^{\infty} |x_A(t)|^2\,\mathrm{d}t$ must go to 0 as $A \to \infty$. Use whatever level of mathematical care and common sense that you feel comfortable with.

(c) Using the energy equation (4.45), argue that

$$\lim_{A\to\infty}\int_{-\infty}^{\infty} |\hat{u}(f) - \hat{u}_A(f)|^2\,\mathrm{d}t = 0.$$

Note: this is only the easy part of the Plancherel theorem. The difficult part is to show the existence of $\hat{u}(f)$. The limit as $A \to \infty$ of the integral $\int_{-A}^{A} u(t)\mathrm{e}^{-2\pi\mathrm{i}ft}\,\mathrm{d}t$ need not exist for all f, and the point of the Plancherel theorem is to forget about this limit for individual f and focus instead on the energy in the difference between the hypothesized $\hat{u}(f)$ and the approximations.

4.26 (Fourier transform for $\mathcal{L}_2$) Assume that $\{u(t) : \mathbb{R} \to \mathbb{C}\}$ and $\{v(t) : \mathbb{R} \to \mathbb{C}\}$ are $\mathcal{L}_2$ and that a and b are complex numbers. Show that $au(t) + bv(t)$ is $\mathcal{L}_2$. For $T > 0$, show that $u(t-T)$ and $u(t/T)$ are $\mathcal{L}_2$ functions.

4.27 (Relation of Fourier series to Fourier integral) Assume that $\{u(t) : [-T/2, T/2] \to \mathbb{C}\}$ is $\mathcal{L}_2$. Without being very careful about the mathematics, the Fourier series expansion of $\{u(t)\}$ is given by

$$u(t) = \lim_{\ell\to\infty} u^{(\ell)}(t), \quad \text{where} \quad u^{(\ell)}(t) = \sum_{k=-\ell}^{\ell} \hat{u}_k e^{2\pi ikt/T} \mathrm{rect}\left(\frac{t}{T}\right);$$

$$\hat{u}_k = \frac{1}{T}\int_{-T/2}^{T/2} u(t)e^{-2\pi ikt/T}\, dt.$$

(a) Does the above limit hold for all $t \in [-T/2, T/2]$? If not, what can you say about the type of convergence?

(b) Does the Fourier transform $\hat{u}(f) = \int_{-T/2}^{T/2} u(t)e^{-2\pi ift}\, dt$ exist for all f? Explain.

(c) The Fourier transform of the finite sum $u^{(\ell)}(t)$ is $\hat{u}^{(\ell)}(f) = \sum_{k=-\ell}^{\ell} \hat{u}_k T\, \mathrm{sinc}(fT - k)$. In the limit $\ell \to \infty$, $\hat{u}(f) = \lim_{\ell\to\infty} \hat{u}^{(\ell)}(f)$, so

$$\hat{u}(f) = \lim_{\ell\to\infty} \sum_{k=-\ell}^{\ell} \hat{u}_k T\, \mathrm{sinc}(fT - k).$$

Give a brief explanation why this equation must hold with equality for all $f \in \mathbb{R}$. Also show that $\{\hat{u}(f) : f \in \mathbb{R}\}$ is completely specified by its values, $\{\hat{u}(k/T) : k \in \mathbb{Z}\}$ at multiples of $1/T$.

4.28 (Sampling) One often approximates the value of an integral by a discrete sum, i.e.

$$\int_{-\infty}^{\infty} g(t)dt \approx \delta \sum_k g(k\delta).$$

(a) Show that if $u(t)$ is a real finite-energy function, lowpass-limited to W Hz, then the above approximation is exact for $g(t) = u^2(t)$ if $\delta \leq 1/2\mathrm{W}$; i.e., show that

$$\int_{-\infty}^{\infty} u^2(t)dt = \delta \sum_k u^2(k\delta).$$

(b) Show that if $g(t)$ is a real finite-energy function, lowpass-limited to W Hz, then for $\delta \leq 1/2\mathrm{W}$,

$$\int_{-\infty}^{\infty} g(t)dt = \delta \sum_k g(k\delta).$$

(c) Show that if $\delta > 1/2\mathrm{W}$, then there exists no such relation in general.

4.29 (Degrees of freedom) This exercise explores how much of the energy of a baseband-limited function $\{u(t) : [-1/2, 1/2] \to \mathbb{R}\}$ can reside outside the region where the sampling coefficients are nonzero. Let $T = 1/2\mathrm{W} = 1$, and let n be a positive even integer. Let $u_k = (-1)^k$ for $-n \leq k \leq n$ and $u_k = 0$ for $|k| > n$. Show that $|u(n + 1/2)|$ increases without bound as the endpoint n is increased. Show that $|u(n+m+1/2)| > |u(n-m-1/2)|$ for all integers m, $0 \leq m < n$. In other words, shifting the sample points by 1/2 leads to most of the sample point energy being outside the interval $[-n, n]$.

4.30 (Sampling theorem for $[\Delta - \mathrm{W}, \Delta + \mathrm{W})]$)

(a) Verify the Fourier transform pair in (4.70). [Hint. Use the scaling and shifting rules on $\mathrm{rect}(f) \leftrightarrow \mathrm{sinc}(t)$.]

(b) Show that the functions making up that expansion are orthogonal. [Hint. Show that the corresponding Fourier transforms are orthogonal.]

(c) Show that the functions in (4.74) are orthogonal.

4.31 (Amplitude-limited functions) Sometimes it is important to generate baseband waveforms with bounded amplitude. This problem explores pulse shapes that can accomplish this

(a) Find the Fourier transform of $g(t) = \mathrm{sinc}^2(\mathrm{W}t)$. Show that $g(t)$ is bandlimited to $f \le \mathrm{W}$ and sketch both $g(t)$ and $\hat{g}(f)$. [Hint. Recall that multiplication in the time domain corresponds to convolution in the frequency domain.]

(b) Let $u(t)$ be a continuous real $\mathcal{L}_2$ function baseband-limited to $f \le \mathrm{W}$ (i.e. a function such that $u(t) = \sum_k u(kT)\,\mathrm{sinc}(t/T - k)$, where $T = 1/2\mathrm{W}$. Let $v(t) = u(t) * g(t)$. Express $v(t)$ in terms of the samples $\{u(kT); k \in \mathbb{Z}\}$ of $u(t)$ and the shifts $\{g(t - kT); k \in \mathbb{Z}\}$ of $g(t)$. [Hint. Use your sketches in part (a) to evaluate $g(t) * \mathrm{sinc}(t/T)$.]

(c) Show that if the T-spaced samples of $u(t)$ are nonnegative, then $v(t) \ge 0$ for all t.

(d) Explain why $\sum_k \mathrm{sinc}(t/T - k) = 1$ for all t.

(e) Using (d), show that $\sum_k g(t - kT) = c$ for all t and find the constant c. [Hint. Use the hint in (b) again.]

(f) Now assume that $u(t)$, as defined in part (b), also satisfies $u(kT) \le 1$ for all $k \in \mathbb{Z}$. Show that $v(t) \le 2$ for all t.

(g) Allow $u(t)$ to be complex now, with $|u(kT)| \le 1$. Show that $|v(t)| \le 2$ for all t.

4.32 (Orthogonal sets) The function $\mathrm{rect}(t/T)$ has the very special property that it, plus its time and frequency shifts, by kT and j/T, respectively, form an orthogonal set. The function $\mathrm{sinc}(t/T)$ has this same property. We explore other functions that are generalizations of $\mathrm{rect}(t/T)$ and which, as you will show in parts (a)–(d), have this same interesting property. For simplicity, choose $T = 1$.

These functions take only the values 0 and 1 and are allowed to be nonzero only over $[-1, 1]$ rather than $[-1/2, 1/2]$ as with $\mathrm{rect}(t)$. Explicitly, the functions considered here satisfy the following constraints:

$$p(t) = p^2(t) \qquad \text{for all } t \quad \text{(0/1 property)}; \qquad (4.108)$$

$$p(t) = 0 \qquad \text{for } |t| > 1; \qquad (4.109)$$

$$p(t) = p(-t) \qquad \text{for all } t \quad \text{(symmetry)}; \qquad (4.110)$$

$$p(t) = 1 - p(t-1) \qquad \text{for } 0 \le t < 1/2. \qquad (4.111)$$

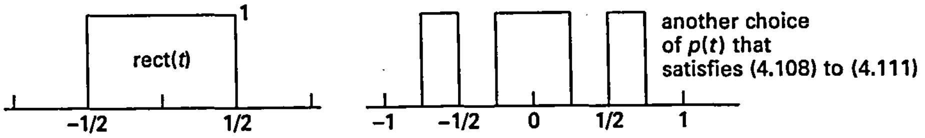


Figure 4.12. Two functions that each satisfy (4.108)–(4.111)

Note: because of property (4.110), condition (4.111) also holds for $1/2 < t \leq 1$. Note also that $p(t)$ at the single points $t = \pm 1/2$ does not affect any orthogonality properties, so you are free to ignore these points in your arguments. Figure 4.12 illustrates two examples of functions satisfying (4.108)–(4.111).

(a) Show that $p(t)$ is orthogonal to $p(t-1)$. [Hint. Evaluate $p(t)p(t-1)$ for each $t \in [0, 1]$ other than $t = 1/2$.]
(b) Show that $p(t)$ is orthogonal to $p(t-k)$ for all integer $k \neq 0$.
(c) Show that $p(t)$ is orthogonal to $p(t-k)e^{i2\pi mt}$ for integer $m \neq 0$ and $k \neq 0$.
(d) Show that $p(t)$ is orthogonal to $p(t)e^{2\pi imt}$ for integer $m \neq 0$. [Hint. Evaluate $p(t)e^{-2\pi imt} + p(t-1)e^{-2\pi im(t-1)}$.]
(e) Let $h(t) = \hat{p}(t)$ where $\hat{p}(f)$ is the Fourier transform of $p(t)$. If $p(t)$ satisfies properties (4.108)–(4.111), does it follow that $h(t)$ has the property that it is orthogonal to $h(t-k)e^{2\pi imt}$ whenever either the integer k or m is nonzero?

Note: almost no calculation is required in this problem.

4.33 (Limits) Construct an example of a sequence of $\mathcal{L}_2$ functions $v^{(m)}(t)$, $m \in \mathbb{Z}$, $m > 0$, such that $\lim_{m\to\infty} v^{(m)}(t) = 0$ for all t but for which $\text{l.i.m.}_{m\to\infty} v^{(m)}(t)$ does not exist. In other words show that pointwise convergence does not imply $\mathcal{L}_2$-convergence. [Hint. Consider time shifts.]

4.34 (Aliasing) Find an example where $\hat{u}(f)$ is 0 for $|f| > 3\mathrm{W}$ and nonzero for $\mathrm{W} < |f| < 3\mathrm{W}$, but where, $s(kT) = v_0(kT)$ for all $k \in \mathbb{Z}$. Here $v_0(kT)$ is defined in (4.77) and $T = 1/2\mathrm{W}$. [Hint. Note that it is equivalent to achieve equality between $\hat{s}(f)$ and $\hat{u}(f)$ for $|f| \leq \mathrm{W}$. Look at Figure 4.10.]

4.35 (Aliasing) The following exercise is designed to illustrate the sampling of an approximately baseband waveform. To avoid messy computation, we look at a waveform baseband-limited to 3/2 which is sampled at rate 1 (i.e. sampled at only 1/3 the rate that it should be sampled at). In particular, let $u(t) = \text{sinc}(3t)$.

(a) Sketch $\hat{u}(f)$. Sketch the function $\hat{v}_m(f) = \text{rect}(f - m)$ for each integer m such that $v_m(f) \neq 0$. Note that $\hat{u}(f) = \sum_m \hat{v}_m(f)$.
(b) Sketch the inverse transforms $v_m(t)$ (real and imaginary parts if complex).
(c) Verify directly from the equations that $u(t) = \sum v_m(t)$. [Hint. This is easier if you express the sine part of the sinc function as a sum of complex exponentials.]
(d) Verify the sinc-weighted sinusoid expansion, (4.73). (There are only three nonzero terms in the expansion.)
(e) For the approximation $s(t) = u(0)\,\text{sinc}(t)$, find the energy in the difference between $u(t)$ and $s(t)$ and interpret the terms.

4.36 (Aliasing) Let $u(t)$ be the inverse Fourier transform of a function $\hat{u}(f)$ which is both $\mathcal{L}_1$ and $\mathcal{L}_2$. Let $v_m(t) = \int \hat{u}(f)\ \text{rect}(fT-m)e^{2\pi i f t}\,df$ and let $v^{(n)}(t) = \sum_{-n}^{n} v_m(t)$.

(a) Show that $|u(t) - v^{(n)}(t)| \le \int_{|f|\ge(2n+1)/T} |\hat{u}(f)|\,df$ and thus that $u(t) = \lim_{n\to\infty} v^{(n)}(t)$ for all t.

(b) Show that the sinc-weighted sinusoid expansion of (4.76) then converges pointwise for all t. [Hint. For any t and any $\varepsilon > 0$, choose n so that $|u(t) - v^{n}(t)| \le \varepsilon/2$. Then for each m, $|m| \le n$, expand $v_m(t)$ in a sampling expansion using enough terms to keep the error less than $\varepsilon/4n+2$.]

4.37 (Aliasing)

(a) Show that $\hat{s}(f)$ in (4.83) is $\mathcal{L}_1$ if $\hat{u}(f)$ is.

(b) Let $\hat{u}(f) = \sum_{k\neq 0} \text{rect}[k^2(f-k)]$. Show that $\hat{u}(f)$ is $\mathcal{L}_1$ and $\mathcal{L}_2$. Let $T = 1$ for $\hat{s}(f)$ and show that $\hat{s}(f)$ is not $\mathcal{L}_2$. [Hint. Sketch $\hat{u}(f)$ and $\hat{s}(f)$.]

(c) Show that $\hat{u}(f)$ does not satisfy $\lim_{|f|\to\infty} \hat{u}(f)|f|^{1+\varepsilon} = 0$.

4.38 (Aliasing) Let $u(t) = \sum_{k\neq 0} \text{rect}[k^2(t-k)]$ and show that $u(t)$ is $\mathcal{L}_2$. Find $s(t) = \sum_k u(k)\ \text{sinc}(t-k)$ and show that it is neither $\mathcal{L}_1$ nor $\mathcal{L}_2$. Find $\sum_k u^2(k)$ and explain why the sampling theorem energy equation (4.66) does not apply here.

6 Channels, modulation, and demodulation

6.1 Introduction

Digital modulation (or channel encoding) is the process of converting an input sequence of bits into a waveform suitable for transmission over a communication channel. Demodulation (channel decoding) is the corresponding process at the receiver of converting the received waveform into a (perhaps noisy) replica of the input bit sequence. Chapter 1 discussed the reasons for using a bit sequence as the interface between an arbitrary source and an arbitrary channel, and Chapters 2 and 3 discussed how to encode the source output into a bit sequence.

Chapters 4 and 5 developed the signal-space view of waveforms. As explained in those chapters, the source and channel waveforms of interest can be represented as real or complex[1] $\mathcal{L}_2$ vectors. Any such vector can be viewed as a conventional function of time, $x(t)$. Given an orthonormal basis $\{\phi_1(t), \phi_2(t), \dots\}$ of $\mathcal{L}_2$, any such $x(t)$ can be represented as

$$x(t) = \sum_j x_j \phi_j(t). \tag{6.1}$$

Each x_j in (6.1) can be uniquely calculated from $x(t)$, and the above series converges in $\mathcal{L}_2$ to $x(t)$. Moreover, starting from any sequence satisfying $\sum_j |x_j|^2 < \infty$, there is an $\mathcal{L}_2$ function $x(t)$ satisfying (6.1) with $\mathcal{L}_2$-convergence. This provides a simple and generic way of going back and forth between functions of time and sequences of numbers. The basic parts of a modulator will then turn out to be a procedure for mapping a sequence of binary digits into a sequence of real or complex numbers, followed by the above approach for mapping a sequence of numbers into a waveform.

[1] As explained later, the actual transmitted waveforms are real. However, they are usually bandpass real waveforms that are conveniently represented as complex baseband waveforms.

In most cases of modulation, the set of waveforms $\phi_1(t), \phi_2(t), \ldots$ in (6.1) will be chosen not as a basis for $\mathcal{L}_2$ but as a basis for some subspace[2] of $\mathcal{L}_2$ such as the set of functions that are baseband-limited to some frequency W_b or passband-limited to some range of frequencies. In some cases, it will also be desirable to use a sequence of waveforms that are not orthonormal.

We can view the mapping from bits to numerical signals and the conversion of signals to a waveform as separate layers. The demodulator then maps the received waveform to a sequence of received signals, which is then mapped to a bit sequence, hopefully equal to the input bit sequence. A major objective in designing the modulator and demodulator is to maximize the rate at which bits enter the encoder, subject to the need to retrieve the original bit stream with a suitably small error rate. Usually this must be done subject to constraints on the transmitted power and bandwidth. In practice there are also constraints on delay, complexity, compatibility with standards, etc., but these need not be a major focus here.

Example 6.1.1 As a particularly simple example, suppose a sequence of binary symbols enters the encoder at T-spaced instants of time. These symbols can be mapped into real numbers using the mapping $0 \to +1$ and $1 \to -1$. The resulting sequence $u_1, u_2, \ldots$ of real numbers is then mapped into a transmitted waveform given by

$$u(t) = \sum_k u_k \operatorname{sinc}\left(\frac{t}{T} - k\right). \tag{6.2}$$

This is baseband-limited to $\mathsf{W}_b = 1/2T$. At the receiver, in the absence of noise, attenuation, and other imperfections, the received waveform is $u(t)$. This can be sampled at times $T, 2T, \ldots$ to retrieve $u_1, u_2, \ldots,$ which can be decoded into the original binary symbols.

The above example contains rudimentary forms of the two layers discussed above. The first is the mapping of binary symbols into numerical signals[3] and the second is the conversion of the sequence of signals into a waveform. In general, the set of T-spaced sinc functions in (6.2) can be replaced by any other set of orthogonal functions (or even nonorthogonal functions). Also, the mapping $0 \to +1$, $1 \to -1$ can be generalized by segmenting the binary stream into b-tuples of binary symbols, which can then be mapped into n-tuples of real or complex numbers. The set of 2^b possible n-tuples resulting from this mapping is called a *signal constellation.*

[2] Equivalently, $\phi_1(t), \phi_2(t), \ldots$ can be chosen as a basis of $\mathcal{L}_2$, but the set of indices for which x_j is allowed to be nonzero can be restricted.

[3] The word *signal* is often used in the communication literature to refer to symbols, vectors, waveforms, or almost anything else. Here we use it only to refer to real or complex numbers (or n-tuples of numbers) in situations where the numerical properties are important. For example, in (6.2) the *signals* (numerical values) $u_1, u_2, \ldots$ determine the real-valued waveform $u(t)$, whereas the binary input *symbols* could be 'Alice' and 'Bob' as easily as 0 and 1.

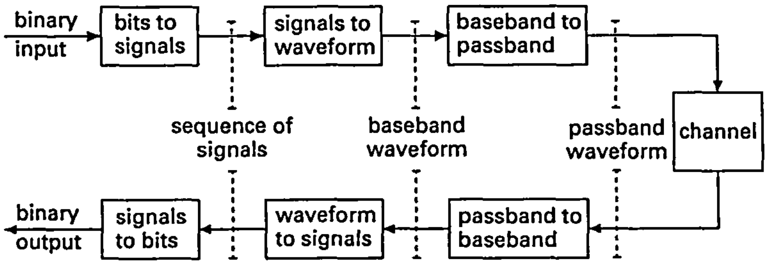


Figure 6.1. Layers of a modulator (channel encoder) and demodulator (channel decoder).

Modulators usually include a third layer, which maps a baseband-encoded waveform, such as $u(t)$ in (6.2), into a passband waveform $x(t) = \Re\{u(t)e^{2\pi i f_c t}\}$ centered on a given carrier frequency f_c. At the decoder, this passband waveform is mapped back to baseband before the other components of decoding are performed. This frequency conversion operation at encoder and decoder is often referred to as modulation and demodulation, but it is more common today to use the word modulation for the entire process of mapping bits to waveforms. Figure 6.1 illustrates these three layers.

We have illustrated the channel as a one-way device going from source to destination. Usually, however, communication goes both ways, so that a physical location can send data to another location and also receive data from that remote location. A physical device that both encodes data going out over a channel and also decodes oppositely directed data coming in from the channel is called a *modem* (for modulator/demodulator). As described in Chapter 1, feedback on the reverse channel can be used to request retransmissions on the forward channel, but in practice this is usually done as part of an automatic retransmission request (ARQ) strategy in the data link control layer. Combining coding with more sophisticated feedback strategies than ARQ has always been an active area of communication and information-theoretic research, but it will not be discussed here for the following reasons:

- it is important to understand communication in a single direction before addressing the complexities of two directions;
- feedback does not increase channel capacity for typical channels (see Shannon (1956));
- simple error detection and retransmission is best viewed as a topic in data networks.

There is an interesting analogy between analog source coding and digital modulation. With analog source coding, an analog waveform is first mapped into a sequence of real or complex numbers (e.g. the coefficients in an orthogonal expansion). This sequence of signals is then quantized into a sequence of symbols from a discrete alphabet, and finally the symbols are encoded into a binary sequence. With modulation, a sequence of bits is encoded into a sequence of signals from a signal constellation. The elements of this constellation are real or complex points in one or several dimensions. This sequence of signal points is then mapped into a waveform by the inverse of the process for converting waveforms into sequences.

6.2 Pulse amplitude modulation (PAM)

Pulse amplitude modulation[4] (PAM) is probably the simplest type of modulation. The incoming binary symbols are first segmented into b-bit blocks. There is a mapping from the set of $M = 2^b$ possible blocks into a signal constellation $\mathcal{A} = \{a_1, a_2, \ldots, a_M\}$ of real numbers. Let R be the rate of incoming binary symbols in bits per second. Then the sequence of b-bit blocks, and the corresponding sequence $u_1, u_2, \ldots$ of M-ary signals, has a rate of $R_s = R/b$ signals/s. The sequence of signals is then mapped into a waveform $u(t)$ by the use of time shifts of a basic pulse waveform $p(t)$, i.e.

$$u(t) = \sum_k u_k p(t - kT), \tag{6.3}$$

where $T = 1/R_s$ is the interval between successive signals. The special case where $b = 1$ is called *binary* PAM and the case $b > 1$ is called *multilevel* PAM. Example 6.1.1 is an example of binary PAM where the basic pulse shape $p(t)$ is a sinc function. Comparing (6.1) with (6.3), we see that PAM is a special case of digital modulation in which the underlying set of functions $\phi_1(t), \phi_2(t), \ldots$ is replaced by functions that are T-spaced time shifts of a basic function $p(t)$.

In what follows, signal constellations (i.e. the outer layer in Figure 6.1) are discussed in Sections 6.2.1 and 6.2.2. Pulse waveforms (i.e. the middle layer in Figure 6.1) are then discussed in Sections 6.2.3 and 6.2.4. In most cases,[5] the pulse waveform $p(t)$ is a baseband waveform, and the resulting modulated waveform $u(t)$ is then modulated up to some passband (i.e. the inner layer in Figure 6.1). Section 6.4 discusses modulation from baseband to passband and back.

6.2.1 Signal constellations

A *standard* M-PAM signal constellation $\mathcal{A}$ (see Figure 6.2) consists of $M = 2^b$ d-spaced real numbers located symmetrically about the origin; i.e.,

$$\mathcal{A} = \left\{ \frac{-d(M-1)}{2}, \ldots, \frac{-d}{2}, \frac{d}{2}, \ldots, \frac{d(M-1)}{2} \right\}.$$

In other words, the signal points are the same as the representation points of a symmetric M-point uniform scalar quantizer.

If the incoming bits are independent equiprobable random symbols (which is a good approximation with effective source coding), then each signal u_k is a sample value of a random variable U_k that is equiprobable over the constellation (alphabet) $\mathcal{A}$. Also the

[4] This terminology comes from analog amplitude modulation, where a baseband waveform is modulated up to some passband for communication. For digital communication, the more interesting problem is turning a bit stream into a waveform at baseband.

[5] Ultra-wide-band modulation (UWB) is an interesting modulation technique where the transmitted waveform is essentially a baseband PAM system over a "baseband" of multiple gigahertz. This is discussed briefly in Chapter 9.

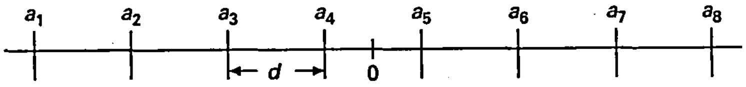


Figure 6.2. An 8-PAM signal set.

sequence $U_1, U_2, \ldots$ is independent and identically distributed (iid). As derived in Exercise 6.1, the mean squared signal value, or "energy per signal," $E_s = \mathsf{E}[U_k^2]$, is then given by

$$E_s = \frac{d^2(M^2-1)}{12} = \frac{d^2(2^{2b}-1)}{12}. \tag{6.4}$$

For example, for $M = 2, 4$, and 8, we have $E_s = d^2/4$, $5d^2/4$, and $21d^2/4$, respectively.

For $b > 2$, $2^{2b} - 1$ is approximately 2^{2b}, so we see that each unit increase in b increases E_s by a factor of 4. Thus, increasing the rate R by increasing b requires impractically large energy for large b.

Before explaining why standard M-PAM is a good choice for PAM and what factors affect the choice of constellation size M and distance d, a brief introduction to channel imperfections is required.

6.2.2 Channel imperfections: a preliminary view

Physical waveform channels are always subject to propagation delay, attenuation, and noise. Many wireline channels can be reasonably modeled using only these degradations, whereas wireless channels are subject to other degradations discussed in Chapter 9. This section provides a preliminary look at delay, then attenuation, and finally noise.

The time reference at a communication receiver is conventionally delayed relative to that at the transmitter. If a waveform $u(t)$ is transmitted, the received waveform (in the absence of other distortion) is $u(t-\tau)$, where τ is the delay due to propagation and filtering. The receiver clock (as a result of tracking the transmitter's timing) is ideally delayed by τ, so that the received waveform, according to the receiver clock, is $u(t)$. With this convention, the channel can be modeled as having no delay, and all equations are greatly simplified. This explains why communication engineers often model filters in the modulator and demodulator as being noncausal, since responses before time 0 can be added to the difference between the two clocks. *Estimating* the above fixed delay at the receiver is a significant problem called timing recovery, but is largely separable from the problem of recovering the transmitted data.

The magnitude of delay in a communication system is often important. It is one of the parameters included in the *quality of service* of a communication system. Delay is important for voice communication and often critically important when the communication is in the feedback loop of a real-time control system. In addition to the fixed delay in time reference between modulator and demodulator, there is also delay in source encoding and decoding. Coding for error correction adds additional delay, which might or might not be counted as part of the modulator/demodulator delay.

Either way, the delays in the source coding and error-correction coding are often much larger than that in the modulator/demodulator proper. Thus this latter delay can be significant, but is usually not of primary significance. Also, as channel speeds increase, the filtering delays in the modulator/demodulator become even less significant.

Amplitudes are usually measured on a different scale at transmitter and receiver. The actual power attenuation suffered in transmission is a product of amplifier gain, antenna coupling losses, antenna directional gain, propagation losses, etc. The process of finding all these gains and losses (and perhaps changing them) is called "the link budget." Such gains and losses are invariably calculated in decibels (dB). Recall that the number of decibels corresponding to a power gain α is defined to be $10\log_{10}\alpha$. The use of a logarithmic measure of gain allows the various components of gain to be added rather than multiplied.

The link budget in a communication system is largely separable from other issues, so the amplitude scale at the transmitter is usually normalized to that at the receiver.

By treating attenuation and delay as issues largely separable from modulation, we obtain a model of the channel in which a baseband waveform $u(t)$ is converted to passband and transmitted. At the receiver, after conversion back to baseband, a waveform $v(t) = u(t) + z(t)$ is received, where $z(t)$ is noise. This noise is a fundamental limitation to communication and arises from a variety of causes, including thermal effects and unwanted radiation impinging on the receiver. Chapter 7 is largely devoted to understanding noise waveforms by modeling them as sample values of random processes. Chapter 8 then explains how best to decode signals in the presence of noise. These issues are briefly summarized here to see how they affect the choice of signal constellation.

For reasons to be described shortly, the basic pulse waveform $p(t)$ used in PAM often has the property that it is orthonormal to all its shifts by multiples of T. In this case, the transmitted waveform $u(t) = \sum_k u_k p(t - k/T)$ is an orthonormal expansion, and, in the absence of noise, the transmitted signals $u_1, u_2, \ldots$ can be retrieved from the baseband waveform $u(t)$ by the inner product operation

$$u_k = \int u(t)p(t - kT)\mathrm{d}t.$$

In the presence of noise, this same operation can be performed, yielding

$$v_k = \int v(t)p(t - kT)\mathrm{d}t = u_k + z_k, \tag{6.5}$$

where $z_k = \int z(t)p(t - kT)\mathrm{d}t$ is the projection of $z(t)$ onto the shifted pulse $p(t - kT)$.

The most common (and often the most appropriate) model for noise on channels is called the additive white Gaussian noise model. As shown in Chapters 7 and 8, the above coefficients $\{z_k;\ k \in \mathbb{Z}\}$ in this model are the sample values of zero-mean, iid Gaussian random variables $\{Z_k;\ k \in \mathbb{Z}\}$. This is true no matter how the orthonormal functions $\{p(t - kT);\ k \in \mathbb{Z}\}$ are chosen, and these noise random variables $\{Z_k;\ k \in \mathbb{Z}\}$ are also independent of the signal random variables $\{U_k;\ k \in \mathbb{Z}\}$. Chapter 8 also shows that the operation in (6.5) is the appropriate operation to go from waveform to signal sequence in the layered demodulator of Figure 6.1.

Now consider the effect of the noise on the choice of M and d in a PAM modulator. Since the transmitted signal reappears at the receiver with a zero-mean Gaussian random variable added to it, any attempt to retrieve U_k from V_k directly with reasonably small probability of error[6] will require d to exceed several standard deviations of the noise. Thus the noise determines how large d must be, and this, combined with the power constraint, determines M.

The relation between error probability and signal-point spacing also helps explain why multi-level PAM systems almost invariably use a standard M-PAM signal set. Because the Gaussian density drops off so fast with increasing distance, the error probability due to confusion of nearest neighbors drops off equally fast. Thus error probability is dominated by the points in the constellation that are closest together. If the signal points are constrained to have some minimum distance d between points, it can be seen that the minimum energy E_s for a given number of points M is achieved by the standard M-PAM set.[7]

To be more specific about the relationship between M, d, and the variance σ^2 of the noise Z_k, suppose that d is selected to be $\alpha\sigma$, where α is chosen to make the detection sufficiently reliable. Then with $M = 2^b$, where b is the number of bits encoded into each PAM signal, (6.4) becomes

$$E_s = \frac{\alpha^2\sigma^2(2^{2b}-1)}{12}; \qquad b = \frac{1}{2}\log\left(1+\frac{12E_s}{\alpha^2\sigma^2}\right). \tag{6.6}$$

This expression looks strikingly similar to Shannon's capacity formula for additive white Gaussian noise, which says that, for the appropriate PAM bandwidth, the capacity per signal is $C = (1/2)\log(1+E_s/\sigma^2)$. The important difference is that in (6.6) α must be increased, thus decreasing b, in order to decrease error probability. Shannon's result, on the other hand, says that error probability can be made arbitrarily small for any number of bits per signal less than C. Both equations, however, show the same basic form of relationship between bits per signal and the signal-to-noise ratio E_s/σ^2. Both equations also say that if there is no noise ($\sigma^2 = 0$), then the the number of transmitted bits per signal can be infinitely large (i.e. the distance d between signal points can be made infinitesimally small). Thus both equations suggest that noise is a fundamental limitation on communication.

6.2.3 Choice of the modulation pulse

As defined in (6.3), the baseband transmitted waveform, $u(t) = \sum_k u_k p(t-kT)$, for a PAM modulator is determined by the signal constellation $\mathcal{A}$, the signal interval T, and the real $\mathcal{L}_2$ modulation pulse $p(t)$.

[6] If error-correction coding is used with PAM, then d can be smaller, but for any given error-correction code, d still depends on the standard deviation of Z_k.

[7] On the other hand, if we choose a set of M signal points to minimize E_s for a given error probability, then the standard M-PAM signal set is not quite optimal (see Exercise 6.3).

It may be helpful to visualize $p(t)$ as the impulse response of a linear time-invariant filter. Then $u(t)$ is the response of that filter to a sequence of T-spaced impulses $\sum_k u_k \delta(t-kT)$. The problem of choosing $p(t)$ for a given T turns out to be largely separable from that of choosing $\mathcal{A}$. The choice of $p(t)$ is also the more challenging and interesting problem.

The following objectives contribute to the choice of $p(t)$.

- $p(t)$ must be 0 for $t < -\tau$ for some finite τ. To see this, assume that the kth input signal to the modulator arrives at time $Tk - \tau$. The contribution of u_k to the transmitted waveform $u(t)$ cannot start until $kT - \tau$, which implies $p(t) = 0$ for $t < -\tau$ as stated. This rules out $\text{sinc}(t/T)$ as a choice for $p(t)$ (although $\text{sinc}(t/T)$ could be truncated at $t = -\tau$ to satisfy the condition).
- In most situations, $\hat{p}(f)$ should be essentially baseband-limited to some bandwidth B_b slightly larger than $\mathsf{W}_b = 1/2T$. We will see shortly that it cannot be baseband-limited to less than $\mathsf{W}_b = 1/2T$, which is called the nominal, or Nyquist, bandwidth. There is usually an upper limit on B_b because of regulatory constraints at bandpass or to allow for other transmission channels in neighboring bands. If this limit were much larger than $\mathsf{W}_b = 1/2T$, then T could be increased, increasing the rate of transmission.
- The retrieval of the sequence $\{u_k; k \in \mathbb{Z}\}$ from the noisy received waveform should be simple and relatively reliable. In the absence of noise, $\{u_k; k \in \mathbb{Z}\}$ should be uniquely specified by the received waveform.

The first condition above makes it somewhat tricky to satisfy the second condition. In particular, the Paley–Wiener theorem (Paley and Wiener, 1934) states that a necessary and sufficient condition for a nonzero $\mathcal{L}_2$ function $p(t)$ to be zero for all $t < 0$ is that its Fourier transform satisfies

$$\int_{-\infty}^{\infty} \frac{|\ln|\hat{p}(f)||}{1+f^2}\, \mathrm{d}f < \infty. \qquad (6.7)$$

Combining this with the shift condition for Fourier transforms, it says that any $\mathcal{L}_2$ function that is 0 for all $t < -\tau$ for any finite delay τ must also satisfy (6.7). This is a particularly strong statement of the fact that functions cannot be both time- and frequency-limited. One consequence of (6.7) is that if $p(t) = 0$ for $t < -\tau$, then $\hat{p}(f)$ must be nonzero except on a set of measure 0. Another consequence is that $\hat{p}(f)$ must go to 0 with increasing f more slowly than exponentially.

The Paley–Wiener condition turns out to be useless as a tool for choosing $p(t)$. First, it distinguishes whether the delay τ is finite or infinite, but gives no indication of its value when finite. Second, if an $\mathcal{L}_2$ function $p(t)$ is chosen with no concern for (6.7), it can then be truncated to be 0 for $t < -\tau$. The resulting $\mathcal{L}_2$ error caused by truncation can be made arbitrarily small by choosing τ to be sufficiently large. The tradeoff between truncation error and delay is clearly improved by choosing $p(t)$ to approach 0 rapidly as $t \to -\infty$.

In summary, we will replace the first objective above with the objective of choosing $p(t)$ to approach 0 rapidly as $t \to -\infty$. The resulting $p(t)$ will then be truncated to satisfy the original objective. Thus $p(t) \leftrightarrow \hat{p}(f)$ will be an approximation to the transmit pulse in what follows. This also means that $\hat{p}(f)$ can be strictly bandlimited to a frequency slightly larger than $1/2T$.

We now turn to the third objective, particularly that of easily retrieving the sequence $u_1, u_2, \ldots$ from $u(t)$ in the absence of noise. This problem was first analyzed in 1928 in a classic paper by Harry Nyquist (Nyquist, 1928). Before looking at Nyquist's results, however, we must consider the demodulator.

6.2.4 PAM demodulation

For the time being, ignore the channel noise. Assume that the time reference and the amplitude scaling at the receiver have been selected so that the received baseband waveform is the same as the transmitted baseband waveform $u(t)$. This also assumes that no noise has been introduced by the channel.

The problem at the demodulator is then to retrieve the transmitted signals $u_1, u_2, \ldots$ from the received waveform $u(t) = \sum_k u_k p(t-kT)$. The middle layer of a PAM demodulator is defined by a signal interval T (the same as at the modulator) and a real $\mathcal{L}_2$ waveform $q(t)$. The demodulator first filters the received waveform using a filter with impulse response $q(t)$. It then samples the output at T-spaced sample times. That is, the received filtered waveform is given by

$$r(t) = \int_{-\infty}^{\infty} u(\tau)q(t-\tau)\,d\tau, \tag{6.8}$$

and the received samples are $r(T), r(2T), \ldots$

Our objective is to choose $p(t)$ and $q(t)$ so that $r(kT) = u_k$ for each k. If this objective is met for all choices of $u_1, u_2, \ldots,$ then the PAM system involving $p(t)$ and $q(t)$ is said to have *no intersymbol interference*. Otherwise, intersymbol interference is said to exist. The reader should verify that $p(t) = q(t) = (1/\sqrt{T})\text{sinc}(t/T)$ is one solution.

This problem of choosing filters to avoid intersymbol interference appears at first to be somewhat artificial. First, the form of the receiver is restricted to be a filter followed by a sampler. Exercise 6.4 shows that if the detection of each signal is restricted to a linear operation on the received waveform, then there is no real loss of generality in further restricting the operation to be a filter followed by a T-spaced sampler. This does not explain the restriction to linear operations, however.

The second artificiality is neglecting the noise, thus neglecting the fundamental limitation on the bit rate. The reason for posing this artificial problem is, first, that avoiding intersymbol interference is significant in choosing $p(t)$, and, second, that there is a simple and elegant solution to this problem. This solution also provides part of the solution when noise is brought into the picture.

Recall that $u(t) = \sum_k u_k p(t-kT)$; thus, from (6.8),

$$r(t) = \int_{-\infty}^{\infty} \sum_k u_k p(\tau - kT)q(t-\tau)\,d\tau. \tag{6.9}$$

Let $g(t)$ be the convolution $g(t) = p(t) * q(t) = \int p(\tau)q(t-\tau)\mathrm{d}\tau$ and assume[8] that $g(t)$ is $\mathcal{L}_2$. We can then simplify (6.9) as follows:

$$r(t) = \sum_k u_k g(t - kT). \tag{6.10}$$

This should not be surprising. The filters $p(t)$ and $q(t)$ are in cascade with each other. Thus $r(t)$ does not depend on which part of the filtering is done in one and which in the other; it is only the convolution $g(t)$ that determines $r(t)$. Later, when channel noise is added, the individual choice of $p(t)$ and $q(t)$ will become important.

There is no intersymbol interference if $r(kT) = u_k$ for each integer k, and from (6.10) this is satisfied if $g(0) = 1$ and $g(kT) = 0$ for each nonzero integer k. Waveforms with this property are said to be *ideal Nyquist* or, more precisely, *ideal Nyquist with interval T*.

Even though the clock at the receiver is delayed by some finite amount relative to that at the transmitter, and each signal u_k can be generated at the transmitter at some finite time before kT, $g(t)$ must still have the property that $g(t) = 0$ for $t < -\tau$ for some finite τ. As before with the transmit pulse $p(t)$, this finite delay constraint will be replaced with the objective that $g(t)$ should approach 0 rapidly as $|t| \to \infty$. Thus the function $\mathrm{sinc}(t/T)$ is ideal Nyquist with interval T, but is unsuitable because of the slow approach to 0 as $|t| \to \infty$.

As another simple example, the function $\mathrm{rect}(t/T)$ is ideal Nyquist with interval T and can be generated with finite delay, but is not remotely close to being baseband-limited.

In summary, we want to find functions $g(t)$ that are ideal Nyquist but are approximately baseband-limited and approximately time-limited. The Nyquist criterion, discussed in Section 6.3, provides a useful frequency characterization of functions that are ideal Nyquist. This characterization will then be used to study ideal Nyquist functions that are approximately baseband-limited and approximately time-limited.

6.3 The Nyquist criterion

The ideal Nyquist property is determined solely by the T-spaced samples of the waveform $g(t)$. This suggests that the results about aliasing should be relevant. Let $s(t)$ be the baseband-limited waveform generated by the samples of $g(t)$, i.e.

$$s(t) = \sum_k g(kT)\,\mathrm{sinc}\left(\frac{t}{T} - k\right). \tag{6.11}$$

If $g(t)$ is ideal Nyquist, then all the above terms except $k = 0$ disappear and $s(t) = \mathrm{sinc}(t/T)$. Conversely, if $s(t) = \mathrm{sinc}(t/T)$, then $g(t)$ must be ideal Nyquist. Thus $g(t)$

[8] By looking at the frequency domain, it is not difficult to construct a $g(t)$ of infinite energy from $\mathcal{L}_2$ functions $p(t)$ and $q(t)$. When we study noise, however, we find that there is no point in constructing such a $g(t)$, so we ignore the possibility.

is ideal Nyquist if and only if $s(t) = \mathrm{sinc}(t/T)$. Fourier transforming this, $g(t)$ is ideal Nyqist if and only if

$$\hat{s}(f) = T\,\mathrm{rect}(fT). \tag{6.12}$$

From the aliasing theorem,

$$\hat{s}(f) = \mathrm{l.i.m.}\sum_m \hat{g}\Big(f + \frac{m}{T}\Big)\,\mathrm{rect}(fT). \tag{6.13}$$

The result of combining (6.12) and (6.13) is the Nyquist criterion.

Theorem 6.3.1 (Nyquist criterion) *Let $\hat{g}(f)$ be $\mathcal{L}_2$ and satisfy the condition $\lim_{|f|\to\infty} \hat{g}(f)|f|^{1+\varepsilon} = 0$ for some $\varepsilon > 0$. Then the inverse transform, $g(t)$, of $\hat{g}(f)$ is ideal Nyquist with interval T if and only if $\hat{g}(f)$ satisfies the "Nyquist criterion" for T, defined as*[9]

$$\mathrm{l.i.m.}\sum_m \hat{g}(f + m/T)\,\mathrm{rect}(fT) = T\,\mathrm{rect}(fT). \tag{6.14}$$

Proof From the aliasing theorem, the baseband approximation $s(t)$ in (6.11) converges pointwise and is $\mathcal{L}_2$. Similarly, the Fourier transform $\hat{s}(f)$ satisfies (6.13). If $g(t)$ is ideal Nyquist, then $s(t) = \mathrm{sinc}(t/T)$. This implies that $\hat{s}(f)$ is $\mathcal{L}_2$-equivalent to $T\,\mathrm{rect}(fT)$, which in turn implies (6.14). Conversely, satisfaction of the Nyquist criterion (6.14) implies that $\hat{s}(f) = T\,\mathrm{rect}(fT)$. This implies $s(t) = \mathrm{sinc}(t/T)$, implying that $g(t)$ is ideal Nyquist. □

There are many choices for $\hat{g}(f)$ that satisfy (6.14), but the ones of major interest are those that are approximately both bandlimited and time-limited. We look specifically at cases where $\hat{g}(f)$ is strictly bandlimited, which, as we have seen, means that $g(t)$ is not strictly time-limited. Before these filters can be used, of course, they must be truncated to be strictly time-limited. It is strange to look for strictly bandlimited and approximately time-limited functions when it is the opposite that is required, but the reason is that the frequency constraint is the more important. The time constraint is usually more flexible and can be imposed as an approximation.

6.3.1 Band-edge symmetry

The *nominal* or *Nyquist bandwidth* associated with a PAM pulse $g(t)$ with signal interval T is defined to be $\mathsf{W}_b = 1/(2T)$. The actual baseband bandwidth[10] B_b is defined as the smallest number B_b such that $\hat{g}(f) = 0$ for $|f| > B_b$. Note that if $\hat{g}(f) = 0$

[9] It can be seen that $\sum_m \hat{g}(f + m/T)$ is periodic and thus the $\mathrm{rect}(fT)$ could be essentially omitted from both sides of (6.14). Doing this, however, would make the limit in the mean meaningless and would also complicate the intuitive understanding of the theorem.

[10] It might be better to call this the design bandwidth, since after the truncation necessary for finite delay, the resulting frequency function is nonzero a.e. However, if the delay is large enough, the energy outside of B_b is negligible. On the other hand, Exercise 6.9 shows that these approximations must be handled with great care.

for $|f| > \mathrm{W_b}$, then the left side of (6.14) is zero except for $m = 0$, so $\hat{g}(f) = T\,\mathrm{rect}(fT)$. This means that $B_b \geq \mathrm{W_b}$, with equality if and only if $g(t) = \mathrm{sinc}(t/T)$.

As discussed above, if $\mathrm{W_b}$ is much smaller than B_b, then $\mathrm{W_b}$ can be increased, thus increasing the rate R_s at which signals can be transmitted. Thus $g(t)$ should be chosen in such a way that B_b exceeds $\mathrm{W_b}$ by a relatively small amount. In particular, we now focus on the case where $\mathrm{W_b} \leq B_b < 2\mathrm{W_b}$.

The assumption $B_b < 2\mathrm{W_b}$ means that $\hat{g}(f) = 0$ for $|f| \geq 2\mathrm{W_b}$. Thus for $0 \leq f \leq \mathrm{W_b}$, $\hat{g}(f + 2m\mathrm{W_b})$ can be nonzero only for $m = 0$ and $m = -1$. Thus the Nyquist criterion (6.14) in this positive frequency interval becomes

$$\hat{g}(f) + \hat{g}(f - 2\mathrm{W_b}) = T \qquad \text{for } 0 \leq f \leq \mathrm{W_b}. \tag{6.15}$$

Since $p(t)$ and $q(t)$ are real, $g(t)$ is also real, so $\hat{g}(f - 2\mathrm{W_b}) = \hat{g}^*(2\mathrm{W_b} - f)$. Substituting this in (6.15) and letting $\Delta = f - \mathrm{W_b}$, (6.15) becomes

$$T - \hat{g}(\mathrm{W_b} + \Delta) = \hat{g}^*(\mathrm{W_b} - \Delta). \tag{6.16}$$

This is sketched and interpreted in Figure 6.3. The figure assumes the typical situation in which $\hat{g}(f)$ is real. In the general case, the figure illustrates the real part of $\hat{g}(f)$ and the imaginary part satisfies $\Im\{\hat{g}(\mathrm{W_b} + \Delta)\} = \Im\{\hat{g}(\mathrm{W_b} - \Delta)\}$.

Figure 6.3 makes it particularly clear that B_b must satisfy $B_b \geq \mathrm{W_b}$ to avoid intersymbol interference. We then see that the choice of $\hat{g}(f)$ involves a tradeoff between making $\hat{g}(f)$ smooth, so as to avoid a slow time decay in $g(t)$, and reducing the excess of B_b over the Nyquist bandwidth $\mathrm{W_b}$. This excess is expressed as a *rolloff factor*,[11] defined to be $(B_b/\mathrm{W_b}) - 1$, usually expressed as a percentage. Thus $\hat{g}(f)$ in the figure has about a 30% rolloff.

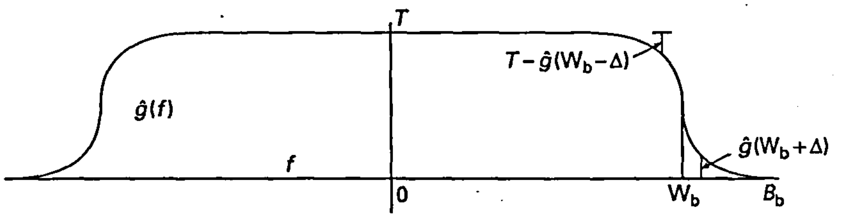

T
$T - \hat{g}(\mathrm{W_b} - \Delta)$
$\hat{g}(f)$
f
$\hat{g}(\mathrm{W_b} + \Delta)$
0
$\mathrm{W_b}$
B_b

Figure 6.3. Band-edge symmetry illustrated for real $\hat{g}(f)$. For each Δ, $0 \leq \Delta \leq \mathrm{W_b}$, $\hat{g}(\mathrm{W_b} + \Delta) = T - \hat{g}(\mathrm{W_b} - \Delta)$. The portion of the curve for $f \geq \mathrm{W_b}$, rotated by 180° around the point $(\mathrm{W_b}, T/2)$, is equal to the portion of the curve for $f \leq \mathrm{W_b}$.

[11] The requirement for a small rolloff actually arises from a requirement on the transmitted pulse $p(t)$, i.e. on the actual bandwidth of the transmitted channel waveform, rather than on the cascade $g(t) = p(t) * q(t)$. The tacit assumption here is that $\hat{p}(f) = 0$ when $\hat{g}(f) = 0$. One reason for this is that it is silly to transmit energy in a part of the spectrum that is going to be completely filtered out at the receiver. We see later that $\hat{p}(f)$ and $\hat{q}(f)$ are usually chosen to have the same magnitude, ensuring that $\hat{p}(f)$ and $\hat{g}(f)$ have the same rolloff.

PAM filters in practice often have *raised cosine* transforms. The raised cosine frequency function, for any given rolloff α between 0 and 1, is defined by

$$\hat{g}_\alpha(f) = \begin{cases} T, & 0 \le |f| \le (1-\alpha)/2T; \\ T\cos^2\left[\frac{\pi T}{2\alpha}\left(|f| - \frac{1-\alpha}{2T}\right)\right], & (1-\alpha)/2T) \le |f| \le (1+\alpha)/2T); \\ 0, & |f| \ge (1+\alpha)/2T). \end{cases} \tag{6.17}$$

The inverse transform of $\hat{g}_\alpha(f)$ can be shown to be (see Exercise 6.8)

$$g_\alpha(t) = \operatorname{sinc}\left(\frac{t}{T}\right)\frac{\cos(\pi\alpha t/T)}{1-4\alpha^2t^2/T^2}, \tag{6.18}$$

which decays asymptotically as $1/t^3$, compared to $1/t$ for sinc(t/T). In particular, for a rolloff $\alpha = 1$, $\hat{g}_\alpha(f)$ is nonzero from $-2\mathsf{W}_\mathsf{b} = -1/T$ to $2\mathsf{W}_\mathsf{b} = 1/T$ and $g_\alpha(t)$ has most of its energy between $-T$ and T. Rolloffs as sharp as 5–10% are used in current practice. The resulting $g_\alpha(t)$ goes to 0 with increasing $|t|$ much faster than sinc(t/T), but the ratio of $g_\alpha(t)$ to sinc(t/T) is a function of $\alpha t/T$ and reaches its first zero at $t = 1.5T/\alpha$. In other words, the required filtering delay is proportional to $1/\alpha$.

The motivation for the raised cosine shape is that $\hat{g}(f)$ should be smooth in order for $g(t)$ to decay quickly in time, but $\hat{g}(f)$ must decrease from T at $\mathsf{W}_\mathsf{b}(1-\alpha)$ to 0 at $\mathsf{W}_\mathsf{b}(1+\alpha)$. As seen in Figure 6.3, the raised cosine function simply rounds off the step discontinuity in rect$(f/2\mathsf{W}_\mathsf{b})$ in such a way as to maintain the Nyquist criterion while making $\hat{g}(f)$ continuous with a continuous derivative, thus guaranteeing that $g(t)$ decays asymptotically with $1/t^3$.

6.3.2 Choosing $\{p(t-kT);\ k \in \mathbb{Z}\}$ as an orthonormal set

In Section 6.3.1, the choice of $\hat{g}(f)$ was described as a compromise between rolloff and smoothness, subject to band-edge symmetry. As illustrated in Figure 6.3, it is not a serious additional constraint to restrict $\hat{g}(f)$ to be real and nonnegative. (Why let $\hat{g}(f)$ go negative or imaginary in making a smooth transition from T to 0?) After choosing $\hat{g}(f) \ge 0$, however, there is still the question of how to choose the transmit filter $p(t)$ and the receive filter $q(t)$ subject to $\hat{p}(f)\hat{q}(f) = \hat{g}(f)$. When studying white Gaussian noise later, we will find that $\hat{q}(f)$ should be chosen to equal $\hat{p}^*(f)$. Thus,[12]

$$|\hat{p}(f)| = |\hat{q}(f)| = \sqrt{\hat{g}(f)}. \tag{6.19}$$

The phase of $\hat{p}(f)$ can be chosen in an arbitrary way, but this determines the phase of $\hat{q}(f) = \hat{p}^*(f)$. The requirement that $\hat{p}(f)\hat{q}(f) = \hat{g}(f) \ge 0$ means that $\hat{q}(f) = \hat{p}^*(f)$. In addition, if $p(t)$ is real then $\hat{p}(-f) = \hat{p}^*(f)$, which determines the phase for negative f in terms of an arbitrary phase for $f > 0$. It is convenient here, however, to be slightly

[12] A function $p(t)$ satisfying (6.19) is often called square root of Nyquist, although it is the magnitude of the transform that is the square root of the transform of an ideal Nyquist pulse.

more general and allow $p(t)$ to be complex. We will prove the following important theorem.

Theorem 6.3.2 (Orthonormal shifts) *Let $p(t)$ be an $\mathcal{L}_2$ function such that $\hat{g}(f) = |\hat{p}(f)|^2$ satisfies the Nyquist criterion for T. Then $\{p(t-kT);\ k \in \mathbb{Z}\}$ is a set of orthonormal functions. Conversely, if $\{p(t-kT);\ k \in \mathbb{Z}\}$ is a set of orthonormal functions, then $|\hat{p}(f)|^2$ satisfies the Nyquist criterion.*

Proof Let $q(t) = p^*(-t)$. Then $g(t) = p(t) * q(t)$, so that

$$g(kT) = \int_{-\infty}^{\infty} p(\tau)q(kT-\tau)\mathrm{d}\tau = \int_{-\infty}^{\infty} p(\tau)p^*(\tau - kT)\mathrm{d}\tau. \tag{6.20}$$

If $\hat{g}(f)$ satisfies the Nyquist criterion, then $g(t)$ is ideal Nyquist and (6.20) has the value 0 for each integer $k \neq 0$ and has the value 1 for $k = 0$. By shifting the variable of integration by jT for any integer j in (6.20), we see also that $\int p(\tau - jT)p^*(\tau - (k+j)T)\mathrm{d}\tau = 0$ for $k \neq 0$ and 1 for $k = 0$. Thus $\{p(t-kT);\ k \in \mathbb{Z}\}$ is an orthonormal set. Conversely, assume that $\{p(t-kT);\ k \in \mathbb{Z}\}$ is an orthonormal set. Then (6.20) has the value 0 for integer $k \neq 0$ and 1 for $k = 0$. Thus $g(t)$ is ideal Nyquist and $\hat{g}(f)$ satisfies the Nyquist criterion. □

Given this orthonormal shift property for $p(t)$, the PAM transmitted waveform $u(t) = \sum_k u_k p(t-kT)$ is simply an orthonormal expansion. Retrieving the coefficient u_k then corresponds to projecting $u(t)$ onto the 1D subspace spanned by $p(t-kT)$. Note that this projection is accomplished by filtering $u(t)$ by $q(t)$ and then sampling at time kT. The filter $q(t)$ is called the *matched filter* to $p(t)$. These filters will be discussed later when noise is introduced into the picture.

Note that we have restricted the pulse $p(t)$ to have unit energy. There is no loss of generality here, since the input signals $\{u_k\}$ can be scaled arbitrarily, and there is no point in having an arbitrary scale factor in both places.

For $|\hat{p}(f)|^2 = \hat{g}(f)$, the actual bandwidth of $\hat{p}(f)$, $\hat{q}(f)$, and $\hat{g}(f)$ are the same, say B_b. Thus if $B_\mathrm{b} < \infty$, we see that $p(t)$ and $q(t)$ can be realized only with infinite delay, which means that both must be truncated. Since $q(t) = p^*(-t)$, they must be truncated for both positive and negative t. We assume that they are truncated at such a large value of delay that the truncation error is negligible. Note that the delay generated by both the transmitter and receiver filter (i.e. from the time that $u_k p(t-kT)$ starts to be formed at the transmitter to the time when u_k is sampled at the receiver) is twice the duration of $p(t)$.

6.3.3 Relation between PAM and analog source coding

The main emphasis in PAM modulation has been that of converting a sequence of T-spaced signals into a waveform. Similarly, the first part of analog source coding is often to convert a waveform into a T-spaced sequence of samples. The major difference is that, with PAM modulation, we have control over the PAM pulse $p(t)$

and thus some control over the class of waveforms. With source coding, we are stuck with whatever class of waveforms describes the source of interest.

For both systems, the nominal bandwidth is $\mathsf{W}_b = 1/2T$, and B_b can be defined as the actual baseband bandwidth of the waveforms. In the case of source coding, $B_b \le \mathsf{W}_b$ is a necessary condition for the sampling approximation $\sum_k u(kT)\,\mathrm{sinc}(t/T-k)$ to recreate perfectly the waveform $u(t)$. The aliasing theorem and the T-spaced sinc-weighted sinusoid expansion were used to analyze the squared error if $B_b > \mathsf{W}_b$.

For PAM, on the other hand, the necessary condition for the PAM demodulator to recreate the initial PAM sequence is $B_b \ge \mathsf{W}_b$. With $B_b > \mathsf{W}_b$, aliasing can be used to advantage, creating an aggregate pulse $g(t)$ that is ideal Nyquist. There is considerable choice in such a pulse, and it is chosen by using contributions from both $f < \mathsf{W}_b$ and $f > \mathsf{W}_b$. Finally we saw that the transmission pulse $p(t)$ for PAM can be chosen so that its T-spaced shifts form an orthonormal set. The sinc functions have this property; however, many other waveforms with slightly greater bandwidth have the same property, but decay much faster with t.

6.4 Modulation: baseband to passband and back

The discussion of PAM in Sections 6.2 and 6.3 focussed on converting a T-spaced sequence of real signals into a real waveform of bandwidth B_b slightly larger than the Nyquist bandwidth $\mathsf{W}_b = 1/2T$. This section focuses on converting that baseband waveform into a passband waveform appropriate for the physical medium, regulatory constraints, and avoiding other transmission bands.

6.4.1 Double-sideband amplitude modulation

The objective of modulating a baseband PAM waveform $u(t)$ to some high-frequency passband around some carrier f_c is simply to shift $u(t)$ up in frequency to $u(t)e^{2\pi i f_c t}$. Thus if $\hat{u}(f)$ is zero except for $-B_b \le f \le B_b$, then the shifted version would be zero except for $f_c - B_b \le f \le f_c + B_b$. This does not quite work since it results in a complex waveform, whereas only real waveforms can actually be transmitted. Thus $u(t)$ is also multiplied by the complex conjugate of $e^{2\pi i f_c t}$, i.e. $e^{-2\pi i f_c t}$, resulting in the following passband waveform:

$$x(t) = u(t)[e^{2\pi i f_c t} + e^{-2\pi i f_c t}] = 2u(t)\cos(2\pi f_c t), \tag{6.21}$$

$$\hat{x}(f) = \hat{u}(f - f_c) + \hat{u}(f + f_c). \tag{6.22}$$

As illustrated in Figure 6.4, $u(t)$ is both translated up in frequency by f_c and also translated down by f_c. Since $x(t)$ must be real, $\hat{x}(f) = \hat{x}^*(-f)$, and the negative frequencies cannot be avoided. Note that the entire set of frequencies in $[-B_b, B_b]$ is both translated up to $[-B_b + f_c, B_b + f_c]$ and down to $[-B_b - f_c, B_b - f_c]$. Thus (assuming $f_c > B_b$) the range of nonzero frequencies occupied by $x(t)$ is twice as large as that occupied by $u(t)$.

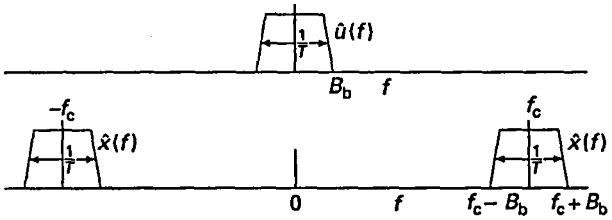


Figure 6.4. Frequency-domain representation of a baseband waveform $u(t)$ shifted up to a passband around the carrier f_c. Note that the baseband bandwidth B_b of $u(t)$ has been doubled to the passband bandwidth $B = 2B_b$ of $x(t)$.

In the communication field, the *bandwidth* of a system is universally defined as the range of *positive* frequencies used in transmission. Since transmitted waveforms are real, the negative frequency part of those waveforms is determined by the positive part and is not counted. This is consistent with our earlier baseband usage, where B_b is the bandwidth of the baseband waveform $u(t)$ in Figure 6.4, and with our new usage for passband waveforms, where $B = 2B_b$ is the bandwidth of $\hat{x}(f)$.

The passband modulation scheme described by (6.21) is called *double-sideband amplitude modulation*. The terminology comes not from the negative frequency band around $-f_c$ and the positive band around f_c, but rather from viewing $[f_c - B_b, f_c + B_b]$ as two sidebands, the upper, $[f_c, f_c + B_b]$, coming from the positive frequency components of $u(t)$ and the lower, $[f_c - B_b, f_c]$ from its negative components. Since $u(t)$ is real, these two bands are redundant and either could be reconstructed from the other.

Double-sideband modulation is quite wasteful of bandwidth since half of the band is redundant. Redundancy is often useful for added protection against noise, but such redundancy is usually better achieved through digital coding.

The simplest and most widely employed solution for using this wasted bandwidth[13] is *quadrature amplitude modulation* (QAM), which is described in Section 6.5. PAM at passband is appropriately viewed as a special case of QAM, and thus the demodulation of PAM from passband to baseband is discussed at the same time as the demodulation of QAM.

6.5 Quadrature amplitude modulation (QAM)

QAM is very similar to PAM except that with QAM the baseband waveform $u(t)$ is chosen to be complex. The complex QAM waveform $u(t)$ is then shifted up to passband

[13] An alternative approach is single-sideband modulation. Here either the positive or negative sideband of a double-sideband waveform is filtered out, thus reducing the transmitted bandwidth by a factor of 2. This used to be quite popular for analog communication, but is harder to implement for digital communication than QAM.

as $u(t)e^{2\pi i f_c t}$. This waveform is complex and is converted into a real waveform for transmission by adding its complex conjugate. The resulting real passband waveform is then given by

$$x(t) = u(t)e^{2\pi i f_c t} + u^*(t)e^{-2\pi i f_c t}. \tag{6.23}$$

Note that the passband waveform for PAM in (6.21) is a special case of this in which $u(t)$ is real. The passband waveform $x(t)$ in (6.23) can also be written in the following equivalent ways:

$$\begin{aligned} x(t) &= 2\Re\{u(t)e^{2\pi i f_c t}\} && (6.24)\\ &= 2\Re\{u(t)\}\cos(2\pi f_c t) - 2\Im\{u(t)\}\sin(2\pi f_c t). && (6.25)\end{aligned}$$

The factor of 2 in (6.24) and (6.25) is an arbitrary scale factor. Some authors leave it out (thus requiring a factor of $1/2$ in (6.23)) and others replace it by $\sqrt{2}$ (requiring a factor of $1/\sqrt{2}$ in (6.23)). This scale factor (however chosen) causes additional confusion when we look at the energy in the waveforms. With the scaling here, $\|x\|^2 = 2\|u\|^2$. Using the scale factor $\sqrt{2}$ solves this problem, but introduces many other problems, not least of which is an extraordinary number of $\sqrt{2}$s in equations. At one level, scaling is a trivial matter, but although the literature is inconsistent, we have tried to be consistent here. One intuitive advantage of the convention here, as illustrated in Figure 6.4, is that the positive frequency part of $x(t)$ is simply $u(t)$ shifted up by f_c.

The remainder of this section provides a more detailed explanation of QAM, and thus also of a number of issues about PAM. A QAM modulator (see Figure 6.5) has the same three layers as a PAM modulator, i.e. first mapping a sequence of bits to a sequence of complex signals, then mapping the complex sequence to a complex baseband waveform, and finally mapping the complex baseband waveform to a real passband waveform.

The demodulator, not surprisingly, performs the inverse of these operations in reverse order, first mapping the received bandpass waveform into a baseband waveform, then recovering the sequence of signals, and finally recovering the binary digits. Each of these layers is discussed in turn.

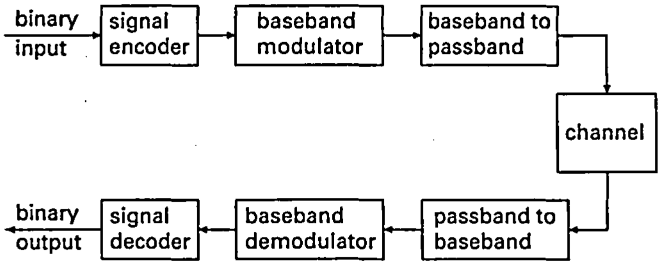


Figure 6.5. QAM modulator and demodulator.

6.5.1 QAM signal set

The input bit sequence arrives at a rate of R bps and is converted, b bits at a time, into a sequence of complex signals u_k chosen from a *signal set* (alphabet, constellation) $\mathcal{A}$ of size $M = |\mathcal{A}| = 2^b$. The *signal rate* is thus $R_s = R/b$ signals/s, and the *signal interval* is $T = 1/R_s = b/R$.

In the case of QAM, the transmitted signals u_k are complex numbers $u_k \in \mathbb{C}$, rather than real numbers. Alternatively, we may think of each signal as a real 2-tuple in $\mathbb{R}^2$.

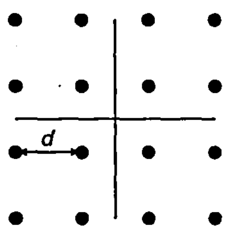
A *standard* $(M' \times M')$*-QAM signal set,* where $M = (M')^2$ is the Cartesian product of two M'-PAM sets; i.e.,

$$\mathcal{A} = \{(a' + \mathrm{i}a'') \mid a' \in \mathcal{A}', a'' \in \mathcal{A}'\},$$

where

$$\mathcal{A}' = \{-d(M'-1)/2, \ldots, -d/2, d/2, \ldots, d(M'-1)/2\}.$$

The signal set $\mathcal{A}$ thus consists of a square array of $M = (M')^2 = 2^b$ signal points located symmetrically about the origin, as illustrated for $M = 16$:



The minimum distance between the 2D points is denoted by d. The average energy per 2D signal, which is denoted by E_s, is simply twice the average energy per dimension:

$$E_s = \frac{d^2[(M')^2 - 1]}{6} = \frac{d^2[M-1]}{6}.$$

In the case of QAM, there are clearly many ways to arrange the signal points other than on a square grid as above. For example, in an M-PSK (phase-shift keyed) signal set, the signal points consist of M equally spaced points on a circle centered on the origin. Thus 4-PSK = 4-QAM. For large M it can be seen that the signal points become very close to each other on a circle so that PSK is rarely used for large M. On the other hand, PSK has some practical advantages because of the uniform signal magnitudes.

As with PAM, the probability of decoding error is primarily a function of the minimum distance d. Not surprisingly, E_s is linear in the signal power of the passband waveform. In wireless systems the signal power is limited both to conserve battery power and to meet regulatory requirements. In wired systems, the power is limited both to avoid crosstalk between adjacent wires and adjacent frequencies, and also to avoid nonlinear effects.

For all of these reasons, it is desirable to choose signal constellations that approximately minimize E_s for a given d and M. One simple result here is that a hexagonal grid of signal points achieves smaller E_s than a square grid for very large M and fixed minimum distance. Unfortunately, finding the optimal signal set to minimize E_s for practical values of M is a messy and ugly problem, and the minima have few interesting properties or symmetries (a possible exception is discussed in Exercise 6.3).

The standard $(M' \times M')$-QAM signal set is almost universally used in practice and will be assumed in what follows.

6.5.2 QAM baseband modulation and demodulation

A QAM baseband modulator is determined by the signal interval T and a complex $\mathcal{L}_2$ waveform $p(t)$. The discrete-time sequence $\{u_k\}$ of complex signal points modulates the amplitudes of a sequence of time shifts $\{p(t-kT)\}$ of the basic pulse $p(t)$ to create a complex transmitted signal $u(t)$ as follows:

$$u(t) = \sum_{k\in\mathbf{Z}} u_k p(t-kT). \tag{6.26}$$

As in the PAM case, we could choose $p(t)$ to be $\text{sinc}(t/T)$, but, for the same reasons as before, $p(t)$ should decay with increasing $|t|$ faster than the sinc function. This means that $\hat{p}(f)$ should be a continuous function that goes to zero rapidly but not instantaneously as f increases beyond $1/2T$. As with PAM, we define $\mathrm{W}_b = 1/2T$ to be the nominal baseband bandwidth of the QAM modulator and B_b to be the actual design bandwidth.

Assume for the moment that the process of conversion to passband, channel transmission, and conversion back to baseband, is ideal, recreating the baseband modulator output $u(t)$ at the input to the baseband demodulator. The baseband demodulator is determined by the interval T (the same as at the modulator) and an $\mathcal{L}_2$ waveform $q(t)$. The demodulator filters $u(t)$ by $q(t)$ and samples the output at T-spaced sample times. Denoting the filtered output by

$$r(t) = \int_{-\infty}^{\infty} u(\tau)q(t-\tau)\mathrm{d}\tau,$$

we see that the received samples are $r(T), r(2T), \ldots$ Note that this is the same as the PAM demodulator except that real signals have been replaced by complex signals. As before, the output $r(t)$ can be represented as

$$r(t) = \sum_k u_k g(t-kT),$$

where $g(t)$ is the convolution of $p(t)$ and $q(t)$. As before, $r(kT) = u_k$ if $g(t)$ is ideal Nyquist, namely if $g(0) = 1$ and $g(kT) = 0$ for all nonzero integer k.

The proof of the Nyquist criterion, Theorem 6.3.1, is valid whether or not $g(t)$ is real. For the reasons explained earlier, however, $\hat{g}(f)$ is usually real and symmetric (as with the raised cosine functions), and this implies that $g(t)$ is also real and symmetric.

Finally, as discussed with PAM, $\hat{p}(f)$ is usually chosen to satisfy $|\hat{p}(f)| = \sqrt{\hat{g}(f)}$. Choosing $\hat{p}(f)$ in this way does not specify the phase of $\hat{p}(f)$, and thus $\hat{p}(f)$ might be real or complex. However $\hat{p}(f)$ is chosen, subject to $|\hat{g}(f)|^2$ satisfying the Nyquist criterion, the set of time shifts $\{p(t-kT)\}$ form an orthonormal set of functions. With this choice also, the baseband bandwidth of $u(t)$, $p(t)$, and $g(t)$ are all the same. Each has a nominal baseband bandwidth given by $1/2T$ and each has an actual baseband bandwidth that exceeds $1/2T$ by some small rolloff factor. As with PAM, $p(t)$ and $q(t)$ must be truncated in time to allow finite delay. The resulting filters are then not quite bandlimited, but this is viewed as a negligible implementation error.

In summary, QAM baseband modulation is virtually the same as PAM baseband modulation. The signal set for QAM is of course complex, and the modulating pulse $p(t)$ can be complex, but the Nyquist results about avoiding intersymbol interference are unchanged.

6.5.3 QAM: baseband to passband and back

Next we discuss modulating the complex QAM baseband waveform $u(t)$ to the passband waveform $x(t)$. Alternative expressions for $x(t)$ are given by (6.23), (6.24), and (6.25), and the frequency representation is illustrated in Figure 6.4.

As with PAM, $u(t)$ has a nominal baseband bandwidth $\mathsf{W}_b = 1/2T$. The actual baseband bandwidth B_b exceeds W_b by some small rolloff factor. The corresponding passband waveform $x(t)$ has a nominal passband bandwidth $\mathsf{W} = 2\mathsf{W}_b = 1/T$ and an actual passband bandwidth $B = 2B_b$. We will assume in everything to follow that $B/2 < f_c$. Recall that $u(t)$ and $x(t)$ are idealized approximations of the true baseband and transmitted waveforms. These true baseband and transmitted waveforms must have finite delay and thus infinite bandwidth, but it is assumed that the delay is large enough that the approximation error is negligible. The assumption[14] $B/2 < f_c$ implies that $u(t)e^{2\pi i f_c t}$ is constrained to positive frequencies and $u(t)e^{-2\pi i f_c t}$ to negative frequencies. Thus the Fourier transform $\hat{u}(f - f_c)$ does not overlap with $\hat{u}(f+f_c)$.

As with PAM, the modulation from baseband to passband is viewed as a two-step process. First $u(t)$ is translated up in frequency by an amount f_c, resulting in a complex passband waveform $x^+(t) = u(t)e^{2\pi i f_c t}$. Next $x^+(t)$ is converted to the real passband waveform $x(t) = [x^+(t)]^* + x^+(t)$.

Assume for now that $x(t)$ is transmitted to the receiver with no noise and no delay. In principle, the received $x(t)$ can be modulated back down to baseband by the reverse of the two steps used in going from baseband to passband. That is, $x(t)$ must first be converted back to the complex positive passband waveform $x^+(t)$, and then $x^+(t)$ must be shifted down in frequency by f_c.

[14] Exercise 6.11 shows that when this assumption is violated, $u(t)$ cannot be perfectly retrieved from $x(t)$, even in the absence of noise. The negligible frequency components of the truncated version of $u(t)$ outside of $B/2$ are assumed to cause negligible error in demodulation.

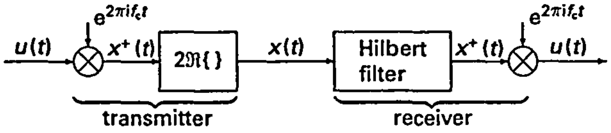


Figure 6.6. Baseband to passband and back.

Mathematically, $x^+(t)$ can be retrieved from $x(t)$ simply by filtering $x(t)$ by a complex filter $h(t)$ such that $\hat{h}(f) = 0$ for $f < 0$ and $\hat{h}(f) = 1$ for $f > 0$. This filter is called a *Hilbert filter*. Note that $h(t)$ is not an $\mathcal{L}_2$ function, but it can be converted to $\mathcal{L}_2$ by making $\hat{h}(f)$ have the value 0 except in the positive passband $[-B/2+f_c, B/2+f_c]$ where it has the value 1. We can then easily retrieve $u(t)$ from $x^+(t)$ simply by a frequency shift. Figure 6.6 illustrates the sequence of operations from $u(t)$ to $x(t)$ and back again.

6.5.4 Implementation of QAM

From an implementation standpoint, the baseband waveform $u(t)$ is usually implemented as two real waveforms, $\Re\{u(t)\}$ and $\Im\{u(t)\}$. These are then modulated up to passband using multiplication by in-phase and out-of-phase carriers as in (6.25), i.e.

$$x(t) = 2\Re\{u(t)\}\cos(2\pi f_c t) - 2\Im\{u(t)\}\sin(2\pi f_c t).$$

There are many other possible implementations, however, such as starting with $u(t)$ given as magnitude and phase. The positive frequency expression $x^+(t) = u(t)e^{2\pi i f_c t}$ is a complex multiplication of complex waveforms which requires four real multiplications rather than the two above used to form $x(t)$ directly. Thus, going from $u(t)$ to $x^+(t)$ to $x(t)$ provides insight but not ease of implementation.

The baseband waveforms $\Re\{u(t)\}$ and $\Im\{u(t)\}$ are easier to generate and visualize if the modulating pulse $p(t)$ is also real. From the discussion of the Nyquist criterion, this is not a fundamental limitation, and there are few reasons for desiring a complex $p(t)$. For real $p(t)$,

$$\Re\{u(t)\} = \sum_k \Re\{u_k\}p(t-kT),$$

$$\Im\{u(t)\} = \sum_k \Im\{u_k\}p(t-kT).$$

Letting $u'_k = \Re\{u_k\}$ and $u''_k = \Im\{u_k\}$, the transmitted passband waveform becomes

$$x(t) = 2\cos(2\pi f_c t)\left(\sum_k u'_k p(t-kT)\right) - 2\sin(2\pi f_c t)\left(\sum_k u''_k p(t-kT)\right). \quad (6.27)$$

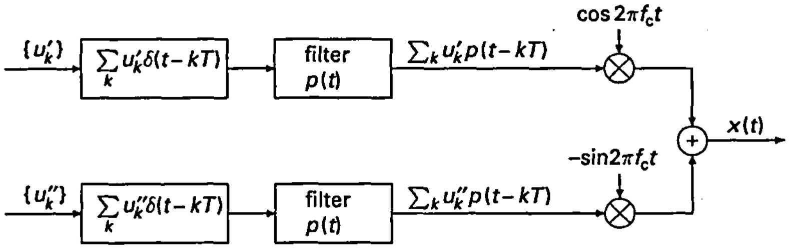


Figure 6.7. DSB-QC modulation.

If the QAM signal set is a standard QAM set, then $\sum_k u'_k p(t-kT)$ and $\sum_k u''_k p(t-kT)$ are parallel baseband PAM systems. They are modulated to passband using "double-sideband" modulation by "quadrature carriers" $\cos(2\pi f_c t)$ and $-\sin(2\pi f_c t)$. These are then summed (with the usual factor of 2), as shown in Figure 6.7. This realization of QAM is called *double-sideband quadrature-carrier* (DSB-QC) modulation.[15]

We have seen that $u(t)$ can be recovered from $x(t)$ by a Hilbert filter followed by shifting down in frequency. A more easily implemented but equivalent procedure starts by multiplying $x(t)$ both by $\cos(2\pi f_c t)$ and by $-\sin(2\pi f_c t)$. Using the trigonometric identities $2\cos^2(\alpha) = 1+\cos(2\alpha)$, $2\sin(\alpha)\cos(\alpha) = \sin(2\alpha)$, and $2\sin^2(\alpha) = 1-\cos(2\alpha)$, these terms can be written as follows:

$$x(t)\cos(2\pi f_c t) = \Re\{u(t)\} + \Re\{u(t)\}\cos(4\pi f_c t) + \Im\{u(t)\}\sin(4\pi f_c t), \quad (6.28)$$

$$-x(t)\sin(2\pi f_c t) = \Im\{u(t)\} - \Re\{u(t)\}\sin(4\pi f_c t) + \Im\{u(t)\}\cos(4\pi f_c t). \quad (6.29)$$

To interpret this, note that multiplying by $\cos(2\pi f_c t) = 1/2\mathrm{e}^{2\pi i f_c t} + 1/2\mathrm{e}^{-2\pi i f_c t}$ both shifts $x(t)$ up[16] and down in frequency by f_c. Thus the positive frequency part of $x(t)$ gives rise to a baseband term and a term around $2f_c$, and the negative frequency part gives rise to a baseband term and a term at $-2f_c$. Filtering out the double-frequency terms then yields $\Re\{u(t)\}$. The interpretation of the sine multiplication is similar.

As another interpretation, recall that $x(t)$ is real and consists of one band of frequencies around f_c and another around $-f_c$. Note also that (6.28) and (6.29) are the real and imaginary parts of $x(t)\mathrm{e}^{-2\pi i f_c t}$, which shifts the positive frequency part of $x(t)$ down to baseband and shifts the negative frequency part down to a band around $-2f_c$. In the Hilbert filter approach, the lower band is filtered out before the frequency shift, and in the approach here it is filtered out after the frequency shift. Clearly the two are equivalent.

[15] The terminology comes from analog modulation in which two real analog waveforms are modulated, respectively, onto cosine and sine carriers. For analog modulation, it is customary to transmit an additional component of carrier from which timing and phase can be recovered. As we see shortly, no such additional carrier is necessary here.

[16] This shift up in frequency is a little confusing, since $x(t)\mathrm{e}^{-2\pi i f_c t} = x(t)\cos(2\pi f_c t) - \mathrm{i}x(t)\sin(2\pi f_c t)$ is only a shift down in frequency. What is happening is that $x(t)\cos(2\pi f_c t)$ is the real part of $x(t)\mathrm{e}^{-2\pi i f_c t}$ and thus needs positive frequency terms to balance the negative frequency terms.

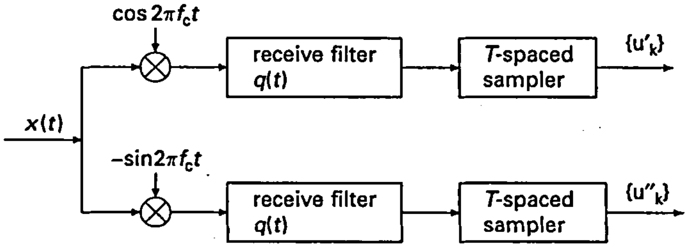


Figure 6.8. DSB-QC demodulation.

It has been assumed throughout that f_c is greater than the baseband bandwidth of $u(t)$. If this is not true, then, as shown in Exercise 6.11, $u(t)$ cannot be retrieved from $x(t)$ by any approach.

Now assume that the baseband modulation filter $p(t)$ is real and a standard QAM signal set is used. Then $\Re\{u(t)\} = \sum u'_k p(t-kT)$ and $\Im\{u(t)\} = \sum u''_k p(t-kT)$ are parallel baseband PAM modulations. Assume also that a receiver filter $q(t)$ is chosen so that $\hat{g}(f) = \hat{p}(f)\hat{q}(f)$ satisfies the Nyquist criterion and all the filters have the common bandwidth $B_b < f_c$. Then, from (6.28), if $x(t)\cos(2\pi f_c t)$ is filtered by $q(t)$, it can be seen that $q(t)$ will filter out the component around $2f_c$. The output from the remaining component $\Re\{u(t)\}$ can then be sampled to retrieve the real signal sequence $u'_1, u'_2, \ldots$ This, plus the corresponding analysis of $-x(t)\sin(2\pi f_c t)$, is illustrated in the DSB-QC receiver in Figure 6.8. Note that the use of the filter $q(t)$ eliminates the need for either filtering out the double-frequency terms or using a Hilbert filter.

The above description of demodulation ignores the noise. As explained in Section 6.3.2, however, if $p(t)$ is chosen so that $\{p(t-kT);\ k \in \mathbb{Z}\}$ is an orthonormal set (i.e. so that $|\hat{p}(f)|^2$ satisfies the Nyquist criterion), then the receiver filter should satisfy $q(t) = p(-t)$. It will be shown later that in the presence of white Gaussian noise, this is the optimal thing to do (in a sense to be described later).

6.6 Signal space and degrees of freedom

Using PAM, real signals can be generated at T-spaced intervals and transmitted in a baseband bandwidth arbitrarily little more than $\mathrm{W}_b = 1/2T$. Thus, over an asymptotically long interval T_0, and in a baseband bandwidth asymptotically close to W_b, $2\mathrm{W}_b T_0$ real signals can be transmitted using PAM.

Using QAM, complex signals can be generated at T-spaced intervals and transmitted in a passband bandwidth arbitrarily little more than $\mathrm{W} = 1/T$. Thus, over an asymptotically long interval T_0, and in a passband bandwidth asymptotically close to W, $\mathrm{W}T_0$ complex signals, and thus $2\mathrm{W}T_0$ real signals, can be transmitted using QAM.

The above description describes PAM at baseband and QAM at passband. To achieve a better comparison of the two, consider an overall large baseband bandwidth W_0 broken into m passbands each of bandwidth W_0/m. Using QAM in each band, we can asymptotically transmit $2W_0T_0$ real signals in a long interval T_0. With PAM used over the entire band W_0, we again asymptotically send $2W_0T_0$ real signals in a duration T_0. We see that, in principle, QAM and baseband PAM with the same overall bandwidth are equivalent in terms of the number of degrees of freedom that can be used to transmit real signals. As pointed out earlier, however, PAM when modulated up to passband uses only half the available degrees of freedom. Also, QAM offers considerably more flexibility since it can be used over an arbitrary selection of frequency bands.

Recall that when we were looking at T-spaced truncated sinusoids and T-spaced sinc-weighted sinusoids, we argued that the class of real waveforms occupying a time interval $(-T_0/2, T_0/2)$ and a frequency interval $(-W_0, W_0)$ has about $2T_0W_0$ degrees of freedom for large W_0, T_0. What we see now is that baseband PAM and passband QAM each employ about $2T_0W_0$ degrees of freedom. In other words, these simple techniques essentially use all the degrees of freedom available in the given bands.

The use of Nyquist theory here has added to our understanding of waveforms that are "essentially" time-and frequency-limited. That is, we can start with a family of functions that are bandlimited within a rolloff factor and then look at asymptotically small rolloffs. The discussion of noise in Chapters 7 and 8 will provide a still better understanding of degrees of freedom subject to essential time and frequency limits.

6.6.1 Distance and orthogonality

Previous sections have shown how to modulate a complex QAM baseband waveform $u(t)$ up to a real passband waveform $x(t)$ and how to retrieve $u(t)$ from $x(t)$ at the receiver. They have also discussed signal constellations that minimize energy for given minimum distance. Finally, the use of a modulation waveform $p(t)$ with orthonormal shifts has connected the energy difference between two baseband signal waveforms, say $u(t) = \sum u_k p(t-kT)$ and $v(t) = \sum_k v_k p(t-kt)$, and the energy difference in the signal points by

$$\|u-v\|^2 = \sum_k |u_k - v_k|^2.$$

Now consider this energy difference at passband. The energy $\|x\|^2$ in the passband waveform $x(t)$ is twice that in the corresponding baseband waveform $u(t)$. Next suppose that $x(t)$ and $y(t)$ are the passband waveforms arising from the baseband waveforms $u(t)$ and $v(t)$, respectively. Then

$$x(t) - y(t) = 2\Re\{u(t)e^{2\pi i f_c t}\} - 2\Re\{v(t)e^{2\pi i f_c t}\} = 2\Re\{[u(t) - v(t)]e^{2\pi i f_c t}\}.$$

Thus $x(t) - y(t)$ is the passband waveform corresponding to $u(t) - v(t)$, so

$$\|x(t) - y(t)\|^2 = 2\|u(t) - v(t)\|^2.$$

This says that, for QAM and PAM, distances between waveforms are preserved (aside from the scale factor of 2 in energy or $\sqrt{2}$ in distance) in going from baseband to passband. Thus distances are preserved in going from signals to baseband waveforms to passband waveforms and back. We will see later that the error probability caused by noise is essentially determined by the distances between the set of passband source waveforms. This error probability is then simply related to the choice of signal constellation and the discrete coding that precedes the mapping of data into signals.

This preservation of distance through the modulation to passband and back is a crucial aspect of the signal-space viewpoint of digital communication. It provides a practical focus to viewing waveforms at baseband and passband as elements of related $\mathcal{L}_2$ inner product spaces.

There is unfortunately a mathematical problem in this very nice story. The set of baseband waveforms forms a complex inner product space, whereas the set of passband waveforms constitutes a real inner product space. The transformation $x(t) = \Re\{u(t)e^{2\pi i f_c t}\}$ is not linear, since, for example, $iu(t)$ does not map into $ix(t)$ for $u(t) \neq 0$. In fact, the notion of a linear transformation does not make much sense, since the transformation goes from complex $\mathcal{L}_2$ to real $\mathcal{L}_2$ and the scalars are different in the two spaces.

Example 6.6.1 As an important example, suppose the QAM modulation pulse is a real waveform $p(t)$ with orthonormal T-spaced shifts. The set of complex baseband waveforms spanned by the orthonormal set $\{p(t-kT);\ k \in \mathbb{Z}\}$ has the form $\sum_k u_k p(t-kT)$, where each u_k is complex. As in (6.27), this is transformed at passband to

$$\sum_k u_k p(t-kT) \to \sum_k 2\Re\{u_k\} p(t-kT)\cos(2\pi f_c t) - 2\sum_k \Im\{u_k\} p(t-kT)\sin(2\pi f_c t).$$

Each baseband function $p(t-kT)$ is modulated to the passband waveform $2p(t-kT)\cos(2\pi f_c t)$. The set of functions $\{p(t-kT)\cos(2\pi f_c t);\ k \in \mathbb{Z}\}$ is not enough to span the space of modulated waveforms, however. It is necessary to add the additional set $\{p(t-kT)\sin(2\pi f_c t);\ k \in \mathbb{Z}\}$. As shown in Exercise 6.15, this combined set of waveforms is an orthogonal set, each with energy 2.

Another way to look at this example is to observe that modulating the baseband function $u(t)$ into the positive passband function $x^+(t) = u(t)e^{2\pi i f_c t}$ is somewhat easier to understand in that the orthonormal set $\{p(t-kT);\ k \in \mathbb{Z}\}$ is modulated to the orthonormal set $\{p(t-kT)e^{2\pi i f_c t};\ k \in \mathbb{Z}\}$, which can be seen to span the space of complex positive frequency passband source waveforms. The additional set of orthonormal waveforms $\{p(t-kT)e^{-2\pi i f_c t};\ k \in \mathbb{Z}\}$ is then needed to span the real passband source waveforms. We then see that the sine/cosine series is simply another way to express this. In the sine/cosine formulation all the coefficients in the series are real, whereas in the complex exponential formulation there is a real and complex coefficient for each term, but they are pairwise-dependent. It will be easier to understand the effects of noise in the sine/cosine formulation.

In the above example, we have seen that each orthonormal function at baseband gives rise to two real orthonormal functions at passband. It can be seen from a degrees-of-freedom argument that this is inevitable no matter what set of orthonormal functions are used at baseband. For a nominal passband bandwidth W, there are 2W real degrees of freedom per second in the baseband complex source waveform, which means there are two real degrees of freedom for each orthonormal baseband waveform. At passband, we have the same 2W degrees of freedom per second, but with a real orthonormal expansion, there is only one real degree of freedom for each orthonormal waveform. Thus there must be two passband real orthonormal waveforms for each baseband complex orthonormal waveform.

The sine/cosine expansion above generalizes in a nice way to an arbitrary set of complex orthonormal baseband functions. Each complex function in this baseband set generates two real functions in an orthogonal passband set. This is expressed precisely in the following theorem, which is proven in Exercise 6.16.

Theorem 6.6.1 *Let $\{\theta_k(t) : k \in \mathbb{Z}\}$ be an orthonormal set limited to the frequency band $[-B/2, B/2]$. Let f_c be greater than $B/2$, and for each $k \in \mathbb{Z}$ let*

$$\psi_{k,1}(t) = \Re\left\{2\theta_k(t)e^{2\pi i f_c t}\right\},$$
$$\psi_{k,2}(t) = \Im\left\{-2\theta_k(t)e^{2\pi i f_c t}\right\}.$$

The set $\{\psi_{k,j};\, k \in \mathbb{Z}, j \in \{1, 2\}\}$ is an orthogonal set of functions, each with energy 2. Furthermore, if $u(t) = \sum_k u_k\theta_k(t)$, then the corresponding passband function $x(t) = 2\Re\{u(t)e^{2\pi i f_c t}\}$ is given by

$$x(t) = \sum_k \Re\{u_k\}\,\psi_{k,1}(t) + \Im\{u_k\}\,\psi_{k,2}(t).$$

This provides a very general way to map any orthonormal set at baseband into a related orthonormal set at passband, with two real orthonormal functions at passband corresponding to each orthonormal function at baseband. It is not limited to any particular type of modulation, and thus will allow us to make general statements about signal space at baseband and passband.

6.7 Carrier and phase recovery in QAM systems

Consider a QAM receiver and visualize the passband-to-baseband conversion as multiplying the positive frequency passband by the complex sinusoid $e^{-2\pi i f_c t}$. If the receiver has a phase error $\phi(t)$ in its estimate of the phase of the transmitted carrier, then it will instead multiply the incoming waveform by $e^{-2\pi i f_c t + i\phi(t)}$. We assume in this analysis that the time reference at the receiver is perfectly known, so that the sampling of the filtered output is carried out at the correct time. Thus the assumption is that the oscillator at the receiver is not quite in phase with the oscillator at the transmitter. Note that the carrier frequency is usually orders of magnitude higher than the baseband bandwidth, and thus a small error in timing is significant in terms of carrier phase but not in terms of sampling. The carrier phase error will rotate the correct complex baseband signal $u(t)$ by $\phi(t)$; i.e. the actual received baseband signal $r(t)$ will be

$$r(t) = e^{i\phi(t)}u(t).$$

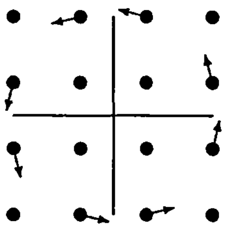


Figure 6.9. Rotation of constellation points by phase error.

If $\phi(t)$ is slowly time-varying relative to the response $q(t)$ of the receiver filter, then the samples $\{r(kT)\}$ of the filter output will be given by

$$r(kT) \approx e^{i\phi(kT)} u_k,$$

as illustrated in Figure 6.9. The phase error $\phi(t)$ is said to come through *coherently.* This phase coherence makes carrier recovery easy in QAM systems.

As can be seen from Figure 6.9, if the phase error is small enough, and the set of points in the constellation are well enough separated, then the phase error can be simply corrected by moving to the closest signal point and adjusting the phase of the demodulating carrier accordingly.

There are two complicating factors here. The first is that we have not taken noise into account yet. When the received signal $y(t)$ is $x(t)+n(t)$, then the output of the T-spaced sampler is not the original signals $\{u_k\}$, but, rather, a noise-corrupted version of them. The second problem is that if a large phase error ever occurs, it cannot be corrected. For example, in Figure 6.9, if $\phi(t) = \pi/2$, then, even in the absence of noise, the received samples always line up with signals from the constellation (but of course not the transmitted signals).

6.7.1 Tracking phase in the presence of noise

The problem of *deciding on* or *detecting* the signals $\{u_k\}$ from the received samples $\{r(kT)\}$ in the presence of noise is a major topic of Chapter 8. Here, however, we have the added complication of both detecting the transmitted signals and tracking and eliminating the phase error.

Fortunately, the problem of decision making and that of phase tracking are largely separable. The oscillators used to generate the modulating and demodulating carriers are relatively stable and have phases which change quite slowly relative to each other. Thus the phase error with any kind of reasonable tracking will be quite small, and thus the data signals can be detected from the received samples almost as if the phase error were zero. The difference between the received sample and the detected data signal will still be nonzero, mostly due to noise but partly due to phase error. However, the noise has zero mean (as we understand later) and thus tends to average out over many sample times. Thus the general approach is to make decisions on the data signals as if the phase error were zero, and then to make slow changes to the phase based on averaging over many sample times. This approach is called *decision-directed carrier recovery*. Note that if we track the phase as phase errors occur, we are also tracking the carrier, in both frequency and phase.

In a decision-directed scheme, assume that the received sample $r(kT)$ is used to make a decision d_k on the transmitted signal point u_k. Also assume that $d_k = u_k$ with very high probability. The apparent phase error for the kth sample is then the difference between the phase of $r(kT)$ and the phase of d_k. Any method for feeding back the apparent phase error to the generator of the sinusoid $\mathrm{e}^{-2\pi \mathrm{i} f_c t + \mathrm{i}\phi(t)}$ in such a way as to reduce the apparent phase error slowly will tend to produce a robust carrier-recovery system.

In one popular method, the feedback signal is taken as the imaginary part of $r(kT)d_k^*$. If the phase angle from d_k to $r(kT)$ is ϕ_k, then

$$r(kT)d_k^* = |r(kT)||d_k|\mathrm{e}^{\mathrm{i}\phi_k},$$

so the imaginary part is $|r(kT)||d_k|\sin\phi_k \approx |r(kT)||d_k|\phi_k$, when ϕ_k is small. Decision-directed carrier recovery based on such a feedback signal can be extremely robust even in the presence of substantial distortion and large initial phase errors. With a second-order phase-locked carrier-recovery loop, it turns out that the carrier frequency f_c can be recovered as well.

6.7.2 Large phase errors

A problem with decision-directed carrier recovery, as with many other approaches, is that the recovered phase may settle into any value for which the received eye pattern (i.e. the pattern of a long string of received samples as viewed on a scope) "looks OK." With $(M \times M)$-QAM signal sets, as in Figure 6.9, the signal set has four-fold symmetry, and phase errors of 90°, 180°, or 270° are not detectable. Simple differential coding methods that transmit the "phase" (quadrantal) part of the signal information as a change of phase from the previous signal rather than as an absolute phase can easily overcome this problem. Another approach is to resynchronize the system frequently by sending some known pattern of signals. This latter approach is frequently used in wireless systems, where fading sometimes causes a loss of phase synchronization.

6.8 Summary of modulation and demodulation

This chapter has used the signal space developed in Chapters 4 and 5 to study the mapping of binary input sequences at a modulator into the waveforms to be transmitted over the channel. Figure 6.1 summarized this process, mapping bits to signals, then signals to baseband waveforms, and then baseband waveforms to passband waveforms. The demodulator goes through the inverse process, going from passband waveforms to baseband waveforms, to signals, to bits. This breaks the modulation process into three layers that can be studied more or less independently.

The development used PAM and QAM throughout, both as widely used systems and as convenient ways to bring out the principles that can be applied more widely.

The mapping from binary digits to signals segments the incoming binary sequence into b-tuples of bits and then maps the set of $M = 2^b$ n-tuples into a constellation of M signal points in $\mathbb{R}^m$ or $\mathbb{C}^m$ for some convenient m. Since the m components of these signal points are going to be used as coefficients in an orthogonal expansion to generate the waveforms, the objectives are to choose a signal constellation with small average energy but with a large distance between each pair of points. PAM is an example where the signal space is $\mathbb{R}^1$, and QAM is an example where the signal space is $\mathbb{C}^1$. For both of these, the standard mapping is the same as the representation points of a uniform quantizer. These are not quite optimal in terms of minimizing the average energy for a given minimum point spacing, but they are almost universally used because of the near optimality and the simplicity.

The mapping of signals into baseband waveforms for PAM chooses a fixed waveform $p(t)$ and modulates the sequence of signals $u_1, u_2, \ldots$ into the baseband waveform $\sum_j u_j p(t-jT)$. One of the objectives in choosing $p(t)$ is to be able to retrieve the sequence $u_1, u_2, \ldots$ from the received waveform. This involves an output filter $q(t)$ which is sampled each T seconds to retrieve $u_1, u_2, \ldots$ The Nyquist criterion was derived, specifying the properties that the product $\hat{g}(f) = \hat{p}(f)\hat{q}(f)$ must satisfy to avoid intersymbol interference. The objective in choosing $\hat{g}(f)$ is a trade-off between the closeness of $\hat{g}(f)$ to $T\,\mathrm{rect}(fT)$ and the time duration of $g(t)$, subject to satisfying the Nyquist criterion. The raised cosine functions are widely used as a good compromise between these dual objectives. For a given real $\hat{g}(f)$, the choice of $\hat{p}(f)$ usually satisfies $\hat{g}(f) = |\hat{p}(f)|^2$, and in this case $\{p(t-kT); k \in \mathbb{Z}\}$ is a set of orthonormal functions.

Most of the remainder of the chapter discussed modulation from baseband to passband. This is an elementary topic in manipulating Fourier transforms, and need not be summarized here.

6.9 Exercises

6.1 (PAM) Consider standard M-PAM and assume that the signals are used with equal probability. Show that the average energy per signal $E_s = \overline{U_k^2}$ is equal to the average energy $\overline{U^2} = d^2M^2/12$ of a uniform continuous distribution over the interval $[-dM/2, dM/2]$, minus the average energy $\overline{(U-U_k)^2} = d^2/12$ of a uniform continuous distribution over the interval $[-d/2, d/2]$:

$$E_s = \frac{d^2(M^2-1)}{12}.$$

This establishes (6.4). Verify the formula for $M = 4$ and $M = 8$.

6.2 (PAM) A discrete memoryless source emits binary equiprobable symbols at a rate of 1000 symbols/s. The symbols from a 1 s interval are grouped into pairs and sent over a bandlimited channel using a standard 4-PAM signal set. The modulation uses a signal interval 0.002 and a pulse $p(t) = \text{sinc}(t/T)$.)

(a) Suppose that a sample sequence $u_1, \dots, u_{500}$ of transmitted signals includes 115 appearances of $3d/2$, 130 appearances of $d/2$, 120 appearances of $-d/2$, and 135 appearances of $-3d/2$. Find the energy in the corresponding transmitted waveform $u(t) = \sum_{k=1}^{500} u_k \,\text{sinc}(t/T - k)$ as a function of d.

(b) What is the bandwidth of the waveform $u(t)$ in part (a)?

(c) Find $\mathsf{E}\left[\int U^2(t)\mathrm{d}t\right]$, where $U(t)$ is the random waveform given by $\sum_{k=1}^{500} U_k \;\text{sinc}(t/T - k)$.

(d) Now suppose that the binary source is not memoryless, but is instead generated by a Markov chain, where

$$\Pr(X_i = 1 \mid X_{i-1} = 1) = \Pr(X_i = 0 \mid X_{i-1} = 0) = 0.9.$$

Assume the Markov chain starts in steady state with $\Pr(X_1 = 1) = 1/2$. Using the mapping $(00 \to a_1)$, $(01 \to a_2)$, $(10 \to a_3)$, $(11 \to a_4)$, find $\mathsf{E}[U_k^2]$ for $1 \le k \le 500$.

(e) Find $\mathsf{E}\left[\int U^2(t)\,\mathrm{d}t\right]$ for this new source.

(f) For the above Markov chain, explain how the above mapping could be changed to reduce the expected energy without changing the separation between signal points.

6.3 (a) Assume that the received signal in a 4-PAM system is $V_k = U_k + Z_k$, where U_k is the transmitted 4-PAM signal at time k. Let Z_k be independent of U_k and Gaussian with density $f_Z(z) = \sqrt{1/2\pi}\ \exp(-z^2/2)$. Assume that the receiver chooses the signal $\tilde{U}_k$ closest to V_k. (It is shown in Chapter 8 that this detection rule minimizes P_e for equiprobable signals.) Find the probability P_e (in terms of Gaussian integrals) that $U_k \neq \tilde{U}_k$.

(b) Evaluate the partial derivitive of P_e with respect to the third signal point a_3 (i.e. the positive inner signal point) at the point where a_3 is equal to its value $d/2$ in standard 4-PAM and all other signal points are kept at their 4-PAM values. [Hint. This does not require any calculation.]

(c) Evaluate the partial derivitive of the signal energy E_s with respect to a_3.

(d) Argue from this that the signal constellation with minimum-error probability for four equiprobable signal points is not 4-PAM, but rather a constellation, where the distance between the inner points is smaller than the distance from inner point to outer point on either side. (This is quite surprising intuitively to the author.)

6.4 (Nyquist) Suppose that the PAM modulated baseband waveform $u(t) = \sum_{k=-\infty}^{\infty} u_k p(t - kT)$ is received. That is, $u(t)$ is known, T is known, and $p(t)$ is known. We want to determine the signals $\{u_k\}$ from $u(t)$. Assume only linear operations can be used. That is, we wish to find some waveform $d_k(t)$ for each integer k such that $\int_{-\infty}^{\infty} u(t)d_k(t)\mathrm{d}t = u_k$.

(a) What properites must be satisfied by $d_k(t)$ such that the above equation is satisfied no matter what values are taken by the other signals, $\ldots, u_{k-2}, u_{k-1}, u_{k+1}, u_{k+2}, \ldots$? These properties should take the form of constraints on the inner products $\langle p(t-kT), d_j(t)\rangle$. Do not worry about convergence, interchange of limits, etc.

(b) Suppose you find a function $d_0(t)$ that satisfies these constraints for $k = 0$. Show that, for each k, a function $d_k(t)$ satisfying these constraints can be found simply in terms of $d_0(t)$.

(c) What is the relationship between $d_0(t)$ and a function $q(t)$ that avoids intersymbol interference in the approach taken in Section 6.3 (i.e. a function $q(t)$ such that $p(t) * q(t)$ is ideal Nyquist)?

You have shown that the filter/sample approach in Section 6.3 is no less general than the arbitrary linear operation approach here. Note that, in the absence of noise and with a known signal constellation, it might be possible to retrieve the signals from the waveform using nonlinear operations even in the presence of intersymbol interference.

6.5 (Nyquist) Let $v(t)$ be a continuous $\mathcal{L}_2$ waveform with $v(0) = 1$ and define $g(t) = v(t)\,\mathrm{sinc}(t/T)$.

(a) Show that $g(t)$ is ideal Nyquist with interval T.

(b) Find $\hat{g}(f)$ as a function of $\hat{v}(f)$.

(c) Give a direct demonstration that $\hat{g}(f)$ satisfies the Nyquist criterion.

(d) If $v(t)$ is baseband-limited to B_b, what is $g(t)$ baseband-limited to?

Note: the usual form of the Nyquist criterion helps in choosing waveforms that avoid intersymbol interference with prescribed rolloff properties in frequency. The approach above show how to avoid intersymbol interference with prescribed attenuation in time and in frequency.

6.6 (Nyquist) Consider a PAM baseband system in which the modulator is defined by a signal interval T and a waveform $p(t)$, the channel is defined by a filter $h(t)$, and the receiver is defined by a filter $q(t)$ which is sampled at T-spaced intervals. The received waveform, after the receiver filter $q(t)$, is then given by $r(t) = \sum_k u_k g(t-kT)$, where $g(t) = p(t) * h(t) * q(t)$.

(a) What property must $g(t)$ have so that $r(kT) = u_k$ for all k and for all choices of input $\{u_k\}$? What is the Nyquist criterion for $\hat{g}(f)$?

(b) Now assume that $T = 1/2$ and that $p(t), h(t), q(t)$ and all their Fourier transforms are restricted to be real. Assume further that $\hat{p}(f)$ and $\hat{h}(f)$ are specified by Figure 6.10, i.e. by

$$\hat{p}(f) = \begin{cases} 1 & |f| \le 0.5; \\ 1.5 - t & 0.5 < |f| \le 1.5; \\ 0 & |f| > 1.5; \end{cases} \qquad \hat{h}(f) = \begin{cases} 1 & |f| \le 0.75; \\ 0 & 0.75 < |f| \le 1; \\ 1 & 1 < |f| \le 1.25; \\ 0 & |f| > 1.25. \end{cases}$$

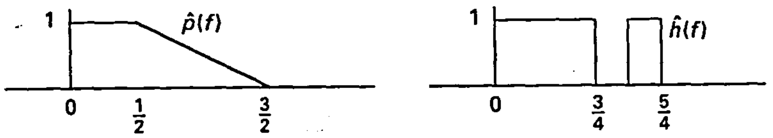


Figure 6.10.

Is it possible to choose a receiver filter transform $\hat{q}(f)$ so that there is no intersymbol interference? If so, give such a $\hat{q}(f)$ and indicate the regions in which your solution is nonunique.

(c) Redo part (b) with the modification that now $\hat{h}(f) = 1$ for $|f| \leq 0.75$ and $\hat{h}(f) = 0$ for $|f| > 0.75$.

(d) Explain the conditions on $\hat{p}(f)\hat{h}(f)$ under which intersymbol interference can be avoided by proper choice of $\hat{q}(f)$. (You may assume, as above, that $\hat{p}(f), \hat{h}(f), p(t)$, and $h(t)$ are all real.)

6.7 (Nyquist) Recall that the rect(t/T) function has the very special property that it, plus its time and frequency shifts by kT and j/T, respectively, form an orthogonal set of functions. The function sinc(t/T) has this same property. This problem is about some other functions that are generalizations of rect(t/T) and which, as you will show in parts (a)–(d), have this same interesting property. For simplicity, choose $T = 1$.

These functions take only the values 0 and 1 and are allowed to be nonzero only over $[-1, 1]$ rather than $[-1/2, 1/2]$ as with rect(t). Explicitly, the functions considered here satisfy the following constraints:

$$p(t) = p^2(t) \qquad \text{for all } t \quad \text{(0/1 property);} \qquad (6.30)$$

$$p(t) = 0 \qquad \text{for } |t| > 1; \qquad (6.31)$$

$$p(t) = p(-t) \qquad \text{for all } t \quad \text{(symmetry);} \qquad (6.32)$$

$$p(t) = 1 - p(t-1) \qquad \text{for } 0 \leq t < 1/2. \qquad (6.33)$$

Two examples of functions $P(t)$ satisfying (6.30)–(6.33) are illustrated in Figure 6.11. Note: because of property (6.32), condition (6.33) also holds for $1/2 < t \leq 1$. Note also that $p(t)$ at the single points $t = \pm 1/2$ does not affect any orthogonality properties, so you are free to ignore these points in your arguments.

(a) Show that $p(t)$ is orthogonal to $p(t-1)$. [Hint. Evaluate $p(t)p(t-1)$ for each $t \in [0, 1]$ other than $t = 1/2$.]

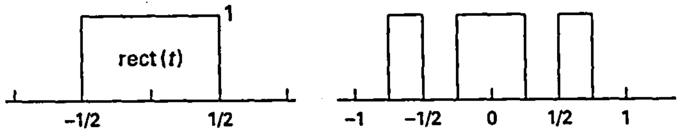


Figure 6.11. Two simple functions $p(t)$ that satisfy (6.30)–(6.33).

(b) Show that $p(t)$ is orthogonal to $p(t-k)$ for all integer $k \neq 0$.

(c) Show that $p(t)$ is orthogonal to $p(t-k)e^{2\pi imt}$ for integer $m \neq 0$ and $k \neq 0$.

(d) Show that $p(t)$ is orthogonal to $p(t)e^{2\pi imt}$ for integer $m \neq 0$. [Hint. Evaluate $p(t)e^{-2\pi imt} + p(t-1)e^{-2\pi im(t-1)}$.]

(e) Let $h(t) = \hat{p}(t)$, where $\hat{p}(f)$ is the Fourier transform of $p(t)$. If $p(t)$ satisfies properties (6.30) to (6.33), does it follow that $h(t)$ has the property that it is orthogonal to $h(t-k)e^{2\pi imt}$ whenever either the integer k or m is nonzero?

Note: almost no calculation is required in this exercise.

6.8 (Nyquist)

(a) For the special case $\alpha = 1$, $T = 1$, verify the formula in (6.18) for $g_1(t)$ given $\hat{g}_1(f)$ in (6.17). [Hint. As an intermediate step, verify that $g_1(t) = \text{sinc}(2t) + (1/2)\,\text{sinc}(2t+1) + (1/2)\,\text{sinc}(2t-1)$.] Sketch $g_1(t)$, in particular showing its value at $mT/2$ for each $m \geq 0$.

(b) For the general case $0 < \alpha < 1$, $T = 1$, show that $\hat{g}_\alpha(f)$ is the convolution of $\text{rect}(f)$ with a half cycle of $\beta\cos(\pi f/\alpha)$ and find β.

(c) Verify (6.18) for $0 < \alpha < 1$, $T = 1$, and then verify for arbitrary $T > 0$.

6.9 (Approximate Nyquist) This exercise shows that approximations to the Nyquist criterion must be treated with great care. Define $\hat{g}_k(f)$, for integer $k \geq 0$ as in Figure 6.12 for $k = 2$. For arbitrary k, there are k small pulses on each side of the main pulse, each of height $1/k$.

(a) Show that $\hat{g}_k(f)$ satisfies the Nyquist criterion for $T = 1$ and for each $k \geq 1$.

(b) Show that $\text{l.i.m.}_{k\to\infty}\, \hat{g}_k(f)$ is simply the central pulse above. That is, this $\mathcal{L}_2$ limit satisfies the Nyquist criterion for $T = 1/2$. To put it another way, $\hat{g}_k(f)$, for large k, satisfies the Nyquist criterion for $T = 1$ using "approximately" the bandwidth $1/4$ rather than the necessary bandwidth $1/2$. The problem is that the $\mathcal{L}_2$ notion of approximation (done carefully here as a limit in the mean of a sequence of approximations) is not always appropriate, and it is often inappropriate with sampling issues.

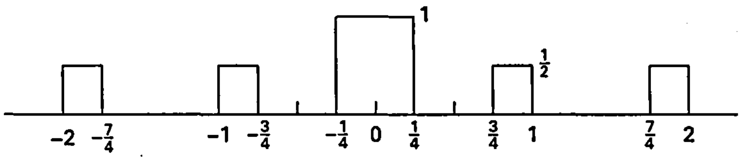


Figure 6.12.

6.10 (Nyquist)

(a) Assume that $\hat{p}(f) = \hat{q}^*(f)$ and $\hat{g}(f) = \hat{p}(f)\hat{q}(f)$. Show that if $p(t)$ is real, then $\hat{g}(f) = \hat{g}(-f)$ for all f.

(b) Under the same assumptions, find an example where $p(t)$ is not real but $\hat{g}(f) \neq \hat{g}(-f)$ and $\hat{g}(f)$ satisifes the Nyquist criterion. [Hint. Show that

$\hat{g}(f) = 1$ for $0 \leq f \leq 1$ and $\hat{g}(f) = 0$ elsewhere satisfies the Nyquist criterion for $T = 1$ and find the corresponding $p(t)$.]

6.11 (Passband)

(a) Let $u_k(t) = \exp(2\pi i f_k t)$ for $k = 1, 2$ and let $x_k(t) = 2\Re\{u_k(t)\exp(2\pi i f_c t)\}$. Assume $f_1 > -f_c$ and find the $f_2 \neq f_1$ such that $x_1(t) = x_2(t)$.

(b) Explain that what you have done is to show that, without the assumption that the bandwidth of $u(t)$ is less than f_c, it is impossible to always retrieve $u(t)$ from $x(t)$, even in the absence of noise.

(c) Let $y(t)$ be a real $\mathcal{L}_2$ function. Show that the result in part (a) remains valid if $u_k(t) = y(t)\exp(2\pi i f_k t)$ (i.e. show that the result in part (a) is valid with a restriction to $\mathcal{L}_2$ functions).

(d) Show that if $u(t)$ is restricted to be real, then $u(t)$ can be retrieved a.e. from $x(t) = 2\Re\{u(t)\exp(2\pi i f_c t)\}$. [Hint. Express $x(t)$ in terms of $\cos(2\pi f_c t)$.]

(e) Show that if the bandwidth of $u(t)$ exceeds f_c, then neither Figure 6.6 nor Figure 6.8 work correctly, even when $u(t)$ is real.

6.12 (QAM)

(a) Let $\theta_1(t)$ and $\theta_2(t)$ be orthonormal complex waveforms. Let $\phi_j(t) = \theta_j(t)e^{2\pi i f_c t}$ for $j = 1, 2$. Show that $\phi_1(t)$ and $\phi_2(t)$ are orthonormal for any f_c.

(b) Suppose that $\theta_2(t) = \theta_1(t - T)$. Show that $\phi_2(t) = \phi_1(t - T)$ if f_c is an integer multiple of $1/T$.

6.13 (QAM)

(a) Assume $B/2 < f_c$. Let $u(t)$ be a real function and let $v(t)$ be an imaginary function, both baseband-limited to $B/2$. Show that the corresponding passband functions, $\Re\{u(t)e^{2\pi i f_c t}\}$ and $\Re\{v(t)e^{2\pi i f_c t}\}$, are orthogonal.

(b) Give an example where the functions in part (a) are not orthogonal if $B/2 > f_c$.

6.14 (a) Derive (6.28) and (6.29) using trigonometric identities.

(b) View the left side of (6.28) and (6.29) as the real and imaginary part, respectively, of $x(t)e^{-2\pi i f_c t}$. Rederive (6.28) and (6.29) using complex exponentials. (Note how much easier this is than part (a).)

6.15 (Passband expansions) Assume that $\{p(t - kT) : k \in \mathbb{Z}\}$ is a set of orthonormal functions. Assume that $\hat{p}(f) = 0$ for $|f| \geq f_c$).

(a) Show that $\{\sqrt{2}p(t - kT)\cos(2\pi f_c t);\ k \in \mathbb{Z}\}$ is an orthonormal set.

(b) Show that $\{\sqrt{2}p(t - kT)\sin(2\pi f_c t);\ k \in \mathbb{Z}\}$ is an orthonormal set and that each function in it is orthonormal to the cosine set in part (a).

6.16 (Passband expansions) Prove Theorem 6.6.1. [Hint. First show that the set of functions $\{\hat{\psi}_{k,1}(f)\}$ and $\{\hat{\psi}_{k,2}(f)\}$ are orthogonal with energy 2 by comparing the integral over negative frequencies with that over positive frequencies.] Indicate explicitly why you need $f_c > B/2$.

6.17 (Phase and envelope modulation) This exercise shows that any real passband waveform can be viewed as a combination of phase and amplitude modulation. Let $x(t)$ be an $\mathcal{L}_2$ real passband waveform of bandwidth B around a carrier frequency $f_c > B/2$. Let $x^+(t)$ be the positive frequency part of $x(t)$ and let $u(t) = x^+(t)\exp\{-2\pi i f_c t\}$.

(a) Express $x(t)$ in terms of $\Re\{u(t)\}, \Im\{u(t)\}, \cos[2\pi f_c t]$, and $\sin[2\pi f_c t]$.

(b) Define $\phi(t)$ implicitly by $e^{i\phi(t)} = u(t)/|u(t)|$. Show that $x(t)$ can be expressed as $x(t) = 2|u(t)|\cos[2\pi f_c t + \phi(t)]$. Draw a sketch illustrating that $2|u(t)|$ is a baseband waveform upperbounding $x(t)$ and touching $x(t)$ roughly once per cycle. Either by sketch or words illustrate that $\phi(t)$ is a phase modulation on the carrier.

(c) Define the *envelope* of a passband waveform $x(t)$ as twice the magnitude of its positive frequency part, i.e. as $2|x^+(t)|$. Without changing the waveform $x(t)$ (or $x^+(t)$) from that before, change the carrier frequency from f_c to some other frequency f_c'. Thus $u'(t) = x^+(t)\exp\{-2\pi i f_c' t\}$. Show that $|x^+(t)| = |u(t)| = |u'(t)|$. Note that you have shown that the envelope does not depend on the assumed carrier frequency, but has the interpretation of part (b).

(d) Show the relationship of the phase $\phi'(t)$ for the carrier f_c' to that for the carrier f_c.

(e) Let $p(t) = |x(t)|^2$ be the power in $x(t)$. Show that if $p(t)$ is lowpass filtered to bandwidth B, the result is $2|u(t)|^2$. Interpret this filtering as a short term average over $|x(t)|^2$ to interpret why the envelope squared is twice the short term average power (and thus why the envelope is $\sqrt{2}$ times the short term root-mean-squared amplitude).

6.18 (Carrierless amplitude-phase modulation (CAP)) We have seen how to modulate a baseband QAM waveform up to passband and then demodulate it by shifting down to baseband, followed by filtering and sampling. This exercise explores the interesting concept of eliminating the baseband operations by modulating and demodulating directly at passband. This approach is used in one of the North American standards for asymmetrical digital subscriber loop (ADSL).

(a) Let $\{u_k\}$ be a complex data sequence and let $u(t) = \sum_k u_k p(t - kT)$ be the corresponding modulated output. Let $\hat{p}(f)$ be equal to $\sqrt{T}$ over $f \in [3/2T, 5/2T]$ and be equal to 0 elsewhere. At the receiver, $u(t)$ is filtered using $p(t)$ and the output $y(t)$ is then T-space sampled at time instants kT. Show that $y(kT) = u_k$ for all $k \in \mathbb{Z}$. Don't worry about the fact that the transmitted waveform $u(t)$ is complex.

(b) Now suppose that $\hat{p}(f) = \sqrt{T}\text{rect}(T(f - f_c))$ for some arbitrary f_c rather than $f_c = 2/T$ as in part (a). For what values of f_c does the scheme still work?

(c) Suppose that $\Re\{u(t)\}$ is now sent over a communication channel. Suppose that the received waveform is filtered by a Hilbert filter before going through the demodulation procedure above. Does the scheme still work?

7 Random processes and noise

7.1 Introduction

Chapter 6 discussed modulation and demodulation, but replaced any detailed discussion of the noise by the assumption that a minimal separation is required between each pair of signal points. This chapter develops the underlying principles needed to understand noise, and Chapter 8 shows how to use these principles in detecting signals in the presence of noise.

Noise is usually the fundamental limitation for communication over physical channels. This can be seen intuitively by accepting for the moment that different possible transmitted waveforms must have a difference of some minimum energy to overcome the noise. This difference reflects back to a required distance between signal points, which, along with a transmitted power constraint, limits the number of bits per signal that can be transmitted.

The transmission rate in bits per second is then limited by the product of the number of bits per signal times the number of signals per second, i.e. the number of degrees of freedom per second that signals can occupy. This intuitive view is substantially correct, but must be understood at a deeper level, which will come from a probabilistic model of the noise.

This chapter and the next will adopt the assumption that the channel output waveform has the form $y(t) = x(t) + z(t)$, where $x(t)$ is the channel input and $z(t)$ is the noise. The channel input $x(t)$ depends on the random choice of binary source digits, and thus $x(t)$ has to be viewed as a particular selection out of an ensemble of possible channel inputs. Similarly, $z(t)$ is a particular selection out of an ensemble of possible noise waveforms.

The assumption that $y(t) = x(t) + z(t)$ implies that the channel attenuation is known and removed by scaling the received signal and noise. It also implies that the input is not filtered or distorted by the channel. Finally it implies that the delay and carrier phase between input and output are known and removed at the receiver.

The noise should be modeled probabilistically. This is partly because the noise is a priori unknown, but can be expected to behave in statistically predictable ways. It is also because encoders and decoders are designed to operate successfully on a variety of different channels, all of which are subject to different noise waveforms. The noise is usually modeled as zero mean, since a mean can be trivially removed.

Modeling the waveforms $x(t)$ and $z(t)$ probabilistically will take considerable care. If $x(t)$ and $z(t)$ were defined only at discrete values of time, such as $\{t = kT; k \in \mathbb{Z}\}$, then they could be modeled as sample values of sequences of random variables (rvs). These sequences of rvs could then be denoted by $X(t) = \{X(kT); k \in \mathbb{Z}\}$ and $Z(t) = \{Z(kT); k \in \mathbb{Z}\}$. The case of interest here, however, is where $x(t)$ and $z(t)$ are defined over the continuum of values of t, and thus a continuum of rvs is required. Such a probabilistic model is known as a *random process* or, synonymously, a *stochastic process*. These models behave somewhat similarly to random sequences, but they behave differently in a myriad of small but important ways.

7.2 Random processes

A *random process* $\{Z(t); t \in \mathbb{R}\}$ is a collection[1] of rvs, one for each $t \in \mathbb{R}$. The parameter t usually models time, and any given instant in time is often referred to as an *epoch*. Thus there is one rv for each epoch. Sometimes the range of t is restricted to some finite interval, $[a, b]$, and then the process is denoted by $\{Z(t); t \in [a, b]\}$.

There must be an underlying sample space Ω over which these rvs are defined. That is, for each epoch $t \in \mathbb{R}$ (or $t \in [a, b]$), the rv $Z(t)$ is a function $\{Z(t, \omega); \omega \in \Omega\}$ mapping sample points $\omega \in \Omega$ to real numbers.

A given sample point $\omega \in \Omega$ within the underlying sample space determines the sample values of $Z(t)$ for each epoch t. The collection of all these sample values for a given sample point ω, i.e. $\{Z(t, \omega); t \in \mathbb{R}\}$, is called a *sample function* $\{z(t): \mathbb{R} \to \mathbb{R}\}$ of the process.

Thus $Z(t, \omega)$ can be viewed as a function of ω for fixed t, in which case it is the rv $Z(t)$, or it can be viewed as a function of t for fixed ω, in which case it is the sample function $\{z(t): \mathbb{R} \to \mathbb{R}\} = \{Z(t, \omega); t \in \mathbb{R}\}$ corresponding to the given ω. Viewed as a function of both t and ω, $\{Z(t, \omega); t \in \mathbb{R}, \omega \in \Omega\}$ is the random process itself; the sample point ω is usually suppressed, leading to the notation $\{Z(t); t \in \mathbb{R}\}$

Suppose a random process $\{Z(t); t \in \mathbb{R}\}$ models the channel noise and $\{z(t) : \mathbb{R} \to \mathbb{R}\}$ is a sample function of this process. At first this seems inconsistent with the traditional elementary view that a random process or set of random variables models an experimental situation a priori (before performing the experiment) and the sample function models the result a posteriori (after performing the experiment). The trouble here is that the experiment might run from $t = -\infty$ to $t = \infty$, so there can be no "before" for the experiment and "after" for the result.

There are two ways out of this perceived inconsistency. First, the notion of "before and after" in the elementary view is inessential; the only important thing is the view

[1] Since a random variable is a mapping from Ω to $\mathbb{R}$, the sample values of a rv are real and thus the sample functions of a random process are real. It is often important to define objects called complex random variables that map Ω to $\mathbb{C}$. One can then define a complex random process as a process that maps each $t \in \mathbb{R}$ into a complex rv. These complex random processes will be important in studying noise waveforms at baseband.

that a multiplicity of sample functions might occur, but only one actually does. This point of view is appropriate in designing a cellular telephone for manufacture. Each individual phone that is sold experiences its own noise waveform, but the device must be manufactured to work over the multiplicity of such waveforms.

Second, whether we view a function of time as going from $-\infty$ to $+\infty$ or going from some large negative to large positive time is a matter of mathematical convenience. We often model waveforms as persisting from $-\infty$ to $+\infty$, but this simply indicates a situation in which the starting time and ending time are sufficiently distant to be irrelevant.

In order to specify a random process $\{Z(t); t \in \mathbb{R}\}$, some kind of rule is required from which joint distribution functions can, at least in principle, be calculated. That is, for all positive integers n, and all choices of n epochs $t_1, t_2, \ldots, t_n$ it must be possible (in principle) to find the joint distribution function,

$$F_{Z(t_1),\ldots,Z(t_n)}(z_1,\ldots,z_n) = \Pr\{Z(t_1) \le z_1, \ldots, Z(t_n) \le z_n\}, \tag{7.1}$$

for all choices of the real numbers $z_1, \ldots, z_n$. Equivalently, if densities exist, it must be possible (in principle) to find the joint density,

$$f_{Z(t_1),\ldots,Z(t_n)}(z_1,\ldots,z_n) = \frac{\partial^n F_{Z(t_1),\ldots,Z(t_n)}(z_1,\ldots,z_n)}{\partial z_1 \cdots \partial z_n}, \tag{7.2}$$

for all real $z_1, \ldots, z_n$. Since n can be arbitrarily large in (7.1) and (7.2), it might seem difficult for a simple rule to specify all these quantities, but a number of simple rules are given in the following examples that specify all these quantities.

7.2.1 Examples of random processes

The following generic example will turn out to be both useful and quite general. We saw earlier that we could specify waveforms by the sequence of coefficients in an orthonormal expansion. In the following example, a random process is similarly specified by a sequence of random variables used as coefficients in an orthonormal expansion.

Example 7.2.1 Let $Z_1, Z_2, \ldots$ be a sequence of random variables (rvs) defined on some sample space Ω and let $\{\phi_1(t)\}, \{\phi_2(t)\}, \ldots$ be a sequence of orthogonal (or orthonormal) real functions. For each $t \in \mathbb{R}$, let the rv $Z(t)$ be defined as $Z(t) = \sum_k Z_k \phi_k(t)$. The corresponding random process is then $\{Z(t); t \in \mathbb{R}\}$. For each t, $Z(t)$ is simply a sum of rvs, so we could, in principle, find its distribution function. Similarly, for each n-tuple $t_1, \ldots, t_n$ of epochs, $Z(t_1), \ldots, Z(t_n)$ is an n-tuple of rvs whose joint distribution could be found in principle. Since $Z(t)$ is a countably infinite sum of rvs, $\sum_{k=1}^{\infty} Z_k \phi_k(t)$, there might be some mathematical intricacies in finding, or even defining, its distribution function. Fortunately, as will be seen, such intricacies do not arise in the processes of most interest here.

It is clear that random processes can be defined as in the above example, but it is less clear that this will provide a mechanism for constructing reasonable models of actual physical noise processes. For the case of Gaussian processes, which will be defined shortly, this class of models will be shown to be broad enough to provide a flexible set of noise models.

The following few examples specialize the above example in various ways.

Example 7.2.2 Consider binary PAM, but view the input signals as independent identically distributed (iid) rvs $U_1, U_2, \ldots$ which take on the values ±1 with probability 1/2 each. Assume that the modulation pulse is $\operatorname{sinc}(t/T)$ so the baseband random process is given by

$$U(t) = \sum_k U_k \operatorname{sinc}\left(\frac{t-kT}{T}\right).$$

At each sampling epoch kT, the rv $U(kT)$ is simply the binary rv U_k. At epochs between the sampling epochs, however, $U(t)$ is a countably infinite sum of binary rvs whose variance will later be shown to be 1, but whose distribution function is quite ugly and not of great interest.

Example 7.2.3 A random variable is said to be zero-mean Gaussian if it has the probability density

$$f_Z(z) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-z^2}{2\sigma^2}\right), \tag{7.3}$$

where σ^2 is the variance of Z. A common model for a noise process $\{Z(t); t \in \mathbb{R}\}$ arises by letting

$$Z(t) = \sum_k Z_k \operatorname{sinc}\left(\frac{t-kT}{T}\right), \tag{7.4}$$

where $\ldots, Z_{-1}, Z_0, Z_1, \ldots$ is a sequence of iid zero-mean Gaussian rvs of variance σ^2. At each sampling epoch kT, the rv $Z(kT)$ is the zero-mean Gaussian rv Z_k. At epochs between the sampling epochs, $Z(t)$ is a countably infinite sum of independent zero-mean Gaussian rvs, which turns out to be itself zero-mean Gaussian of variance σ^2. Section 7.3 considers sums of Gaussian rvs and their interrelations in detail. The sample functions of this random process are simply sinc expansions and are limited to the baseband $[-1/2T, 1/2T]$. This example, as well as Example 7.2.2, brings out the following mathematical issue: the expected energy in $\{Z(t); t \in \mathbb{R}\}$ turns out to be infinite. As discussed later, this energy can be made finite either by truncating $Z(t)$ to some finite interval much larger than any time of interest or by similarly truncating the sequence $\{Z_k; k \in \mathbb{Z}\}$.

Another slightly disturbing aspect of this example is that this process cannot be "generated" by a sequence of Gaussian rvs entering a generating device that multiplies them by T-spaced sinc functions and adds them. The problem is the same as the problem with sinc functions in Chapter 6: they extend forever, and thus the process cannot be generated with finite delay. This is not of concern here, since we are not trying to generate random processes, only to show that interesting processes can be defined. The approach here will be to define and analyze a wide variety of random processes, and then to see which are useful in modeling physical noise processes.

Example 7.2.4 Let $\{Z(t); t \in [-1, 1]\}$ be defined by $Z(t) = tZ$ for all $t \in [-1, 1]$, where Z is a zero-mean Gaussian rv of variance 1. This example shows that random processes can be very degenerate; a sample function of this process is fully specified by the sample value $z(t)$ at $t = 1$. The sample functions are simply straight lines through the origin with random slope. This illustrates that the sample functions of a random process do not necessarily "look" random.

7.2.2 The mean and covariance of a random process

Often the first thing of interest about a random process is the mean at each epoch t and the covariance between any two epochs t and τ. The mean, $\mathsf{E}[Z(t)] = \overline{Z}(t)$, is simply a real-valued function of t, and can be found directly from the distribution function $F_{Z(t)}(z)$ or density $f_{Z(t)}(z)$. It can be verified that $\overline{Z}(t)$ is 0 for all t for Examples 7.2.2, 7.2.3, and 7.2.4. For Example 7.2.1, the mean cannot be specified without specifying more about the random sequence and the orthogonal functions.

The covariance[2] is a real-valued function of the epochs t and τ. It is denoted by $\mathsf{K}_Z(t,\tau)$ and defined by

$$\mathsf{K}_Z(t,\tau) = \mathsf{E}\left[\left(Z(t) - \overline{Z}(t)\right)\left(Z(\tau) - \overline{Z}(\tau)\right)\right]. \tag{7.5}$$

This can be calculated (in principle) from the joint distribution function $F_{Z(t),Z(\tau)}(z_1, z_2)$ or from the density $f_{Z(t),Z(\tau)}(z_1, z_2)$. To make the covariance function look a little simpler, we usually split each random variable $Z(t)$ into its mean, $\overline{Z}(t)$, and its fluctuation, $\widetilde{Z}(t) = Z(t) - \overline{Z}(t)$. The covariance function is then given by

$$\mathsf{K}_Z(t,\tau) = \mathsf{E}\left[\widetilde{Z}(t)\widetilde{Z}(\tau)\right]. \tag{7.6}$$

The random processes of most interest to us are used to model noise waveforms and usually have zero mean, in which case $Z(t) = \widetilde{Z}(t)$. In other cases, it often aids intuition to separate the process into its mean (which is simply an ordinary function) and its fluctuation, which by definition has zero mean.

The covariance function for the generic random process in Example 7.2.1 can be written as follows:

$$\mathsf{K}_Z(t,\tau) = \mathsf{E}\left[\sum_k \widetilde{Z}_k \phi_k(t) \sum_m \widetilde{Z}_m \phi_m(\tau)\right]. \tag{7.7}$$

If we assume that the rvs $Z_1, Z_2, \ldots$ are iid with variance σ^2, then $\mathsf{E}[\widetilde{Z}_k \widetilde{Z}_m] = 0$ for all $k \neq m$ and $\mathsf{E}[\widetilde{Z}_k \widetilde{Z}_m] = \sigma^2$ for $k = m$. Thus, ignoring convergence questions, (7.7) simplifies to

$$\mathsf{K}_Z(t,\tau) = \sigma^2 \sum_k \phi_k(t)\phi_k(\tau). \tag{7.8}$$

[2] This is often called the *autocovariance* to distinguish it from the covariance between two processes; we will not need to refer to this latter type of covariance.

For the sampling expansion, where $\phi_k(t) = \text{sinc}(t/T - k)$, it can be shown (see (7.48)) that the sum in (7.8) is simply $\text{sinc}[(t-\tau)/T]$. Thus for Examples 7.2.2 and 7.2.3, the covariance is given by

$$\mathsf{K}_Z(t, \tau) = \sigma^2 \,\text{sinc}\left(\frac{t-\tau}{T}\right)$$

where $\sigma^2 = 1$ for the binary PAM case of Example 7.2.2. Note that this covariance depends only on $t - \tau$ and not on the relationship between t or τ and the sampling points kT. These sampling processes are considered in more detail later.

7.2.3 Additive noise channels

The communication channels of greatest interest to us are known as *additive noise channels*. Both the channel input and the noise are modeled as random processes, $\{X(t);\ t \in \mathbb{R}\}$ and $\{Z(t);\ t \in \mathbb{R}\}$, both on the same underlying sample space Ω. The channel output is another random process, $\{Y(t);\ t \in \mathbb{R}\}$ and $Y(t) = X(t) + Z(t)$. This means that, for each epoch t, the random variable $Y(t)$ is equal to $X(t) + Z(t)$.

Note that one could always define the noise on a channel as the difference $Y(t) - X(t)$ between output and input. The notion of *additive noise* inherently also includes the assumption that the processes $\{X(t);\ t \in \mathbb{R}\}$ and $\{Z(t);\ t \in \mathbb{R}\}$ are statistically independent.[3]

As discussed earlier, the additive noise model $Y(t) = X(t) + Z(t)$ implicitly assumes that the channel attenuation, propagation delay, and carrier frequency and phase are perfectly known and compensated for. It also assumes that the input waveform is not changed by any disturbances other than the noise $Z(t)$.

Additive noise is most frequently modeled as a Gaussian process, as discussed in Section 7.3. Even when the noise is not modeled as Gaussian, it is often modeled as some modification of a Gaussian process. Many rules of thumb in engineering and statistics about noise are stated without any mention of Gaussian processes, but often are valid only for Gaussian processes.

7.3 Gaussian random variables, vectors, and processes

This section first defines Gaussian random variables (rvs), then jointly Gaussian random vectors (rvs), and finally Gaussian random processes. The covariance function and joint density function for Gaussian rvs are then derived. Finally, several equivalent conditions for rvs to be jointly Gaussian are derived.

A rv W is a *normalized Gaussian* rv, or more briefly a *normal*[4] rv, if it has the probability density

$$f_w(w) = \frac{1}{\sqrt{2\pi}} \exp\left(\frac{-w^2}{2}\right).$$

[3] More specifically, this means that, for all $k > 0$, all epochs $t_1, \ldots, t_k$ and all epochs $\tau_1, \ldots, \tau_k$ the rvs $X(t_1), \ldots, X(t_k)$ are statistically independent of $Z(\tau_1), \ldots, Z(\tau_k)$.

[4] Some people use normal rv as a synonym for Gaussian rv.

This density is symmetric around 0, and thus the mean of W is 0. The variance is 1, which is probably familiar from elementary probability and is demonstrated in Exercise 7.1. A random variable Z is a *Gaussian* rv if it is a scaled and shifted version of a normal rv, i.e. if $Z = \sigma W + \bar{Z}$ for a normal rv W. It can be seen that $\bar{Z}$ is the mean of Z and σ^2 is the variance.[5] The density of Z (for $\sigma^2 > 0$) is given by

$$f_z(z) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-(z-\bar{Z})^2}{2\sigma^2}\right). \tag{7.9}$$

A Gaussian rv Z of mean $\bar{Z}$ and variance σ^2 is denoted by $Z \sim \mathcal{N}(\bar{Z}, \sigma^2)$. The Gaussian rvs used to represent noise almost invariably have zero mean. Such rvs have the density $f_Z(z) = (1/\sqrt{2\pi\sigma^2}) \exp(-z^2/2\sigma^2)$, and are denoted by $Z \sim \mathcal{N}(0, \sigma^2)$.

Zero-mean Gaussian rvs are important in modeling noise and other random phenomena for the following reasons:

- they serve as good approximations to the sum of many independent zero-mean rvs (recall the central limit theorem);
- they have a number of extremal properties – as discussed later, they are, in several senses, the most random rvs for a given variance;
- they are easy to manipulate analytically, given a few simple properties;
- they serve as representative channel noise models, which provide insight about more complex models.

Definition 7.3.1 A set of n random variables $Z_1, \ldots, Z_n$ is *zero-mean jointly Gaussian* if there is a set of iid normal rvs $W_1, \ldots, W_\ell$ such that each Z_k, $1 \le k \le n$, can be expressed as

$$Z_k = \sum_{m=1}^{\ell} a_{km} W_m; \qquad 1 \le k \le n, \tag{7.10}$$

where $\{a_{km};\ 1 \le k \le n, 1 \le m \le \ell\}$ is an array of real numbers. More generally, $Z'_1, \ldots, Z'_n$ are *jointly Gaussian* if $Z'_k = Z_k + \bar{Z}'_k$, where the set $Z_1, \ldots, Z_n$ is zero-mean jointly Gaussian and $\bar{Z}'_1, \ldots, \bar{Z}'_n$ is a set of real numbers.

It is convenient notationally to refer to a set of n random variables $Z_1, \ldots, Z_n$ as a random vector[6] (rv) $Z = (Z_1, \ldots, Z_n)^{\mathsf{T}}$. Letting A be the n by ℓ real matrix with elements $\{a_{km};\ 1 \le k \le n, 1 \le m \le \ell\}$, (7.10) can then be represented more compactly as

$$Z = \mathsf{A} W, \tag{7.11}$$

where W is an ℓ-tuple of iid normal rvs. Similarly, the jointly Gaussian random vector Z' can be represented as $Z' = \mathsf{A} Z + \bar{Z}'$, where $\bar{Z}'$ is an n-vector of real numbers.

[5] It is convenient to define Z to be Gaussian even in the deterministic case where $\sigma = 0$, but then (7.9) is invalid.

[6] The class of random vectors for a given n over a given sample space satisfies the axioms of a vector space, but here the vector notation is used simply as a notational convenience.

In the remainder of this chapter, all random variables, random vectors, and random processes are assumed to be zero-mean unless explicitly designated otherwise. In other words, only the fluctuations are analyzed, with the means added at the end.[7]

It is shown in Exercise 7.2 that any sum $\sum_m a_{km} W_m$ of iid normal rvs $W_1, \dots, W_n$ is a Gaussian rv, so that each Z_k in (7.10) is Gaussian. Jointly Gaussian means much more than this, however. The random variables $Z_1, \dots, Z_n$ must also be related as linear combinations of the same set of iid normal variables. Exercises 7.3 and 7.4 illustrate some examples of pairs of random variables which are individually Gaussian but not jointly Gaussian. These examples are slightly artificial, but illustrate clearly that the joint density of jointly Gaussian rvs is much more constrained than the possible joint densities arising from constraining marginal distributions to be Gaussian.

The definition of jointly Gaussian looks a little contrived at first, but is in fact very natural. Gaussian rvs often make excellent models for physical noise processes because noise is often the summation of many small effects. The central limit theorem is a mathematically precise way of saying that the sum of a very large number of independent small zero-mean random variables is approximately zero-mean Gaussian. Even when different sums are statistically dependent on each other, they are different linear combinations of a common set of independent small random variables. Thus, the jointly Gaussian assumption is closely linked to the assumption that the noise is the sum of a large number of small, essentially independent, random disturbances. Assuming that the underlying variables are Gaussian simply makes the model analytically clean and tractable.

An important property of any jointly Gaussian n-dimensional rv Z is the following: for any real m by n real matrix B, the rv $Y = \mathsf{B}Z$ is also jointly Gaussian. To see this, let $Z = \mathsf{A}W$, where W is a normal rv. Then

$$Y = \mathsf{B}Z = \mathsf{B}(\mathsf{A}W) = (\mathsf{BA})W. \tag{7.12}$$

Since BA is a real matrix, Y is jointly Gaussian. A useful application of this property arises when A is diagonal, so Z has arbitrary independent Gaussian components. This implies that $Y = \mathsf{B}Z$ is jointly Gaussian whenever a rv Z has independent Gaussian components.

Another important application is where B is a 1 by n matrix and Y is a random variable. Thus every linear combination $\sum_{k=1}^{n} b_k Z_k$ of a jointly Gaussian rv $Z = (Z_1, \dots, Z_n)^{\mathsf{T}}$ is Gaussian. It will be shown later in this section that this is an if and only if property; that is, if every linear combination of a rv Z is Gaussian, then Z is jointly Gaussian.

We now have the machinery to define zero-mean Gaussian processes.

Definition 7.3.2 $\{Z(t); t \in \mathbb{R}\}$ is a *zero-mean Gaussian process* if, for all positive integers n and all finite sets of epochs $t_1, \dots, t_n$, the set of random variables $Z(t_1), \dots, Z(t_n)$ is a (zero-mean) jointly Gaussian set of random variables.

[7] When studying estimation and conditional probabilities, means become an integral part of many arguments, but these arguments will not be central here.

If the covariance, $\mathsf{K}_Z(t,\tau) = \mathsf{E}[Z(t)Z(\tau)]$, is known for each pair of epochs t, τ, then, for any finite set of epochs $t_1, \dots, t_n$, $\mathsf{E}[Z(t_k)Z(t_m)]$ is known for each pair (t_k, t_m) in that set. Sections 7.3.1 and 7.3.2 will show that the joint probability density for any such set of (zero-mean) jointly Gaussian rvs depends only on the covariances of those variables. This will show that a zero-mean Gaussian process is specified by its covariance function. A nonzero-mean Gaussian process is similarly specified by its covariance function and its mean.

7.3.1 The covariance matrix of a jointly Gaussian random vector

Let an n-tuple of (zero-mean) rvs $Z_1, \dots, Z_n$ be represented as a rv $\boldsymbol{Z} = (Z_1, \dots, Z_n)^{\mathsf{T}}$. As defined earlier, $\boldsymbol{Z}$ is jointly Gaussian if $\boldsymbol{Z} = \mathsf{A}\boldsymbol{W}$, where $\boldsymbol{W} = (W_1, W_2, \dots, W_\ell)^{\mathsf{T}}$ is a vector of iid normal rvs and A is an n by ℓ real matrix. Each rv Z_k, and all linear combinations of $Z_1, \dots, Z_n$, are Gaussian.

The covariance of two (zero-mean) rvs Z_1, Z_2 is $\mathsf{E}[Z_1Z_2]$. For a rv $\boldsymbol{Z} = (Z_1, \dots Z_n)^{\mathsf{T}}$ the covariance between all pairs of random variables is very conveniently represented by the n by n covariance matrix

$$\mathsf{K}_{\boldsymbol{Z}} = \mathsf{E}[\boldsymbol{Z}\boldsymbol{Z}^{\mathsf{T}}].$$

Appendix 7.11.1 develops a number of properties of covariance matrices (including the fact that they are identical to the class of nonnegative definite matrices). For a vector $\boldsymbol{W} = W_1, \dots, W_\ell$ of independent normalized Gaussian rvs, $\mathsf{E}[W_jW_m] = 0$ for $j \neq m$ and 1 for $j = m$. Thus

$$\mathsf{K}_{\boldsymbol{W}} = \mathsf{E}[\boldsymbol{W}\boldsymbol{W}^{\mathsf{T}}] = \mathsf{I}_\ell,$$

where I_ℓ is the ℓ by ℓ identity matrix. For a zero-mean jointly Gaussian vector $\boldsymbol{Z} = \mathsf{A}\boldsymbol{W}$, the covariance matrix is thus given by

$$\mathsf{K}_{\boldsymbol{Z}} = \mathsf{E}[\mathsf{A}\boldsymbol{W}\boldsymbol{W}^{\mathsf{T}}\mathsf{A}^{\mathsf{T}}] = \mathsf{A}\mathsf{E}[\boldsymbol{W}\boldsymbol{W}^{\mathsf{T}}]\mathsf{A}^{\mathsf{T}} = \mathsf{A}\mathsf{A}^{\mathsf{T}}. \tag{7.13}$$

7.3.2 The probability density of a jointly Gaussian random vector

The *probability density*, $f_{\boldsymbol{Z}}(\boldsymbol{z})$, of a rv $\boldsymbol{Z} = (Z_1, Z_2, \dots, Z_n)^{\mathsf{T}}$ is the joint probability density of the components $Z_1, \dots, Z_n$. An important example is the iid rv $\boldsymbol{W}$, where the components W_k, $1 \le k \le n$, are iid and normal, $W_k \sim \mathcal{N}(0,1)$. By taking the product of the n densities of the individual rvs, the density of $\boldsymbol{W} = (W_1, W_2, \dots, W_n)^{\mathsf{T}}$ is given by

$$f_{\boldsymbol{W}}(\boldsymbol{w}) = \frac{1}{(2\pi)^{n/2}} \exp\left(\frac{-w_1^2 - w_2^2 - \cdots - w_n^2}{2}\right) = \frac{1}{(2\pi)^{n/2}} \exp\left(\frac{-\|\boldsymbol{w}\|^2}{2}\right). \tag{7.14}$$

This shows that the density of W at a sample value w depends only on the squared distance $\|w\|^2$ of the sample value from the origin. That is, $f_W(w)$ is spherically symmetric around the origin, and points of equal probability density lie on concentric spheres around the origin.

Consider the transformation $Z = AW$, where Z and W each have n components and A is n by n. If we let $a_1, a_2, \dots, a_n$ be the n columns of A, then this means that $Z = \sum_m a_m W_m$. That is, for any sample values $w_1, \dots, w_n$ for W, the corresponding sample value for Z is $z = \sum_m a_m w_m$. Similarly, if we let $b_1, \dots, b_n$ be the rows of A, then $Z_k = b_k W$.

Let $\mathcal{B}_\delta$ be a cube, δ on a side, of the sample values of W defined by $\mathcal{B}_\delta = \{w : 0 \le w_k \le \delta;\ 1 \le k \le n\}$ (see Figure 7.1). The set $\mathcal{B}'_\delta$ of vectors $z = \mathsf{A}w$ for $w \in \mathcal{B}_\delta$ is a parallelepiped whose sides are the vectors $\delta a_1, \dots, \delta a_n$. The determinant, det(A), of A has the remarkable geometric property that its magnitude, $|\det(\mathsf{A})|$, is equal to the volume of the parallelepiped with sides a_k; $1 \le k \le n$. Thus the unit cube $\mathcal{B}_\delta$, with volume δ^n, is mapped by A into a parallelepiped of volume $|\det \mathsf{A}|\delta^n$.

Assume that the columns $a_1, \dots, a_n$ of A are linearly independent. This means that the columns must form a basis for $\mathbb{R}^n$, and thus that every vector z is some linear combination of these columns, i.e. that $z = \mathsf{A}w$ for some vector w. The matrix A must then be invertible, i.e. there is a matrix A^{-1} such that $\mathsf{A}\mathsf{A}^{-1} = \mathsf{A}^{-1}\mathsf{A} = \mathsf{I}_n$, where I_n is the n by n identity matrix. The matrix A maps the unit vectors of $\mathbb{R}^n$ into the vectors $a_1, \dots, a_n$ and the matrix A^{-1} maps $a_1, \dots, a_n$ back into the unit vectors.

If the columns of A are not linearly independent, i.e. A is not invertible, then A maps the unit cube in $\mathbb{R}^n$ into a subspace of dimension less than n. In terms of Figure 7.1, the unit cube would be mapped into a straight line segment. The area, in 2D space, of a straight line segment is 0, and more generally the volume in n-space of any lower-dimensional set of points is 0. In terms of the determinant, $\det \mathsf{A} = 0$ for any noninvertible matrix A.

Assuming again that A is invertible, let z be a sample value of Z and let $w = \mathsf{A}^{-1}z$ be the corresponding sample value of W. Consider the incremental cube $w + \mathcal{B}_\delta$ cornered at w. For δ very small, the probability $P_\delta(w)$ that W lies in this cube is $f_W(w)\delta^n$ plus terms that go to zero faster than δ^n as $\delta \to 0$. This cube around w maps into a parallelepiped of volume $\delta^n|\det(\mathsf{A})|$ around z, and no other sample value of W maps into this parallelepiped. Thus $P_\delta(w)$ is also equal to $f_Z(z)\delta^n|\det(\mathsf{A})|$

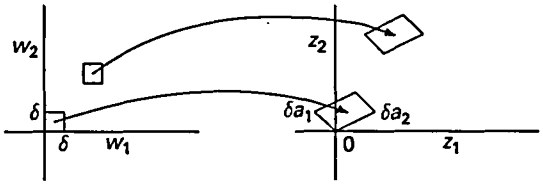


Figure 7.1. Example illustrating how $Z = AW$ maps cubes into parallelepipeds. Let $Z_1 = -W_1 + 2W_2$ and $Z_2 = W_1 + W_2$. The figure shows the set of sample pairs z_1, z_2 corresponding to $0 \le w_1 \le \delta$ and $0 \le w_2 \le \delta$. It also shows a translation of the same cube mapping into a translation of the same parallelepiped.

plus negligible terms. Going to the limit $\delta \to 0$, we have

$$f_Z(z)|\det(\mathsf{A})| = \lim_{\delta\to 0} \frac{P_\delta(w)}{\delta^n} = f_W(w). \tag{7.15}$$

Since $w = \mathsf{A}^{-1}z$, we obtain the explicit formula

$$f_Z(z) = \frac{f_W(\mathsf{A}^{-1}z)}{|\det(\mathsf{A})|}. \tag{7.16}$$

This formula is valid for any random vector W with a density, but we are interested in the vector W of iid Gaussian rvs, $\mathcal{N}(0,1)$. Substituting (7.14) into (7.16), we obtain

$$f_Z(z) = \frac{1}{(2\pi)^{n/2}|\det(\mathsf{A})|} \exp\left(\frac{-\|\mathsf{A}^{-1}z\|^2}{2}\right), \tag{7.17}$$

$$= \frac{1}{(2\pi)^{n/2}|\det(\mathsf{A})|} \exp\left(-\frac{1}{2}z^{\mathsf{T}}(\mathsf{A}^{-1})^{\mathsf{T}}\mathsf{A}^{-1}z\right). \tag{7.18}$$

We can simplify this somewhat by recalling from (7.13) that the covariance matrix of Z is given by $\mathsf{K}_Z = \mathsf{A}\mathsf{A}^{\mathsf{T}}$. Thus, $\mathsf{K}_Z^{-1} = (\mathsf{A}^{-1})^{\mathsf{T}}\mathsf{A}^{-1}$. Substituting this into (7.18) and noting that $\det(\mathsf{K}_Z) = |\det(\mathsf{A})|^2$, we obtain

$$f_Z(z) = \frac{1}{(2\pi)^{n/2}\sqrt{\det(\mathsf{K}_Z)}} \exp\left(-\frac{1}{2}z^{\mathsf{T}}\mathsf{K}_Z^{-1}z\right). \tag{7.19}$$

Note that this probability density depends only on the covariance matrix of Z and not directly on the matrix A.

The density in (7.19) does rely, however, on A being nonsingular. If A is singular, then at least one of its rows is a linear combination of the other rows, and thus, for some m, $1 \le m \le n$, Z_m is a linear combination of the other Z_k. The random vector Z is still jointly Gaussian, but the joint probability density does not exist (unless one wishes to view the density of Z_m as a unit impulse at a point specified by the sample values of the other variables). It is possible to write out the distribution function for this case, using step functions for the dependent rvs, but it is not worth the notational mess. It is more straightforward to face the problem and find the density of a maximal set of linearly independent rvs, and specify the others as deterministic linear combinations.

It is important to understand that there is a large difference between rvs being *statistically dependent* and *linearly dependent*. If they are linearly dependent, then one or more are deterministic functions of the others, whereas statistical dependence simply implies a probabilistic relationship.

These results are summarized in the following theorem.

Theorem 7.3.1 (Density for jointly Gaussian rvs) *Let Z be a (zero-mean) jointly Gaussian rv with a nonsingular covariance matrix K_Z. Then the probability density $f_Z(z)$ is given by (7.19). If K_Z is singular, then $f_Z(z)$ does not exist, but the density in (7.19) can be applied to any set of linearly independent rvs out of $Z_1, \dots, Z_n$.*

For a zero-mean Gaussian process $Z(t)$, the covariance function $\mathsf{K}_Z(t, \tau)$ specifies $\mathsf{E}[Z(t_k)Z(t_m)]$ for arbitrary epochs t_k and t_m and thus specifies the covariance matrix for any finite set of epochs $t_1, \ldots, t_n$. From Theorem was 7.3.1, this also specifies the joint probability distribution for that set of epochs. Thus the covariance function specifies all joint probability distributions for all finite sets of epochs, and thus specifies the process in the sense[8] of Section 7.2. In summary, we have the following important theorem.

Theorem 7.3.2 (Gaussian process) *A zero-mean Gaussian process is specified by its covariance function* $\mathsf{K}(t, \tau)$.

7.3.3 Special case of a 2D zero-mean Gaussian random vector

The probability density in (7.19) is now written out in detail for the 2D case. Let $\mathsf{E}[Z_1^2] = \sigma_1^2$, $\mathsf{E}[Z_2^2] = \sigma_2^2$, and $\mathsf{E}[Z_1 Z_2] = \kappa_{12}$. Thus

$$\mathsf{K}_Z = \begin{bmatrix} \sigma_1^2 & \kappa_{12} \\ \kappa_{12} & \sigma_2^2 \end{bmatrix}.$$

Let ρ be the *normalized covariance* $\rho = \kappa_{12}/(\sigma_1\sigma_2)$. Then $\det(\mathsf{K}_Z) = \sigma_1^2\sigma_2^2 - \kappa_{12}^2 = \sigma_1^2\sigma_2^2(1-\rho^2)$. Note that ρ must satisfy $|\rho| \le 1$ with strict inequality if K_Z is nonsingular:

$$\mathsf{K}_Z^{-1} = \frac{1}{\sigma_1^2\sigma_2^2 - \kappa_{12}^2}\begin{bmatrix} \sigma_2^2 & -\kappa_{12} \\ -\kappa_{12} & \sigma_1^2 \end{bmatrix} = \frac{1}{1-\rho^2}\begin{bmatrix} 1/\sigma_1^2 & -\rho/(\sigma_1\sigma_2) \\ -\rho/(\sigma_1\sigma_2) & 1/\sigma_2^2 \end{bmatrix};$$

$$\begin{aligned} f_Z(z) &= \frac{1}{2\pi\sqrt{\sigma_1^2\sigma_2^2 - \kappa_{12}^2}}\exp\left(\frac{-z_1^2\sigma_2^2 + 2z_1z_2\kappa_{12} - z_2^2\sigma_1^2}{2(\sigma_1^2\sigma_2^2 - \kappa_{12}^2)}\right) \\ &= \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}}\exp\left(\frac{-(z_1/\sigma_1)^2 + 2\rho(z_1/\sigma_1)(z_2/\sigma_2) - (z_2/\sigma_2)^2}{2(1-\rho^2)}\right). \end{aligned} \tag{7.20}$$

Curves of equal probability density in the plane correspond to points where the argument of the exponential function in (7.20) is constant. This argument is quadratic, and thus points of equal probability density form an ellipse centered on the origin. The ellipses corresponding to different values of probability density are concentric, with larger ellipses corresponding to smaller densities.

If the normalized covariance ρ is 0, the axes of the ellipse are the horizontal and vertical axes of the plane; if $\sigma_1 = \sigma_2$, the ellipse reduces to a circle; and otherwise the ellipse is elongated in the direction of the larger standard deviation. If $\rho > 0$, the density in the first and third quadrants is increased at the expense of the second

[8] As will be discussed later, focusing on the pointwise behavior of a random process at all finite sets of epochs has some of the same problems as specifying a function pointwise rather than in terms of $\mathcal{L}_2$-equivalence. This can be ignored for the present.

and fourth, and thus the ellipses are elongated in the first and third quadrants. This is reversed, of course, for $\rho < 0$.

The main point to be learned from this example, however, is that the detailed expression for two dimensions in (7.20) is messy. The messiness gets far worse in higher dimensions. Vector notation is almost essential. One should reason directly from the vector equations and use standard computer programs for calculations.

7.3.4 $Z = \mathsf{A}W$, where A is orthogonal

An n by n real matrix A for which $\mathsf{A}\mathsf{A}^{\mathsf{T}} = \mathsf{I}_n$ is called an *orthogonal matrix* or *orthonormal matrix* (orthonormal is more appropriate, but orthogonal is more common). For $Z = \mathsf{A}W$, where W is iid normal and A is orthogonal, $\mathsf{K}_Z = \mathsf{A}\mathsf{A}^{\mathsf{T}} = \mathsf{I}_n$. Thus $\mathsf{K}_Z^{-1} = \mathsf{I}_n$ also, and (7.19) becomes

$$f_Z(z) = \frac{\exp(-(1/2)z^{\mathsf{T}}z)}{(2\pi)^{n/2}} = \prod_{k=1}^{n} \frac{\exp(-z_k^2/2)}{\sqrt{2\pi}}. \qquad (7.21)$$

This means that A transforms W into a random vector Z with the same probability density, and thus the components of Z are still normal and iid. To understand this better, note that $\mathsf{A}\mathsf{A}^{\mathsf{T}} = \mathsf{I}_n$ means that A^{T} is the inverse of A and thus that $\mathsf{A}^{\mathsf{T}}\mathsf{A} = \mathsf{I}_n$. Letting a_m be the mth column of A, the equation $\mathsf{A}^{\mathsf{T}}\mathsf{A} = \mathsf{I}_n$ means that $a_m^{\mathsf{T}}a_j = \delta_{mj}$ for each m, j, $1 \le m, j \le n$, i.e. that the columns of A are orthonormal. Thus, for the 2D example, the unit vectors e_1, e_2 are mapped into orthonormal vectors a_1, a_2, so that the transformation simply rotates the points in the plane. Although it is difficult to visualize such a transformation in higher-dimensional space, it is still called a rotation, and has the property that $||\mathsf{A}w||^2 = w^{\mathsf{T}}\mathsf{A}^{\mathsf{T}}\mathsf{A}w$, which is just $w^{\mathsf{T}}w = ||w||^2$. Thus, each point w maps into a point $\mathsf{A}w$ at the same distance from the origin as itself.

Not only are the columns of an orthogonal matrix orthonormal, but also the rows, say $\{b_k;\ 1 \le k \le n\}$ are orthonormal (as is seen directly from $\mathsf{A}\mathsf{A}^{\mathsf{T}} = \mathsf{I}_n$). Since $Z_k = b_k W$, this means that, for any set of orthonormal vectors $b_1, \dots, b_n$, the random variables $Z_k = b_k W$ are normal and iid for $1 \le k \le n$.

7.3.5 Probability density for Gaussian vectors in terms of principal axes

This section describes what is often a more convenient representation for the probability density of an n-dimensional (zero-mean) Gaussian rv Z with a nonsingular covariance matrix K_Z. As shown in Appendix 7.11.1, the matrix K_Z has n real orthonormal eigenvectors, $q_1, \dots, q_n$, with corresponding nonnegative (but not necessarily distinct[9]) real eigenvalues, $\lambda_1, \dots, \lambda_n$. Also, for any vector z, it is shown that $z^{\mathsf{T}}\mathsf{K}_Z^{-1}z$ can be

[9] If an eigenvalue λ has multiplicity m, it means that there is an m-dimensional subspace of vectors q satisfying $\mathsf{K}_Z q = \lambda q$; in this case, any orthonormal set of m such vectors can be chosen as the m eigenvectors corresponding to that eigenvalue.

expressed as $\sum_k \lambda_k^{-1}|\langle z, q_k\rangle|^2$. Substituting this in (7.19), we have

$$f_Z(z) = \frac{1}{(2\pi)^{n/2}\sqrt{\det(K_Z)}} \exp\left(-\sum_k \frac{|\langle z, q_k\rangle|^2}{2\lambda_k}\right). \tag{7.22}$$

Note that $\langle z, q_k\rangle$ is the projection of z in the direction q_k, where q_k is the kth of n orthonormal directions. The determinant of an n by n real matrix can be expressed in terms of the n eigenvalues (see Appendix 7.11.1) as $\det(K_Z) = \prod_{k=1}^{n} \lambda_k$. Thus (7.22) becomes

$$f_Z(z) = \prod_{k=1}^{n} \frac{1}{\sqrt{2\pi\lambda_k}} \exp\left(\frac{-|\langle z, q_k\rangle|^2}{2\lambda_k}\right). \tag{7.23}$$

This is the product of n Gaussian densities. It can be interpreted as saying that the Gaussian rvs $\{\langle Z, q_k\rangle; 1 \le k \le n\}$ are statistically independent with variances $\{\lambda_k; 1 \le k \le n\}$. In other words, if we represent the rv Z using $q_1, \dots, q_n$ as a basis, then the components of Z in that coordinate system are independent random variables. The orthonormal eigenvectors are called *principal axes* for Z.

This result can be viewed in terms of the contours of equal probability density for Z (see Figure 7.2). Each such contour satisfies

$$c = \sum_k \frac{|\langle z, q_k\rangle|^2}{2\lambda_k},$$

where c is proportional to the log probability density for that contour. This is the equation of an ellipsoid centered on the origin, where q_k is the kth axis of the ellipsoid and $\sqrt{2c\lambda_k}$ is the length of that axis.

The probability density formulas in (7.19) and (7.23) suggest that, for every covariance matrix K, there is a jointly Gaussian rv that has that covariance, and thus has that probability density. This is in fact true, but to verify it we must demonstrate that for every covariance matrix K there is a matrix A (and thus a rv $Z = AW$) such that $K = AA^T$. There are many such matrices for any given K, but a particularly convenient one is given in (7.84). As a function of the eigenvectors and eigenvalues of K, it is $A = \sum_k \sqrt{\lambda_k} q_k q_k^T$. Thus, for every nonsingular covariance matrix K, there is a jointly Gaussian rv whose density satisfies (7.19) and (7.23).

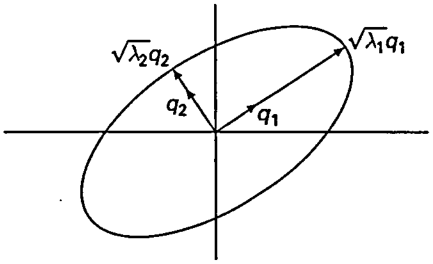


Figure 7.2. Contours of equal probability density. Points z on the q_1 axis are points for which $\langle z, q_2\rangle = 0$ and points on the q_2 axis satisfy $\langle z, q_1\rangle = 0$. Points on the illustrated ellipse satisfy $z^T K_Z^{-1} z = 1$.

7.3.6 Fourier transforms for joint densities

As suggested in Exercise 7.2, Fourier transforms of probability densities are useful for finding the probability density of sums of independent random variables. More generally, for an n-dimensional rv Z, we can define the n-dimensional Fourier transform of $f_Z(z)$ as follows:

$$\hat{f}_Z(s) = \int \cdots \int f_Z(z) \exp(-2\pi i s^\mathsf{T} z)\, dz_1 \cdots dz_n = \mathsf{E}[\exp(-2\pi i s^\mathsf{T} Z)]. \tag{7.24}$$

If Z is jointly Gaussian, this is easy to calculate. For any given $s \neq 0$, let $X = s^\mathsf{T} Z = \sum_k s_k Z_k$. Thus X is Gaussian with variance $\mathsf{E}[s^\mathsf{T} Z Z^\mathsf{T} s] = s^\mathsf{T} \mathsf{K}_Z s$. From Exercise 7.2,

$$\hat{f}_X(\theta) = \mathsf{E}[\exp(-2\pi i \theta s^\mathsf{T} Z)] = \exp\left(-\frac{(2\pi\theta)^2 s^\mathsf{T} \mathsf{K}_Z s}{2}\right). \tag{7.25}$$

Comparing (7.25) for $\theta = 1$ with (7.24), we see that

$$\hat{f}_Z(s) = \exp\left(-\frac{(2\pi)^2 s^\mathsf{T} \mathsf{K}_Z s}{2}\right). \tag{7.26}$$

The above derivation also demonstrates that $\hat{f}_Z(s)$ is determined by the Fourier transform of each linear combination of the elements of Z. In other words, if an arbitrary rv Z has covariance K_Z and has the property that all linear combinations of Z are Gaussian, then the Fourier transform of its density is given by (7.26). Thus, assuming that the Fourier transform of the density uniquely specifies the density, Z must be jointly Gaussian if all linear combinations of Z are Gaussian.

A number of equivalent conditions have now been derived under which a (zero-mean) random vector Z is jointly Gaussian. In summary, each of the following are necessary and sufficient conditions for a rv Z with a nonsingular covariance K_Z to be jointly Gaussian:

- $Z = \mathsf{A}W$, where the components of W are iid normal and $\mathsf{K}_Z = \mathsf{A}\mathsf{A}^\mathsf{T}$;
- Z has the joint probability density given in (7.19);
- Z has the joint probability density given in (7.23);
- all linear combinations of Z are Gaussian random variables.

For the case where K_Z is singular, the above conditions are necessary and sufficient for any linearly independent subset of the components of Z.

This section has considered only zero-mean random variables, vectors, and processes. The results here apply directly to the fluctuation of arbitrary random variables, vectors, and processes. In particular, the probability density for a jointly Gaussian rv Z with a nonsingular covariance matrix K_Z and mean vector $\overline{Z}$ is given by

$$f_Z(z) = \frac{1}{(2\pi)^{n/2}\sqrt{\det(\mathsf{K}_Z)}} \exp\left(-\frac{1}{2}(z - \overline{Z})^\mathsf{T} \mathsf{K}_Z^{-1} (z - \overline{Z})\right). \tag{7.27}$$

7.4 Linear functionals and filters for random processes

This section defines the important concept of linear functionals of arbitrary random processes $\{Z(t); t \in \mathbb{R}\}$ and then specializes to Gaussian random processes, where the results of the Section 7.3 can be used. Assume that the sample functions $Z(t, \omega)$ of $Z(t)$ are real $\mathcal{L}_2$ waveforms. These sample functions can be viewed as vectors in the $\mathcal{L}_2$ space of real waveforms. For any given real $\mathcal{L}_2$ waveform $g(t)$, there is an inner product,

$$\langle Z(t, \omega), g(t) \rangle = \int_{-\infty}^{\infty} Z(t, \omega) g(t) \mathrm{d}t.$$

By the Schwarz inequality, the magnitude of this inner product in the space of real $\mathcal{L}_2$ functions is upperbounded by $\|Z(t, \omega)\| \|g(t)\|$ and is thus a finite real value for each ω. This then maps sample points ω into real numbers and is thus a random variable,[10] denoted by $V = \int_{-\infty}^{\infty} Z(t) g(t) \mathrm{d}t$. This rv V is called a *linear functional* of the process $\{Z(t); t \in \mathbb{R}\}$.

As an example of the importance of linear functionals, recall that the demodulator for both PAM and QAM contains a filter $q(t)$ followed by a sampler. The output of the filter at a sampling time kT for an input $u(t)$ is $\int u(t) q(kT - t) \mathrm{d}t$. If the filter input also contains additive noise $Z(t)$, then the output at time kT also contains the linear functional $\int Z(t) q(kT - t) \mathrm{d}t$.

Similarly, for any random process $\{Z(t); t \in \mathbb{R}\}$ (again assuming $\mathcal{L}_2$ sample functions) and any real $\mathcal{L}_2$ function $h(t)$, consider the result of passing $Z(t)$ through the filter with impulse response $h(t)$. For any $\mathcal{L}_2$ sample function $Z(t, \omega)$, the filter output at any given time τ is the inner product

$$\langle Z(t, \omega), h(\tau - t) \rangle = \int_{-\infty}^{\infty} Z(t, \omega) h(\tau - t) \mathrm{d}t.$$

For each real τ, this maps sample points ω into real numbers, and thus (aside from measure-theoretic issues)

$$V(\tau) = \int Z(t) h(\tau - t) \mathrm{d}t \tag{7.28}$$

is a rv for each τ. This means that $\{V(\tau); \tau \in \mathbb{R}\}$ is a random process. This is called the *filtered process* resulting from passing $Z(t)$ through the filter $h(t)$. Not much can be said about this general problem without developing a great deal of mathematics, so instead we restrict ourselves to Gaussian processes and other relatively simple examples.

For a Gaussian process, we would hope that a linear functional is a Gaussian random variable. The following examples show that some restrictions are needed even for the class of Gaussian processes.

[10] One should use measure theory over the sample space Ω to interpret these mappings carefully, but this is unnecessary for the simple types of situations here and would take us too far afield.

Example 7.4.1 Let $Z(t) = tX$ for all $t \in \mathbb{R}$, where $X \sim \mathcal{N}(0, 1)$. The sample functions of this Gaussian process have infinite energy with probability 1. The output of the filter also has infinite energy except for very special choices of $h(t)$.

Example 7.4.2 For each $t \in [0, 1]$, let $W(t)$ be a Gaussian rv, $W(t) \sim \mathcal{N}(0, 1)$. Assume also that $\mathsf{E}[W(t)W(\tau)] = 0$ for each $t \neq \tau \in [0, 1]$. The sample functions of this process are discontinuous everywhere.[11] We do not have the machinery to decide whether the sample functions are integrable, let alone whether the linear functionals above exist; we discuss this example further later.

In order to avoid the mathematical issues in Example 7.4.2, along with a host of other mathematical issues, we start with Gaussian processes defined in terms of orthonormal expansions.

7.4.1 Gaussian processes defined over orthonormal expansions

Let $\{\phi_k(t);\ k \geq 1\}$ be a countable set of real orthonormal functions and let $\{Z_k;\ k \geq 1\}$ be a sequence of independent Gaussian random variables, $\{\mathcal{N}(0, \sigma_k^2)\}$. Consider the Gaussian process defined by

$$Z(t) = \sum_{k=1}^{\infty} Z_k \phi_k(t). \tag{7.29}$$

Essentially all zero-mean Gaussian processes of interest can be defined this way, although we will not prove this. Clearly a mean can be added if desired, but zero-mean processes are assumed in what follows. First consider the simple case in which σ_k^2 is nonzero for only finitely many values of k, say $1 \leq k \leq n$. In this case, for each $t \in \mathbb{R}$, $Z(t)$ is a finite sum, given by

$$Z(t) = \sum_{k=1}^{n} Z_k \phi_k(t), \tag{7.30}$$

of independent Gaussian rvs and thus is Gaussian. It is also clear that $Z(t_1), Z(t_2), \ldots, Z(t_\ell)$ are jointly Gaussian for all $\ell, t_1, \ldots, t_\ell$, so $\{Z(t); t \in \mathbb{R}\}$ is in fact a Gaussian random process. The energy in any sample function, $z(t) = \sum_k z_k \phi_k(t)$, is $\sum_{k=1}^{n} z_k^2$. This is finite (since the sample values are real and thus finite), so every sample function is $\mathcal{L}_2$. The covariance function is then easily calculated to be

$$\mathsf{K}_Z(t, \tau) = \sum_{k,m} \mathsf{E}[Z_k Z_m]\phi_k(t)\phi_m(\tau) = \sum_{k=1}^{n} \sigma_k^2\, \phi_k(t)\phi_k(\tau). \tag{7.31}$$

Next consider the linear functional $\int Z(t)g(t)\,\mathrm{d}t$, where $g(t)$ is a real $\mathcal{L}_2$ function:

$$V = \int_{-\infty}^{\infty} Z(t)g(t)\mathrm{d}t = \sum_{k=1}^{n} Z_k \int_{-\infty}^{\infty} \phi_k(t)g(t)\mathrm{d}t. \tag{7.32}$$

[11] Even worse, the sample functions are not measurable. This process would not even be called a random process in a measure-theoretic formulation, but it provides an interesting example of the occasional need for a measure-theoretic formulation.

Since V is a weighted sum of the zero-mean independent Gaussian rvs $Z_1, \dots, Z_n$, V is also Gaussian with variance given by

$$\sigma_V^2 = \mathsf{E}[V^2] = \sum_{k=1}^{n} \sigma_k^2 |\langle \phi_k, g \rangle|^2. \tag{7.33}$$

Next consider the case where n is infinite but $\sum_k \sigma_k^2 < \infty$. The sample functions are still $\mathcal{L}_2$ (at least with probability 1). Equations (7.29) – (7.33) are still valid, and Z is still a Gaussian rv. We do not have the machinery to prove this easily, although Exercise 7.7 provides quite a bit of insight into why these results are true.

Finally, consider a finite set of $\mathcal{L}_2$ waveforms $\{g_m(t);\ 1 \le m \le \ell\}$ and let $V_m = \int_{-\infty}^{\infty} Z(t) g_m(t)\, dt$. By the same argument as above, V_m is a Gaussian rv for each m. Furthermore, since each linear combination of these variables is also a linear functional, it is also Gaussian, so $\{V_1, \dots, V_\ell\}$ is jointly Gaussian.

7.4.2 Linear filtering of Gaussian processes

We can use the same argument as in Section 7.4.1 to look at the output of a linear filter (see Figure 7.3) for which the input is a Gaussian process $\{Z(t);\ t \in \mathbb{R}\}$. In particular, assume that $Z(t) = \sum_k Z_k \phi_k(t)$, where $Z_1, Z_2, \dots$ is an independent sequence $\{Z_k \sim \mathcal{N}(0, \sigma_k^2\}$ satisfying $\sum_k \sigma_k^2 < \infty$ and where $\phi_1(t), \phi_2(t), \dots$ is a sequence of orthonormal functions.

Assume that the impulse response $h(t)$ of the filter is a real $\mathcal{L}_1$ and $\mathcal{L}_2$ waveform. Then, for any given sample function $Z(t, \omega) = \sum_k Z_k(\omega) \phi_k(t)$ of the input, the filter output at any epoch τ is given by

$$V(\tau, \omega) = \int_{-\infty}^{\infty} Z(t, \omega) h(\tau - t) \mathrm{d}t = \sum_k Z_k(\omega) \int_{-\infty}^{\infty} \phi_k(t) h(\tau - t) \mathrm{d}t. \tag{7.34}$$

Each integral on the right side of (7.34) is an $\mathcal{L}_2$ function of τ (see Exercise 7.5). It follows from this (see Exercise 7.7) that $\int_{-\infty}^{\infty} Z(t, \omega) h(\tau - t) \mathrm{d}t$ is an $\mathcal{L}_2$ waveform with probability 1. For any given epoch τ, (7.34) maps sample points ω to real values, and thus $V(\tau, \omega)$ is a sample value of a random variable $V(\tau)$ defined as

$$V(\tau) = \int_{-\infty}^{\infty} Z(t) h(\tau - t) \mathrm{d}t = \sum_k Z_k \int_{-\infty}^{\infty} \phi_k(t) h(\tau - t) \mathrm{d}t. \tag{7.35}$$

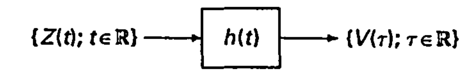


Figure 7.3. Filtered random process.

This is a Gaussian rv for each epoch τ. For any set of epochs $\tau_1, \ldots, \tau_\ell$, we see that $V(\tau_1), \ldots, V(\tau_\ell)$ are jointly Gaussian. Thus $\{V(\tau); \tau \in \mathbb{R}\}$ is a Gaussian random process.

We summarize Sections 7.4.1 and 7.4.2 in the following theorem.

Theorem 7.4.1 *Let $\{Z(t); t \in \mathbb{R}\}$ be a Gaussian process $Z(t) = \sum_k Z_k \phi_k(t)$, where $\{Z_k; k \geq 1\}$ is a sequence of independent Gaussian rvs $\mathcal{N}(0, \sigma_k^2)$, where $\sum \sigma_k^2 < \infty$ and $\{\phi_k(t); k \geq 1\}$ is an orthonormal set. Then*

- *for any set of $\mathcal{L}_2$ waveforms $g_1(t), \ldots, g_\ell(t)$, the linear functionals $\{Z_m; 1 \leq m \leq \ell\}$ given by $Z_m = \int_{-\infty}^{\infty} Z(t) g_m(t)\,\mathrm{d}t$ are zero-mean jointly Gaussian;*
- *for any filter with real $\mathcal{L}_1$ and $\mathcal{L}_2$ impulse response $h(t)$, the filter output $\{V(\tau); \tau \in \mathbb{R}\}$ given by (7.35) is a zero-mean Gaussian process.*

These are important results. The first, concerning sets of linear functionals, is important when we represent the input to the channel in terms of an orthonormal expansion; the noise can then often be expanded in the same orthonormal expansion. The second, concerning linear filtering, shows that when the received signal and noise are passed through a linear filter, the noise at the filter output is simply another zero-mean Gaussian process. This theorem is often summarized by saying that linear operations preserve Gaussianity.

7.4.3 Covariance for linear functionals and filters

Assume that $\{Z(t); t \in \mathbb{R}\}$ is a random process and that $g_1(t), \ldots, g_\ell(t)$ are real $\mathcal{L}_2$ waveforms. We have seen that if $\{Z(t); t \in \mathbb{R}\}$ is Gaussian, then the linear functionals $V_1, \ldots, V_\ell$ given by $V_m = \int_{-\infty}^{\infty} Z(t) g_m(t)\mathrm{d}t$ are jointly Gaussian for $1 \leq m \leq \ell$. We now want to find the covariance for each pair V_j, V_m of these random variables. The result does not depend on the process $Z(t)$ being Gaussian. The computation is quite simple, although we omit questions of limits, interchanges of order of expectation and integration, etc. A more careful derivation could be made by returning to the sampling-theorem arguments before, but this would somewhat obscure the ideas. Assuming that the process $Z(t)$ has zero mean,

$$\mathsf{E}[V_j V_m] = \mathsf{E}\left[\int_{-\infty}^{\infty} Z(t) g_j(t)\,\mathrm{d}t \int_{-\infty}^{\infty} Z(\tau) g_m(\tau)\mathrm{d}\tau\right] \tag{7.36}$$

$$= \int_{t=-\infty}^{\infty} \int_{\tau=-\infty}^{\infty} g_j(t)\mathsf{E}[Z(t)Z(\tau)] g_m(\tau)\mathrm{d}t\,\mathrm{d}\tau \tag{7.37}$$

$$= \int_{t=-\infty}^{\infty} \int_{\tau=-\infty}^{\infty} g_j(t)\mathsf{K}_Z(t, \tau) g_m(\tau)\mathrm{d}t\,\mathrm{d}\tau. \tag{7.38}$$

Each covariance term (including $\mathsf{E}[V_m^2]$ for each m) then depends only on the covariance function of the process and the set of waveforms $\{g_m; 1 \leq m \leq \ell\}$.

The convolution $V(r) = \int Z(t)h(r-t)\mathrm{d}t$ is a linear functional at each time r, so the covariance for the filtered output of $\{Z(t); t \in \mathbb{R}\}$ follows in the same way as the results above. The output $\{V(r)\}$ for a filter with a real $\mathcal{L}_2$ impulse response h is given by (7.35), so the covariance of the output can be found as follows:

$$\begin{aligned}\mathsf{K}_V(r,s) &= \mathsf{E}[V(r)V(s)]\\ &= \mathsf{E}\left[\int_{-\infty}^{\infty} Z(t)h(r-t)\mathrm{d}t \int_{-\infty}^{\infty} Z(\tau)h(s-\tau)\mathrm{d}\tau\right]\\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} h(r-t)\mathsf{K}_Z(t,\tau)h(s-\tau)\mathrm{d}t\,\mathrm{d}\tau. \end{aligned} \tag{7.39}$$

7.5 Stationarity and related concepts

Many of the most useful random processes have the property that the location of the time origin is irrelevant, i.e. they "behave" the same way at one time as at any other time. This property is called *stationarity*, and such a process is called a *stationary process*.

Since the location of the time origin must be irrelevant for stationarity, random processes that are defined over any interval other than $(-\infty, \infty)$ cannot be stationary. Thus, assume a process that is defined over $(-\infty, \infty)$.

The next requirement for a random process $\{Z(t);\, t \in \mathbb{R}\}$ to be stationary is that $Z(t)$ must be identically distributed for all epochs $t \in \mathbb{R}$. This means that, for any epochs t and $t+\tau$, and for any real number x, $\Pr\{Z(t) \le x\} = \Pr\{Z(t+\tau) \le x\}$. This does not mean that $Z(t)$ and $Z(t+\tau)$ are the same random variables; for a given sample outcome ω of the experiment, $Z(t,\omega)$ is typically unequal to $Z(t+\tau,\omega)$. It simply means that $Z(t)$ and $Z(t+\tau)$ have the same distribution function, i.e.

$$F_{Z(t)}(x) = F_{Z(t+\tau)}(x) \qquad \text{for all } x. \tag{7.40}$$

This is still not enough for stationarity, however. The joint distributions over any set of epochs must remain the same if all those epochs are shifted to new epochs by an arbitrary shift τ. This includes the previous requirement as a special case, so we have the following definition.

Definition 7.5.1 A random process $\{Z(t);\, t \in \mathbb{R}\}$ is *stationary* if, for all positive integers ℓ, for all sets of epochs $t_1, \dots, t_\ell \in \mathbb{R}$, for all amplitudes $z_1, \dots, z_\ell$, and for all shifts $\tau \in \mathbb{R}$,

$$F_{Z(t_1),\dots,Z(t_\ell)}(z_1, \dots, z_\ell) = F_{Z(t_1+\tau),\dots,Z(t_\ell+\tau)}(z_1, \dots, z_\ell). \tag{7.41}$$

For the typical case where densities exist, this can be rewritten as

$$f_{Z(t_1),\dots,Z(t_\ell)}(z_1, \dots, z_\ell) = f_{Z(t_1+\tau),\dots,Z(t_\ell+\tau)}(z_1, \dots, z_\ell) \tag{7.42}$$

for all $z_1, \dots, z_\ell \in \mathbb{R}$.

For a (zero-mean) Gaussian process, the joint distribution of $Z(t_1), \dots, Z(t_\ell)$ depends only on the covariance of those variables. Thus, this distribution will be the same as that of $Z(t_1+\tau), \dots, Z(t_\ell+\tau)$ if $\mathsf{K}_Z(t_m, t_j) = \mathsf{K}_Z(t_m+\tau, t_j+\tau)$ for $1 \le m, j \le \ell$. This condition will be satisfied for all τ, all ℓ, and all $t_1, \dots, t_\ell$ (verifying that $\{Z(t)\}$ is stationary) if $\mathsf{K}_Z(t_1, t_2) = \mathsf{K}_Z(t_1+\tau, t_2+\tau)$ for all τ and all t_1, t_2. This latter condition will be satisfied if $\mathsf{K}_Z(t_1, t_2) = \mathsf{K}_Z(t_1-t_2, 0)$ for all t_1, t_2. We have thus shown that a zero-mean Gaussian process is stationary if

$$\mathsf{K}_Z(t_1, t_2) = \mathsf{K}_Z(t_1-t_2, 0) \qquad \text{for all } t_1, t_2 \in \mathbb{R}. \tag{7.43}$$

Conversely, if (7.43) is not satisfied for some choice of t_1, t_2, then the joint distribution of $Z(t_1), Z(t_2)$ must be different from that of $Z(t_1-t_2), Z(0)$, and the process is not stationary. The following theorem summarizes this.

Theorem 7.5.1 *A zero-mean Gaussian process $\{Z(t);\ t\in\mathbb{R}\}$ is stationary if and only if (7.43) is satisfied.*

An obvious consequence of this is that a Gaussian process with a nonzero mean is stationary if and only if its mean is constant and its fluctuation satisfies (7.43).

7.5.1 Wide-sense stationary (WSS) random processes

There are many results in probability theory that depend only on the covariances of the random variables of interest (and also the mean if nonzero). For random processes, a number of these classical results are simplified for stationary processes, and these simplifications depend only on the mean and covariance of the process rather than full stationarity. This leads to the following definition.

Definition 7.5.2 A random process $\{Z(t);\ t\in\mathbb{R}\}$ is *wide-sense stationary (WSS)* if $\mathsf{E}[Z(t_1)] = \mathsf{E}[Z(0)]$ and $\mathsf{K}_Z(t_1, t_2) = \mathsf{K}_Z(t_1-t_2, 0)$ for all $t_1, t_2 \in \mathbb{R}$.

Since the covariance function $\mathsf{K}_Z(t+\tau, t)$ of a WSS process is a function of only one variable τ, we will often write the covariance function as a function of one variable, namely $\tilde{\mathsf{K}}_Z(\tau)$ in place of $\mathsf{K}_Z(t+\tau, t)$. In other words, the single variable in the single-argument form represents the difference between the two arguments in two-argument form. Thus, for a WSS process, $\mathsf{K}_Z(t, \tau) = \mathsf{K}_Z(t-\tau, 0) = \tilde{\mathsf{K}}_Z(t-\tau)$.

The random processes defined as expansions of T-spaced sinc functions have been discussed several times. In particular, let

$$V(t) = \sum_k V_k \operatorname{sinc}\left(\frac{t-kT}{T}\right), \tag{7.44}$$

where $\{\dots, V_{-1}, V_0, V_1, \dots\}$ is a sequence of (zero-mean) iid rvs. As shown in (7.8), the covariance function for this random process is given by

$$\mathsf{K}_V(t, \tau) = \sigma_V^2 \sum_k \operatorname{sinc}\left(\frac{t-kT}{T}\right) \operatorname{sinc}\left(\frac{\tau-kT}{T}\right), \tag{7.45}$$

where σ_V^2 is the variance of each V_k. The sum in (7.45), as shown below, is a function only of $t-\tau$, leading to the following theorem.

Theorem 7.5.2 (Sinc expansion) *The random process in (7.44) is WSS. In addition, if the rvs $\{V_k; k \in \mathbb{Z}\}$ are iid Gaussian, the process is stationary. The covariance function is given by*

$$\tilde{\mathsf{K}}_V(t-\tau) = \sigma_V^2 \operatorname{sinc}\left(\frac{t-\tau}{T}\right). \tag{7.46}$$

Proof From the sampling theorem, any $\mathcal{L}_2$ function $u(t)$, baseband-limited to $1/2T$, can be expanded as

$$u(t) = \sum_k u(kT) \operatorname{sinc}\left(\frac{t-kT}{T}\right). \tag{7.47}$$

For any given τ, take $u(t)$ to be $\operatorname{sinc}[(t-\tau)/T]$. Substituting this in (7.47), we obtain

$$\operatorname{sinc}\left(\frac{t-\tau}{T}\right) = \sum_k \operatorname{sinc}\left(\frac{kT-\tau}{T}\right) \operatorname{sinc}\left(\frac{t-kT}{T}\right) = \sum_k \operatorname{sinc}\left(\frac{\tau-kT}{T}\right) \operatorname{sinc}\left(\frac{t-kT}{T}\right). \tag{7.48}$$

Substituting this in (7.45) shows that the process is WSS with the stated covariance. As shown in Section 7.4.1, $\{V(t); t \in \mathbb{R}\}$ is Gaussian if the rvs $\{V_k\}$ are Gaussian. From Theorem 7.5.1, this Gaussian process is stationary since it is WSS. □

Next consider another special case of the sinc expansion in which each V_k is binary, taking values ±1 with equal probability. This corresponds to a simple form of a PAM transmitted waveform. In this case, $V(kT)$ must be ± 1, whereas, for values of t between the sample points, $V(t)$ can take on a wide range of values. Thus this process is WSS but cannot be stationary. Similarly, any discrete distribution for each V_k creates a process that is WSS but not stationary.

There are not many important models of *noise* processes that are WSS but not stationary,[12] despite the above example and the widespread usage of the term WSS. Rather, the notion of wide-sense stationarity is used to make clear, for some results, that they depend only on the mean and covariance, thus perhaps making it easier to understand them.

The Gaussian sinc expansion brings out an interesting theoretical non sequitur. Assuming that $\sigma_V^2 > 0$, i.e. that the process is not the trivial process for which $V(t) = 0$ with probability 1 for all t, the expected energy in the process (taken over all time) is infinite. It is not difficult to convince oneself that the sample functions of this process have infinite energy with probability 1. Thus, stationary noise models are simple to work with, but the sample functions of these processes do not fit into the $\mathcal{L}_2$ theory of waveforms that has been developed. Even more important than the issue of infinite energy, stationary noise models make unwarranted assumptions about the very distant

[12] An important exception is interference from other users, which, as the above sinc expansion with binary signals shows, can be WSS but not stationary. Even in this case, if the interference is modeled as just part of the noise (rather than specifically as interference), the nonstationarity is usually ignored.

past and future. The extent to which these assumptions affect the results about the present is an important question that must be asked.

The problem here is not with the peculiarities of the Gaussian sinc expansion. Rather it is that stationary processes have constant power over all time, and thus have infinite energy. One practical solution[13] to this is simple and familiar. The random process is simply truncated in any convenient way. Thus, when we say that noise is stationary, we mean that it is stationary within a much longer time interval than the interval of interest for communication. This is not very precise, and the notion of *effective stationarity* is now developed to formalize this notion of a truncated stationary process.

7.5.2 Effectively stationary and effectively WSS random processes

Definition 7.5.3 A (zero-mean) random process is *effectively stationary within* $[-T_0/2, T_0/2]$ if the joint probability assignment for $t_1, \dots, t_n$ is the same as that for $t_1+\tau, t_2+\tau, \dots, t_n+\tau$ whenever $t_1, \dots, t_n$ and $t_1+\tau, t_2+\tau, \dots, t_n+\tau$ are all contained in the interval $[-T_0/2, T_0/2]$. It is *effectively WSS within* $[-T_0/2, T_0/2]$ if $\mathsf{K}_Z(t,\tau)$ is a function only of $t-\tau$ for $t, \tau \in [-T_0/2, T_0/2]$. A random process with nonzero mean is effectively stationary (effectively WSS) if its mean is constant within $[-T_0/2, T_0/2]$ and its fluctuation is effectively stationary (WSS) within $[-T_0/2, T_0/2]$.

One way to view a stationary (WSS) random process is in the limiting sense of a process that is effectively stationary (WSS) for all intervals $[-T_0/2, T_0/2]$. For operations such as linear functionals and filtering, the nature of this limit as T_0 becomes large is quite simple and natural, whereas, for frequency-domain results, the effect of finite T_0 is quite subtle.

For an effectively WSS process within $[-T_0/2, T_0/2]$, the covariance within $[-T_0/2, T_0/2]$ is a function of a single parameter, $\mathsf{K}_Z(t,\tau) = \tilde{\mathsf{K}}_Z(t-\tau)$ for $t, \tau \in [-T_0/2, T_0/2]$. As illustrated by Figure 7.4, however, that $t-\tau$ can range from $-T_0$ (for $t=-T_0/2$, $\tau=T_0/2$) to T_0 (for $t=T_0/2$, $\tau=-T_0/2$).

Since a Gaussian process is determined by its covariance function and mean, it is effectively stationary within $[-T_0/2, T_0/2]$ if it is effectively WSS.

Note that the difference between a stationary and effectively stationary random process for large T_0 is primarily a difference in the model and not in the situation being modeled. If two models have a significantly different behavior over the time intervals of interest, or, more concretely, if noise in the distant past or future has a significant effect, then the entire modeling issue should be rethought.

[13] There is another popular solution to this problem. For any $\mathcal{L}_2$ function $g(t)$, the energy in $g(t)$ outside of $[-T_0/2, T_0/2]$ vanishes as $T_0 \to \infty$, so intuitively the effect of these tails on the linear functional $\int g(t)Z(t)dt$ vanishes as $T_0 \to 0$. This provides a nice intuitive basis for ignoring the problem, but it fails, both intuitively and mathematically, in the frequency domain.

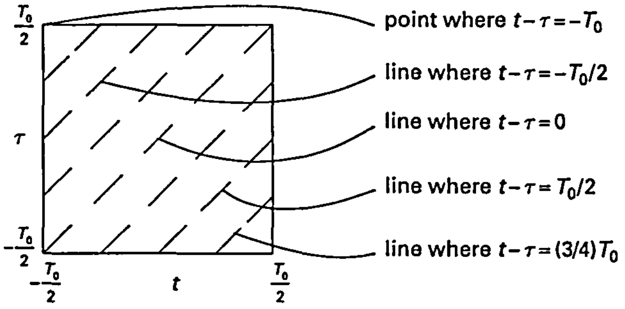

$\frac{T_0}{2}$
τ
$-\frac{T_0}{2}$
$-\frac{T_0}{2}$
t
$\frac{T_0}{2}$
point where $t-\tau=-T_0$
line where $t-\tau=-T_0/2$
line where $t-\tau=0$
line where $t-\tau=T_0/2$
line where $t-\tau=(3/4)T_0$

Figure 7.4. Relationship of the two-argument covariance function $\mathsf{K}_Z(t,\tau)$ and the one-argument function $\tilde{\mathsf{K}}_Z(t-\tau)$ for an effectively WSS process; $\mathsf{K}_Z(t,\tau)$ is constant on each dashed line above. Note that, for example, the line for which $t-\tau=(3/4)T_0$ applies only for pairs (t,τ) where $t \geq T_0/2$ and $\tau \leq -T_0/2$. Thus $\tilde{\mathsf{K}}_Z((3/4)T_0)$ is not necessarily equal to $\mathsf{K}_Z((3/4)T_0, 0)$. It can be easily verified, however, that $\tilde{\mathsf{K}}_Z(\alpha T_0)=\mathsf{K}_Z(\alpha T_0, 0)$ for all $\alpha \leq 1/2$.

7.5.3 Linear functionals for effectively WSS random processes

The covariance matrix for a set of linear functionals and the covariance function for the output of a linear filter take on simpler forms for WSS or effectively WSS processes than the corresponding forms for general processes derived in Section 7.4.3.

Let $Z(t)$ be a zero-mean WSS random process with covariance function $\tilde{\mathsf{K}}_Z(t-\tau)$ for $t,\tau \in [-T_0/2, T_0/2]$, and let $g_1(t), g_2(t), \ldots, g_\ell(t)$ be a set of $\mathcal{L}_2$ functions nonzero only within $[-T_0/2, T_0/2]$. For the conventional WSS case, we can take $T_0=\infty$. Let the linear functional V_m be given by $\int_{-T_0/2}^{T_0/2} Z(t)g_m(t)\,\mathrm{d}t$ for $1 \leq m \leq \ell$. The covariance $\mathsf{E}[V_m V_j]$ is then given by

$$\begin{aligned}\mathsf{E}[V_m V_j] &= \mathsf{E}\left[\int_{-T_0/2}^{T_0/2} Z(t)g_m(t)\mathrm{d}t \int_{-\infty}^{\infty} Z(\tau)g_j(\tau)\mathrm{d}\tau\right]\\ &= \int_{-T_0/2}^{T_0/2}\int_{-T_0/2}^{T_0/2} g_m(t)\tilde{\mathsf{K}}_Z(t-\tau)g_j(\tau)\mathrm{d}\tau\,\mathrm{d}t. \qquad (7.49)\end{aligned}$$

Note that this depends only on the covariance where $t,\tau \in [-T_0/2, T_0/2]$, i.e. where $\{Z(t)\}$ is effectively WSS. This is not surprising, since we would not expect V_m to depend on the behavior of the process outside of where $g_m(t)$ is nonzero.

7.5.4 Linear filters for effectively WSS random processes

Next consider passing a random process $\{Z(t);\, t \in \mathbb{R}\}$ through a linear time-invariant filter whose impulse response $h(t)$ is $\mathcal{L}_2$. As pointed out in (7.28), the output of the filter is a random process $\{V(\tau);\, \tau \in \mathbb{R}\}$ given by

$$V(\tau) = \int_{-\infty}^{\infty} Z(t_1)h(\tau-t_1)\mathrm{d}t_1.$$

Note that $V(\tau)$ is a linear functional for each choice of τ. The covariance function evaluated at t, τ is the covariance of the linear functionals $V(t)$ and $V(\tau)$. Ignoring questions of orders of integration and convergence,

$$\mathsf{K}_V(t, \tau) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} h(t - t_1)\mathsf{K}_Z(t_1, t_2)h(\tau - t_2)\mathrm{d}t_1\,\mathrm{d}t_2. \tag{7.50}$$

First assume that $\{Z(t);\ t \in \mathbb{R}\}$ is WSS in the conventional sense. Then $\mathsf{K}_Z(t_1, t_2)$ can be replaced by $\tilde{\mathsf{K}}_Z(t_1 - t_2)$. Replacing $t_1 - t_2$ by s (i.e. t_1 by $t_2 + s$), we have

$$\mathsf{K}_V(t, \tau) = \int_{-\infty}^{\infty} \left[\int_{-\infty}^{\infty} h(t - t_2 - s)\tilde{\mathsf{K}}_Z(s)\mathrm{d}s\right] h(\tau - t_2)\mathrm{d}t_2.$$

Replacing t_2 by $\tau + \mu$ yields

$$\mathsf{K}_V(t, \tau) = \int_{-\infty}^{\infty} \left[\int_{-\infty}^{\infty} h(t - \tau - \mu - s)\tilde{\mathsf{K}}_Z(s)\mathrm{d}s\right] h(-\mu)\mathrm{d}\mu. \tag{7.51}$$

Thus $\mathsf{K}_V(t, \tau)$ is a function only of $t - \tau$. This means that $\{V(t);\ t \in \mathbb{R}\}$ is WSS. This is not surprising; passing a WSS random process through a linear time-invariant filter results in another WSS random process.

If $\{Z(t);\ t \in \mathbb{R}\}$ is a Gaussian process, then, from Theorem 7.4.1, $\{V(t);\ t \in \mathbb{R}\}$ is also a Gaussian process. Since a Gaussian process is determined by its covariance function, it follows that if $Z(t)$ is a stationary Gaussian process, then $V(t)$ is also a stationary Gaussian process.

We do not have the mathematical machinery to carry out the above operations carefully over the infinite time interval.[14] Rather, it is now assumed that $\{Z(t);\ t \in \mathbb{R}\}$ is effectively WSS within $[-T_0/2, T_0/2]$. It will also be assumed that the impulse response $h(t)$ above is time-limited in the sense that, for some finite A, $h(t) = 0$ for $|t| > A$.

Theorem 7.5.3 *Let $\{Z(t);\ t \in \mathbb{R}\}$ be effectively WSS within $[-T_0/2, T_0/2]$ and have sample functions that are $\mathcal{L}_2$ within $[-T_0/2, T_0/2]$ with probability 1. Let $Z(t)$ be the input to a filter with an $\mathcal{L}_2$ time-limited impulse response $\{h(t)\colon [-A, A] \to \mathbb{R}\}$. Then, for $T_0/2 > A$, the output random process $\{V(t);\ t \in \mathbb{R}\}$ is WSS within $[-T_0/2 + A, T_0/2 - A]$ and its sample functions within $[-T_0/2 + A, T_0/2 - A]$ are $\mathcal{L}_2$ with probability 1.*

Proof Let $z(t)$ be a sample function of $Z(t)$ and assume $z(t)$ is $\mathcal{L}_2$ within $[-T_0/2, T_0/2]$. Let $v(\tau) = \int z(t)h(\tau - t)\,\mathrm{d}t$ be the corresponding filter output. For each $\tau \in [-T_0/2 + A, T_0/2 - A]$, $v(\tau)$ is determined by $z(t)$ in the range $t \in [-T_0/2, T_0/2]$.

[14] More important, we have no justification for modeling a process over the infinite time interval. Later, however, after building up some intuition about the relationship of an infinite interval to a very large interval, we can use the simpler equations corresponding to infinite intervals.

Thus, if we replace $z(t)$ by $z_0(t) = z(t)\,\mathrm{rect}[T_0]$, the filter output, say $v_0(\tau)$, will equal $v(\tau)$ for $\tau \in [-T_0/2+A,\ T_0/2-A]$. The time-limited function $z_0(t)$ is $\mathcal{L}_1$ as well as $\mathcal{L}_2$. This implies that the Fourier transform $\hat{z}_0(f)$ is bounded, say by $\hat{z}_0(f) \le B$, for each f. Since $\hat{v}_0(f) = \hat{z}_0(f)\hat{h}(f)$, we see that

$$\int |\hat{v}_0(f)|^2\,\mathrm{d}f = \int |\hat{z}_0(f)|^2 |\hat{h}(f)|^2\,\mathrm{d}f \le B^2 \int |\hat{h}(f)|^2\,\mathrm{d}f < \infty.$$

This means that $\hat{v}_0(f)$, and thus also $v_0(t)$, is $\mathcal{L}_2$. Now $v_0(t)$, when truncated to $[-T_0/2+A,\ T_0/2-A]$ is equal to $v(t)$ truncated to $[-T_0/2+A,\ T_0/2-A]$, so the truncated version of $v(t)$ is $\mathcal{L}_2$. Thus the sample functions of $\{V(t)\}$, truncated to $[-T_0/2+A,\ T_0/2-A]$, are $\mathcal{L}_2$ with probability 1.

Finally, since $\{Z(t);\, t \in \mathbb{R}\}$ can be truncated to $[-T_0/2,\ T_0/2]$ with no lack of generality, it follows that $\mathsf{K}_Z(t_1, t_2)$ can be truncated to $t_1, t_2 \in [-T_0/2,\ T_0/2]$. Thus, for $t, \tau \in [-T_0/2+A,\ T_0/2-A]$, (7.50) becomes

$$\mathsf{K}_V(t, \tau) = \int_{-T_0/2}^{T_0/2} \int_{-T_0/2}^{T_0/2} h(t - t_1)\tilde{\mathsf{K}}_Z(t_1 - t_2)h(\tau - t_2)\mathrm{d}t_1\,\mathrm{d}t_2. \qquad (7.52)$$

The argument in (7.50) and (7.51) shows that $V(t)$ is effectively WSS within $[-T_0/2+A,\ T_0/2-A]$. □

Theorem 7.5.3, along with the effective WSS result about linear functionals, shows us that results about WSS processes can be used within finite intervals. The result in the theorem about the interval of effective stationarity being reduced by filtering should not be too surprising. If we truncate a process and then pass it through a filter, the filter spreads out the effect of the truncation. For a finite-duration filter, however, as assumed here, this spreading is limited.

The notion of stationarity (or effective stationarity) makes sense as a modeling tool where T_0 is very much larger than other durations of interest, and in fact where there is no need for explicit concern about how long the process is going to be stationary.

Theorem 7.5.3 essentially tells us that we can have our cake and eat it too. That is, transmitted waveforms and noise processes can be truncated, thus making use of both common sense and $\mathcal{L}_2$ theory, but at the same time insights about stationarity can still be relied upon. More specifically, random processes can be modeled as stationary, without specifying a specific interval $[-T_0/2,\ T_0/2]$ of effective stationarity, because stationary processes can now be viewed as asymptotic versions of finite-duration processes.

Appendices 7.11.2 and 7.11.3 provide a deeper analysis of WSS processes truncated to an interval. The truncated process is represented as a Fourier series with random variables as coefficients. This gives a clean interpretation of what happens as the interval size is increased without bound, and also gives a clean interpretation of the effect of time-truncation in the frequency domain. Another approach to a truncated process is the Karhunen–Loeve expansion, which is discussed in Appendix 7.11.4.

7.6 Stationarity in the frequency domain

Stationary and WSS zero-mean processes, and particularly Gaussian processes, are often viewed more insightfully in the frequency domain than in the time domain. An effectively WSS process over $[-T_0/2, T_0/2]$ has a single-variable covariance function $\tilde{\mathsf{K}}_Z(\tau)$ defined over $[-T_0, T_0]$. A WSS process can be viewed as a process that is effectively WSS for each T_0. The energy in such a process, truncated to $[-T_0/2, T_0/2]$, is linearly increasing in T_0, but the covariance simply becomes defined over a larger and larger interval as $T_0 \to \infty$. We assume in what follows that this limiting covariance is $\mathcal{L}_2$. This does not appear to rule out any but the most pathological processes.

First we look at linear functionals and linear filters, ignoring limiting questions and convergence issues and assuming that T_0 is "large enough." We will refer to the random processes as stationary, while still assuming $\mathcal{L}_2$ sample functions.

For a zero-mean WSS process $\{Z(t); t \in \mathbb{R}\}$ and a real $\mathcal{L}_2$ function $g(t)$, consider the linear functional $V = \int g(t)Z(t)\mathrm{d}t$. From (7.49),

$$\mathsf{E}[V^2] = \int_{-\infty}^{\infty} g(t)\left[\int_{-\infty}^{\infty} \tilde{\mathsf{K}}_Z(t-\tau)g(\tau)\mathrm{d}\tau\right]\mathrm{d}t \tag{7.53}$$

$$= \int_{-\infty}^{\infty} g(t)\left[\tilde{\mathsf{K}}_Z * g\right](t)\mathrm{d}t, \tag{7.54}$$

where $\tilde{\mathsf{K}}_Z * g$ denotes the convolution of the waveforms $\tilde{\mathsf{K}}_Z(t)$ and $g(t)$. Let $S_Z(f)$ be the Fourier transform of $\tilde{\mathsf{K}}_Z(t)$. The function $S_Z(f)$ is called the *spectral density* of the stationary process $\{Z(t); t \in \mathbb{R}\}$. Since $\tilde{\mathsf{K}}_Z(t)$ is $\mathcal{L}_2$, real, and symmetric, its Fourier transform is also $\mathcal{L}_2$, real, and symmetric, and, as shown later, $S_Z(f) \geq 0$. It is also shown later that $S_Z(f)$ at each frequency f can be interpreted as the power per unit frequency at f.

Let $\theta(t) = [\tilde{\mathsf{K}}_Z * g](t)$ be the convolution of $\tilde{\mathsf{K}}_Z$ and g. Since g and K_Z are real, $\theta(t)$ is also real, so $\theta(t) = \theta^*(t)$. Using Parseval's theorem for Fourier transforms,

$$\mathsf{E}[V^2] = \int_{-\infty}^{\infty} g(t)\theta^*(t)\mathrm{d}t = \int_{-\infty}^{\infty} \hat{g}(f)\hat{\theta}^*(f)\mathrm{d}f.$$

Since $\theta(t)$ is the convolution of K_Z and g, we see that $\hat{\theta}(f) = S_Z(f)\hat{g}(f)$. Thus,

$$\mathsf{E}[V^2] = \int_{-\infty}^{\infty} \hat{g}(f)S_Z(f)\hat{g}^*(f)\mathrm{d}f = \int_{-\infty}^{\infty} |\hat{g}(f)|^2 S_Z(f)\mathrm{d}f. \tag{7.55}$$

Note that $\mathsf{E}[V^2] \geq 0$ and that this holds for all real $\mathcal{L}_2$ functions $g(t)$. The fact that $g(t)$ is real constrains the transform $\hat{g}(f)$ to satisfy $\hat{g}(f) = \hat{g}^*(-f)$, and thus $|\hat{g}(f)| = |\hat{g}(-f)|$ for all f. Subject to this constraint and the constraint that $|\hat{g}(f)|$ be $\mathcal{L}_2$, $|\hat{g}(f)|$ can be chosen as any $\mathcal{L}_2$ function. Stated another way, $\hat{g}(f)$ can be chosen arbitrarily for $f \geq 0$ subject to being $\mathcal{L}_2$.

Since $S_Z(f) = S_Z(-f)$, (7.55) can be rewritten as follows:

$$\mathsf{E}[V^2] = \int_0^\infty 2\,|\hat{g}(f)|^2 S_Z(f)\,\mathrm{d}f.$$

Since $\mathsf{E}[V^2] \geq 0$ and $|\hat{g}(f)|$ is arbitrary, it follows that $S_Z(f) \geq 0$ for all $f \in \mathbb{R}$.

The conclusion here is that the spectral density of any WSS random process must be nonnegative. Since $S_Z(f)$ is also the Fourier transform of $\tilde{\mathsf{K}}(t)$, this means that a necessary property of any single-variable covariance function is that it have a nonnegative Fourier transform.

Next, let $V_m = \int g_m(t)Z(t)\,\mathrm{d}t$, where the function $g_m(t)$ is real and $\mathcal{L}_2$ for $m = 1, 2$. From (7.49), we have

$$\mathsf{E}[V_1V_2] = \int_{-\infty}^{\infty} g_1(t)\left[\int_{-\infty}^{\infty} \tilde{\mathsf{K}}_Z(t-\tau)g_2(\tau)\mathrm{d}\tau\right]\mathrm{d}t \tag{7.56}$$

$$= \int_{-\infty}^{\infty} g_1(t)\left[\tilde{\mathsf{K}} * g_2\right](t)\mathrm{d}t. \tag{7.57}$$

Let $\hat{g}_m(f)$ be the Fourier transform of $g_m(t)$ for $m = 1, 2$, and let $\theta(t) = [\tilde{\mathsf{K}}_Z(t) * g_2](t)$ be the convolution of $\tilde{\mathsf{K}}_Z$ and g_2. Let $\hat{\theta}(f) = S_Z(f)\hat{g}_2(f)$ be its Fourier transform. As before, we have

$$\mathsf{E}[V_1V_2] = \int \hat{g}_1(f)\hat{\theta}^*(f)\mathrm{d}f = \int \hat{g}_1(f)S_Z(f)\hat{g}_2^*(f)\mathrm{d}f. \tag{7.58}$$

There is a remarkable feature in the above expression. If $\hat{g}_1(f)$ and $\hat{g}_2(f)$ have no overlap in frequency, then $\mathsf{E}[V_1V_2] = 0$. In other words, for any stationary process, two linear functionals over different frequency ranges must be uncorrelated. If the process is Gaussian, then the linear functionals are independent. This means in essence that Gaussian noise in different frequency bands must be independent. That this is true simply because of stationarity is surprising. Appendix 7.11.3 helps to explain this puzzling phenomenon, especially with respect to effective stationarity.

Next, let $\{\phi_m(t);\ m \in \mathbb{Z}\}$ be a set of real orthonormal functions and let $\{\hat{\phi}_m(f)\}$ be the corresponding set of Fourier transforms. Letting $V_m = \int Z(t)\phi_m(t)\mathrm{d}t$, (7.58) becomes

$$\mathsf{E}[V_mV_j] = \int \hat{\phi}_m(f)S_Z(f)\hat{\phi}_j^*(f)\mathrm{d}f. \tag{7.59}$$

If the set of orthonormal functions $\{\phi_m(t);\ m \in \mathbb{Z}\}$ is limited to some frequency band, and if $S_Z(f)$ is constant, say with value $N_0/2$ in that band, then

$$\mathsf{E}[V_mV_j] = \frac{N_0}{2}\int \hat{\phi}_m(f)\hat{\phi}_j^*(f)\mathrm{d}f. \tag{7.60}$$

By Parseval's theorem for Fourier transforms, we have $\int \hat{\phi}_m(f)\hat{\phi}_j^*(f)\mathrm{d}f = \delta_{mj}$, and thus

$$\mathsf{E}[V_mV_j] = \frac{N_0}{2}\delta_{mj}. \tag{7.61}$$

The rather peculiar-looking constant $N_0/2$ is explained in Section 7.7. For now, however, it is possible to interpret the meaning of the spectral density of a noise process.

Suppose that $S_Z(f)$ is continuous and approximately constant with value $S_Z(f_c)$ over some narrow band of frequencies around f_c, and suppose that $\phi_1(t)$ is constrained to that narrow band. Then the variance of the linear functional $\int_{-\infty}^{\infty} Z(t)\phi_1(t)\mathrm{d}t$ is approximately $S_Z(f_c)$. In other words, $S_Z(f_c)$ in some fundamental sense describes the energy in the noise per degree of freedom at the frequency f_c. Section 7.7 interprets this further.

7.7 White Gaussian noise

Physical noise processes are very often reasonably modeled as zero-mean, stationary, and Gaussian. There is one further simplification that is often reasonable. This is that the covariance between the noise at two epochs dies out very rapidly as the interval between those epochs increases. The interval over which this covariance is significantly nonzero is often very small relative to the intervals over which the signal varies appreciably. This means that the covariance function $\tilde{\mathsf{K}}_Z(\tau)$ looks like a short-duration pulse around $\tau = 0$.

We know from linear system theory that $\int \tilde{\mathsf{K}}_Z(t-\tau)g(\tau)\mathrm{d}\tau$ is equal to $g(t)$ if $\tilde{\mathsf{K}}_Z(t)$ is a unit impulse. Also, this integral is approximately equal to $g(t)$ if $\tilde{\mathsf{K}}_Z(t)$ has unit area and is a narrow pulse relative to changes in $g(t)$. It follows that, under the same circumstances, (7.56) becomes

$$\mathsf{E}[V_1 V_2^*] = \int_t \int_\tau g_1(t)\tilde{\mathsf{K}}_Z(t-\tau)g_2(\tau)\mathrm{d}\tau\,\mathrm{d}t \approx \int g_1(t)g_2(t)\mathrm{d}t. \tag{7.62}$$

This means that if the covariance function is very narrow relative to the functions of interest, then its behavior relative to those functions is specified by its area. In other words, the covariance function can be viewed as an impulse of a given magnitude. We refer to a zero-mean WSS Gaussian random process with such a narrow covariance function as *white Gaussian noise (WGN)*. The area under the covariance function is called the *intensity* or the *spectral density* of the WGN and is denoted by the symbol $N_0/2$. Thus, for $\mathcal{L}_2$ functions $g_1(t), g_2(t), \ldots$ in the range of interest, and for WGN (denoted by $\{W(t); t \in \mathbb{R}\}$) of intensity $N_0/2$, the random variable $V_m = \int W(t)g_m(t)\mathrm{d}t$ has variance given by

$$\mathsf{E}[V_m^2] = (N_0/2)\int g_m^2(t)\mathrm{d}t. \tag{7.63}$$

Similarly, the rvs V_j and V_m have covariance given by

$$\mathsf{E}[V_j V_m] = (N_0/2)\int g_j(t)g_m(t)\mathrm{d}t. \tag{7.64}$$

Also, $V_1, V_2, \ldots$ are jointly Gaussian.

The most important special case of (7.63) and (7.64) is to let $\phi_j(t)$ be a set of orthonormal functions and let $W(t)$ be WGN of intensity $N_0/2$. Let $V_m = \int \phi_m(t)W(t)\mathrm{d}t$. Then, from (7.63) and (7.64),

$$\mathsf{E}[V_j V_m] = (N_0/2)\delta_{jm}. \tag{7.65}$$

This is an important equation. It says that if the noise can be modeled as WGN, then when the noise is represented in terms of *any* orthonormal expansion, the resulting rvs are iid. Thus, we can represent signals in terms of an arbitrary orthonormal expansion, and represent WGN in terms of the same expansion, and the result is iid Gaussian rvs.

Since the coefficients of a WGN process in any orthonormal expansion are iid Gaussian, it is common to also refer to a random vector of iid Gaussian rvs as WGN.

If $\mathsf{K}_W(t)$ is approximated by $(N_0/2)\delta(t)$, then the spectral density is approximated by $S_W(f) = N_0/2$. If we are concerned with a particular band of frequencies, then we are interested in $S_W(f)$ being constant within that band, and in this case $\{W(t); t \in \mathbb{R}\}$ can be represented as white noise within that band. If this is the only band of interest, we can model[15] $S_W(f)$ as equal to $N_0/2$ everywhere, in which case the corresponding model for the covariance function is $(N_0/2)\delta(t)$.

The careful reader will observe that WGN has not really been defined. What has been said, in essence, is that if a stationary zero-mean Gaussian process has a covariance function that is very narrow relative to the variation of all functions of interest, or a spectral density that is constant within the frequency band of interest, then we can pretend that the covariance function is an impulse times $N_0/2$, where $N_0/2$ is the value of $S_W(f)$ within the band of interest. Unfortunately, according to the definition of random process, there cannot be any Gaussian random process $W(t)$ whose covariance function is $\tilde{\mathsf{K}}(t) = (N_0/2)\delta(t)$. The reason for this dilemma is that $\mathsf{E}[W^2(t)] = \mathsf{K}_W(0)$. We could interpret $\mathsf{K}_W(0)$ to be either undefined or ∞, but either way $W(t)$ cannot be a random variable (although we could think of it taking on only the values plus or minus ∞).

Mathematicians view WGN as a generalized random process, in the same sense as the unit impulse $\delta(t)$ is viewed as a generalized function. That is, the impulse function $\delta(t)$ is not viewed as an ordinary function taking the value 0 for $t \neq 0$ and the value ∞ at $t = 0$. Rather, it is viewed in terms of its effect on other, better behaved, functions $g(t)$, where $\int_{-\infty}^{\infty} g(t)\delta(t)\,\mathrm{d}t = g(0)$. In the same way, WGN is not viewed in terms of rvs at each epoch of time. Rather, it is viewed as a generalized zero-mean random process for which linear functionals are jointly Gaussian, for which variances and covariances are given by (7.63) and (7.64), and for which the covariance is formally taken to be $(N_0/2)\delta(t)$.

Engineers should view WGN within the context of an overall bandwidth and time interval of interest, where the process is effectively stationary within the time interval and has a constant spectral density over the band of interest. Within that context, the spectral density can be viewed as constant, the covariance can be viewed as an impulse, and (7.63) and (7.64) can be used.

The difference between the engineering view and the mathematical view is that the engineering view is based on a context of given time interval and bandwidth of

[15] This is not as obvious as it sounds, and will be further discussed in terms of the theorem of irrelevance in Chapter 8.

interest, whereas the mathematical view is based on a very careful set of definitions and limiting operations within which theorems can be stated without explicitly defining the context. Although the ability to prove theorems without stating the context is valuable, any application must be based on the context.

When we refer to signal space, what is usually meant is this overall bandwidth and time interval of interest, i.e. the context above. As we have seen, the bandwidth and the time interval cannot both be perfectly truncated, and because of this signal space cannot be regarded as strictly finite-dimensional. However, since the time interval and bandwidth are essentially truncated, visualizing signal space as finite-dimensional with additive WGN is often a reasonable model.

7.7.1 The sinc expansion as an approximation to WGN

Theorem 7.5.2 treated the process $Z(t) = \sum_k Z_k \operatorname{sinc}(t - kT/T)$, where each rv $\{Z_k; k \in \mathbb{Z}\}$ is iid and $\mathcal{N}(0, \sigma^2)$. We found that the process is zero-mean Gaussian and stationary with covariance function $\tilde{K}_Z(t-\tau) = \sigma^2 \operatorname{sinc}[(t-\tau)/T]$. The spectral density for this process is then given by

$$S_Z(f) = \sigma^2 T \operatorname{rect}(fT). \tag{7.66}$$

This process has a constant spectral density over the baseband bandwidth $\mathsf{W}_\mathsf{b} = 1/2T$, so, by making T sufficiently small, the spectral density is constant over a band sufficiently large to include all frequencies of interest. Thus this process can be viewed as WGN of spectral density $N_0/2 = \sigma^2 T$ for any desired range of frequencies $\mathsf{W}_\mathsf{b} = 1/2T$ by making T sufficiently small. Note, however, that to approximate WGN of spectral density $N_0/2$, the noise power, i.e. the variance of $Z(t)$ is $\sigma^2 = \mathsf{W}N_0$. In other words, σ^2 must increase with increasing W. This also says that N_0 is the noise power per unit *positive frequency*. The spectral density, $N_0/2$, is defined over both positive and negative frequencies, and so becomes N_0 when positive and negative frequencies are combined, as in the standard definition of bandwidth.[16]

If a sinc process is passed through a linear filter with an arbitrary impulse response $h(t)$, the output is a stationary Gaussian process with spectral density $|\hat{h}(f)|^2\sigma^2 T \operatorname{rect}(fT)$. Thus, by using a sinc process plus a linear filter, a stationary Gaussian process with any desired nonnegative spectral density within any desired finite bandwith can be generated. In other words, stationary Gaussian processes with arbitrary covariances (subject to $S(f) \geq 0$) can be generated from orthonormal expansions of Gaussian variables.

Since the sinc process is stationary, it has sample waveforms of infinite energy. As explained in Section 7.5.2, this process may be truncated to achieve an effectively stationary process with $\mathcal{L}_2$ sample waveforms. Appendix 7.11.3 provides some insight

[16] One would think that this field would have found a way to be consistent about counting only positive frequencies or positive and negative frequencies. However, the word bandwidth is so widely used among the mathophobic, and Fourier analysis is so necessary for engineers, that one must simply live with such minor confusions.

about how an effectively stationary Gaussian process over an interval T_0 approaches stationarity as $T_0 \to \infty$.

The sinc process can also be used to understand the strange, everywhere uncorrelated, process in Example 7.4.2. Holding $\sigma^2 = 1$ in the sinc expansion as T approaches 0, we get a process whose limiting covariance function is 1 for $t - \tau = 0$ and 0 elsewhere. The corresponding limiting spectral density is 0 everywhere. What is happening is that the power in the process (i.e. $\tilde{K}_Z(0)$) is 1, but that power is being spread over a wider and wider band as $T \to 0$, so the power per unit frequency goes to 0.

To explain this in another way, note that any measurement of this noise process must involve filtering over some very small, but nonzero, interval. The output of this filter will have zero variance. Mathematically, of course, the limiting covariance is $\mathcal{L}_2$-equivalent to 0, so again the mathematics[17] corresponds to engineering reality.

7.7.2 Poisson process noise

The sinc process of Section 7.7.1 is very convenient for generating noise processes that approximate WGN in an easily used formulation. On the other hand, this process is not very believable[18] as a physical process. A model that corresponds better to physical phenomena, particularly for optical channels, is a sequence of very narrow pulses which arrive according to a Poisson distribution in time.

The Poisson distribution, for our purposes, can be simply viewed as a limit of a discrete-time process where the time axis is segmented into intervals of duration Δ and a pulse of width Δ arrives in each interval with probability $\Delta\rho$, independent of every other interval. When such a process is passed through a linear filter, the fluctuation of the output at each instant of time is approximately Gaussian if the filter is of sufficiently small bandwidth to integrate over a very large number of pulses. One can similarly argue that linear combinations of filter outputs tend to be approximately Gaussian, making the process an approximation of a Gaussian process.

We do not analyze this carefully, since our point of view is that WGN, over limited bandwidths, is a reasonable and canonical approximation to a large number of physical noise processes. After understanding how this affects various communication systems, one can go back and see whether the model is appropriate for the given physical noise process. When we study wireless communication, we will find that the major problem is not that the noise is poorly approximated by WGN, but rather that the channel itself is randomly varying.

[17] This process also cannot be satisfactorily defined in a measure-theoretic way.

[18] To many people, defining these sinc processes with their easily analyzed properties, but no physical justification, is more troublesome than our earlier use of discrete memoryless sources in studying source coding. In fact, the approach to modeling is the same in each case: first understand a class of easy-to-analyze but perhaps impractical processes, then build on that understanding to understand practical cases. Actually, sinc processes have an advantage here: the bandlimited stationary Gaussian random processes defined this way (although not the method of generation) are widely used as practical noise models, whereas there are virtually no uses of discrete memoryless sources as practical source models.

7.8 Adding noise to modulated communication

Consider the QAM communication problem again. A complex $\mathcal{L}_2$ baseband waveform $u(t)$ is generated and modulated up to passband as a real waveform $x(t) = 2\Re[u(t)e^{2\pi i f_c t}]$. A sample function $w(t)$ of a random noise process $W(t)$ is then added to $x(t)$ to produce the output $y(t) = x(t) + w(t)$, which is then demodulated back to baseband as the received complex baseband waveform $v(t)$.

Generalizing QAM somewhat, assume that $u(t)$ is given by $u(t) = \sum_k u_k \theta_k(t)$, where the functions $\theta_k(t)$ are complex orthonormal functions and the sequence of symbols $\{u_k; k \in \mathbb{Z}\}$ are complex numbers drawn from the symbol alphabet and carrying the information to be transmitted. For each symbol u_k, $\Re(u_k)$ and $\Im(u_k)$ should be viewed as sample values of the random variables $\Re(U_k)$ and $\Im(U_k)$. The joint probability distributions of these rvs is determined by the incoming random binary digits and how they are mapped into symbols. The *complex random variable*[19] $\Re(U_k) + i\Im(U_k)$ is then denoted by U_k.

In the same way, $\Re(\sum_k U_k \theta_k(t))$ and $\Im(\sum_k U_k \theta_k(t))$ are random processes denoted, respectively, by $\Re(U(t))$ and $\Im(U(t))$. We then call $U(t) = \Re(U(t)) + i\Im(U(t))$ for $t \in \mathbb{R}$ a *complex random process*. A complex random process $U(t)$ is defined by the joint distribution of $U(t_1), U(t_2), \ldots, U(t_n)$ for all choices of $n, t_1, \ldots, t_n$. This is equivalent to defining both $\Re(U(t))$ and $\Im(U(t))$ as joint processes.

Recall from the discussion of the Nyquist criterion that if the QAM transmit pulse $p(t)$ is chosen to be square root of Nyquist, then $p(t)$ and its T-spaced shifts are orthogonal and can be normalized to be orthonormal. Thus a particularly natural choice here is $\theta_k(t) = p(t - kT)$ for such a p. Note that this is a generalization of Chapter 6 in the sense that $\{U_k; k \in \mathbb{Z}\}$ is a sequence of complex rvs using random choices from the signal constellation rather than some given sample function of that random sequence. The transmitted passband (random) waveform is then given by

$$X(t) = \sum_k 2\Re\{U_k \theta_k(t) \exp(2\pi i f_c t)\}. \tag{7.67}$$

Recall that the transmitted waveform has twice the power of the baseband waveform. Now define the following:

$$\psi_{k,1}(t) = \Re\{2\theta_k(t)\exp(2\pi i f_c t)\};$$
$$\psi_{k,2}(t) = \Im\{-2\theta_k(t)\exp(2\pi i f_c t)\}.$$

Also, let $U_{k,1} = \Re(U_k)$ and $U_{k,2} = \Im(U_k)$. Then

$$X(t) = \sum_k [U_{k,1}\psi_{k,1}(t) + U_{k,2}\psi_{k,2}(t)].$$

[19] Recall that a rv is a mapping from sample points to real numbers, so that a complex rv is a mapping from sample points to complex numbers. Sometimes in discussions involving both rvs and complex rvs, it helps to refer to rvs as real rvs, but the modifier "real" is superfluous.

As shown in Theorem 6.6.1, the set of bandpass functions $\{\psi_{k,\ell};\, k \in \mathbb{Z}, \ell \in \{1,2\}\}$ are orthogonal, and each has energy equal to 2. This again assumes that the carrier frequency f_c is greater than all frequencies in each baseband function $\theta_k(t)$.

In order for $u(t)$ to be $\mathcal{L}_2$, assume that the number of orthogonal waveforms $\theta_k(t)$ is arbitrarily large but finite, say $\theta_1(t), \dots, \theta_n(t)$. Thus $\{\psi_{k,\ell}\}$ is also limited to $1 \le k \le n$.

Assume that the noise $\{W(t);\, t \in \mathbb{R}\}$ is white over the band of interest and effectively stationary over the time interval of interest, but has $\mathcal{L}_2$ sample functions.[20] Since $\{\psi_{k,l};\, 1 \le k \le n, \ell = 1,2\}$ is a finite real orthogonal set, the projection theorem can be used to express each sample noise waveform $\{w(t);\, t \in \mathbb{R}\}$ as

$$w(t) = \sum_{k=1}^{n} [z_{k,1}\psi_{k,1}(t) + z_{k,2}\psi_{k,2}(t)] + w_{\perp}(t), \tag{7.68}$$

where $w_{\perp}(t)$ is the component of the sample noise waveform perpendicular to the space spanned by $\{\psi_{k,l};\, 1 \le k \le n, \ell = 1,2\}$. Let $Z_{k,\ell}$ be the rv with sample value $z_{k,\ell}$. Then each rv $Z_{k,\ell}$ is a linear functional on $W(t)$. Since $\{\psi_{k,l};\, 1 \le k \le n, \ell = 1,2\}$ is an orthogonal set, the rvs $Z_{k,\ell}$ are iid Gaussian rvs. Let $W_{\perp}(t)$ be the random process corresponding to the sample function $w_{\perp}(t)$ above. Expanding $\{W_{\perp}(t);\, t \in \mathbb{R}\}$ in an orthonormal expansion orthogonal to $\{\psi_{k,l};\, 1 \le k \le n, \ell = 1,2\}$, the coefficients are assumed to be independent of the $Z_{k,\ell}$, at least over the time and frequency band of interest. What happens to these coefficients outside of the region of interest is of no concern, other than assuming that $W_{\perp}(t)$ is independent of $U_{k,\ell}$ and $Z_{k,\ell}$ for $1 \le k \le n$ and $\ell = \{1,2\}$. The received waveform $Y(t) = X(t) + W(t)$ is then given by

$$Y(t) = \sum_{k=1}^{n} \left[(U_{k,1} + Z_{k,1})\psi_{k,1}(t) + (U_{k,2} + Z_{k,2})\psi_{k,2}(t)\right] + W_{\perp}(t).$$

When this is demodulated,[21] the baseband waveform is represented as the complex waveform,

$$V(t) = \sum_{k} (U_k + Z_k)\theta_k(t) + Z_{\perp}(t), \tag{7.69}$$

where each Z_k is a complex rv given by $Z_k = Z_{k,1} + \mathrm{i}Z_{k,2}$ and the baseband residual noise $Z_{\perp}(t)$ is independent of $\{U_k, Z_k;\, 1 \le k \le n\}$. The variance of each real rv $Z_{k,1}$ and $Z_{k,2}$ is taken by convention to be $N_0/2$. We follow this convention because we are measuring the input power at baseband; as mentioned many times, the power at passband is scaled to be twice that at baseband. The point here is that N_0 is not a physical constant; rather, it is the noise power per unit positive frequency *in the units used to represent the signal power.*

[20] Since the set of orthogonal waveforms $\theta_k(t)$ is not necessarily time- or frequency-limited, the assumption here is that the noise is white over a much larger time and frequency interval than the nominal bandwidth and time interval used for communication. This assumption is discussed further in Chapter 8.

[21] Some filtering is necessary before demodulation to remove the residual noise that is far out of band, but we do not want to analyze that here.

7.8.1 Complex Gaussian random variables and vectors

Noise waveforms, after demodulation to baseband, are usually complex and are thus represented, as in (7.69), by a sequence of complex random variables which is best regarded as a complex random vector (rv). It is possible to view any such n-dimensional complex rv $Z = Z_{\mathrm{re}} + \mathrm{i}Z_{\mathrm{im}}$ as a $2n$-dimensional real rv

$$\begin{bmatrix} Z_{\mathrm{re}} \\ Z_{\mathrm{im}} \end{bmatrix}, \text{ where } Z_{\mathrm{re}} = \Re(Z) \text{ and } Z_{\mathrm{im}} = \Im(Z).$$

For many of the same reasons that it is desirable to work directly with a complex baseband waveform rather than a pair of real passband waveforms, it is often beneficial to work directly with complex rvs.

A complex rv $Z = Z_{\mathrm{re}} + \mathrm{i}Z_{\mathrm{im}}$ is *Gaussian* if Z_{re} and Z_{im} are jointly Gaussian; Z is *circularly symmetric Gaussian*[22] if it is Gaussian and in addition Z_{re} and Z_{im} are iid. In this case (assuming zero mean as usual), the amplitude of Z is a Rayleigh-distributed rv and the phase is uniformly distributed; thus the joint density is circularly symmetric. A circularly symmetric complex Gaussian rv Z is fully described by its mean $\bar{Z}$ (which we continue to assume to be 0 unless stated otherwise) and its variance $\sigma^2 = \mathsf{E}[\tilde{Z}\tilde{Z}^*]$. A circularly symmetric complex Gaussian rv Z of mean $\bar{Z}$ and variance σ^2 is denoted by $Z \sim \mathcal{CN}(\bar{Z}, \sigma^2)$.

A complex random vector $\boldsymbol{Z}$ is a jointly Gaussian rv if the real and imaginary components of $\boldsymbol{Z}$ collectively are jointly Gaussian; it is also circularly symmetric if the density of the fluctuation $\tilde{\boldsymbol{Z}}$ (i.e. the joint density of the real and imaginary parts of the components of $\tilde{\boldsymbol{Z}}$) is the same[23] as that of $\mathrm{e}^{\mathrm{i}\theta}\tilde{\boldsymbol{Z}}$ for all phase angles θ.

An important example of a circularly symmetric Gaussian rv is $\boldsymbol{Z} = (Z_1, \ldots, Z_n)^{\mathsf{T}}$, where the real and imaginary components collectively are iid and $\mathcal{N}(0,1)$. Because of the circular symmetry of each Z_k, multiplying $\boldsymbol{Z}$ by $\mathrm{e}^{\mathrm{i}\theta}$ simply rotates each Z_k and the probability density does not change. The probability density is just that of $2n$ iid $\mathcal{N}(0,1)$ rvs, which is

$$f_{\boldsymbol{Z}}(\boldsymbol{z}) = \frac{1}{(2\pi)^n} \exp\left(\frac{\sum_{k=1}^{n} -|z_k|^2}{2} \right), \tag{7.70}$$

where we have used the fact that $|z_k|^2 = \Re(z_k)^2 + \Im(z_k)^2$ for each k to replace a sum over $2n$ terms with a sum over n terms.

Another much more general example is to let A be an arbitrary complex n by n matrix and let the complex rv $\boldsymbol{Y}$ be defined by

$$\boldsymbol{Y} = \mathsf{A}\boldsymbol{Z}, \tag{7.71}$$

[22] This is sometimes referred to as complex proper Gaussian.

[23] For a single complex rv Z with Gaussian real and imaginary parts, this phase-invariance property is enough to show that the real and imaginary parts are jointly Gaussian, and thus that Z is circularly symmetric Gaussian. For a random vector with Gaussian real and imaginary parts, phase invariance as defined here is not sufficient to ensure the jointly Gaussian property. See Exercise 7.14 for an example.

where Z has iid real and imaginary normal components as above. The complex rv defined in this way has jointly Gaussian real and imaginary parts. To see this, represent (7.71) as the following real linear transformation of $2n$ real space:

$$\begin{bmatrix} Y_{\text{re}} \\ Y_{\text{im}} \end{bmatrix} = \begin{bmatrix} \mathsf{A}_{\text{re}} & -\mathsf{A}_{\text{im}} \\ \mathsf{A}_{\text{im}} & \mathsf{A}_{\text{re}} \end{bmatrix} \begin{bmatrix} Z_{\text{re}} \\ Z_{\text{im}} \end{bmatrix}, \tag{7.72}$$

where $Y_{\text{re}} = \Re(Y)$, $Y_{\text{im}} = \Im(Y)$, $\mathsf{A}_{\text{re}} = \Re(\mathsf{A})$, and $\mathsf{A}_{\text{im}} = \Im(\mathsf{A})$.

The rv Y is also circularly symmetric.[24] To see this, note that $e^{i\theta}Y = e^{i\theta}\mathsf{A}Z = \mathsf{A}e^{i\theta}Z$. Since Z is circularly symmetric, the density at any given sample value z (i.e. the density for the real and imaginary parts of z) is the same as that at $e^{i\theta}z$. This in turn implies[25] that the density at y is the same as that at $e^{i\theta}y$.

The covariance matrix of a complex rv Y is defined as

$$\mathsf{K}_Y = \mathsf{E}[YY^{\dagger}], \tag{7.73}$$

where $Y^{\dagger}$ is defined as $Y^{\mathsf{T}*}$. For a random vector Y defined by (7.71), $\mathsf{K}_Y = \mathsf{A}\mathsf{A}^{\dagger}$.

Finally, for a circularly symmetric complex Gaussian vector as defined in (7.71), the probability density is given by

$$f_Y(y) = \frac{1}{(2\pi)^n \det(\mathsf{K}_Y)} \exp\left(-y^{\dagger}\mathsf{K}_Y y\right). \tag{7.74}$$

It can be seen that complex circularly symmetric Gaussian vectors behave quite similarly to (real) jointly Gaussian vectors. Both are defined by their covariance matrices, the properties of the covariance matrices are almost identical (see Appendix 7.11.1), the covariance can be expressed as $\mathsf{A}\mathsf{A}^{\dagger}$, where A describes a linear transformation from iid components, and the transformation A preserves the circularly symmetric Gaussian property in the complex case and the jointly Gaussian property in the real case.

An arbitrary (zero-mean) complex Gaussian rv is not specified by its variance, since $\mathsf{E}[Z_{\text{re}}^2]$ might be different from $\mathsf{E}[Z_{\text{im}}^2]$. Similarly, an arbitrary (zero-mean) complex Gaussian vector is not specified by its covariance matrix. In fact, arbitrary Gaussian complex n-vectors are usually best viewed as $2n$-dimensional real vectors; the simplifications from dealing with complex Gaussian vectors directly are primarily constrained to the circularly symmetric case.

7.9 Signal-to-noise ratio

There are a number of different measures of signal power, noise power, energy per symbol, energy per bit, and so forth, which are defined here. These measures are explained

[24] Conversely, as we will see later, all circularly symmetric jointly Gaussian rvs can be defined this way.

[25] This is not as simple as it appears, and is shown more carefully in the exercises. It is easy to become facile at working in $\mathbb{R}^n$ and $\mathbb{C}^n$, but going back and forth between $\mathbb{R}^{2n}$ and $\mathbb{C}^n$ is tricky and inelegant (witness (7.72) and (7.71)).

in terms of QAM and PAM, but they also apply more generally. In Section 7.8, a fairly general set of orthonormal functions was used, and here a specific set is assumed. Consider the orthonormal functions $p_k(t) = p(t-kT)$ as used in QAM, and use a nominal passband bandwidth $\mathsf{W} = 1/T$. Each QAM symbol U_k can be assumed to be iid with energy $E_s = \mathsf{E}[|U_k|^2]$. This is the signal energy per two real dimensions (i.e. real plus imaginary). The noise energy per two real dimensions is defined to be N_0. Thus the signal-to-noise ratio is defined to be

$$\mathrm{SNR} = \frac{E_s}{N_0} \qquad \text{for QAM.} \tag{7.75}$$

For baseband PAM, using real orthonormal functions satisfying $p_k(t) = p(t-kT)$, the signal energy per signal is $E_s = \mathsf{E}[|U_k|^2]$. Since the signal is 1D, i.e. real, the noise energy per dimension is defined to be $N_0/2$. Thus, the SNR is defined to be

$$\mathrm{SNR} = \frac{2E_s}{N_0} \qquad \text{for PAM.} \tag{7.76}$$

For QAM there are W complex degrees of freedom per second, so the signal power is given by $P = E_s\mathsf{W}$. For PAM at baseband, there are $2\mathsf{W}$ degrees of freedom per second, so the signal power is $P = 2E_s\mathsf{W}$. Thus, in each case, the SNR becomes

$$\mathrm{SNR} = \frac{P}{N_0\mathsf{W}} \qquad \text{for QAM and PAM.} \tag{7.77}$$

We can interpret the denominator here as the overall noise power in the bandwidth W, so SNR is also viewed as the signal power divided by the noise power in the nominal band. For those who like to minimize the number of formulas they remember, all of these equations for SNR follow from a basic definition as the signal energy per degree of freedom divided by the noise energy per degree of freedom.

PAM and QAM each use the same signal energy for each degree of freedom (or at least for each complex pair of degrees of freedom), whereas other systems might use the available degrees of freedom differently. For example, PAM with baseband bandwidth W occupies bandwidth $2\mathsf{W}$ if modulated to passband, and uses only half the available degrees of freedom. For these situations, SNR can be defined in several different ways depending on the context. As another example, frequency-hopping is a technique used both in wireless and in secure communication. It is the same as QAM, except that the carrier frequency f_c changes pseudo-randomly at intervals long relative to the symbol interval. Here the bandwidth W might be taken as the bandwidth of the underlying QAM system, or as the overall bandwidth within which f_c hops. The SNR in (7.77) is quite different in the two cases.

The appearance of W in the denominator of the expression for SNR in (7.77) is rather surprising and disturbing at first. It says that if more bandwidth is allocated to a communication system with the same available power, then SNR *decreases*. This is because the signal energy per degree of freedom decreases when it is spread over more degrees of freedom, but the noise is everywhere. We will see later that the net gain can be made positive.

Another important parameter is the rate R; this is the number of transmitted bits per second, which is the number of bits per symbol, $\log_2|\mathcal{A}|$, times the number of symbols per second. Thus

$$R = \mathrm{W}\log_2|\mathcal{A}| \quad \text{for QAM;} \qquad R = 2\mathrm{W}\log_2|\mathcal{A}| \quad \text{for PAM.} \tag{7.78}$$

An important parameter is the *spectral efficiency* of the system, which is defined as $\rho = R/\mathrm{W}$. This is the transmitted number of bits per second in each unit frequency interval. For QAM and PAM, ρ is given by (7.78) to be

$$\rho = \log_2|\mathcal{A}| \quad \text{for QAM;} \qquad \rho = 2\log_2|\mathcal{A}| \quad \text{for PAM.} \tag{7.79}$$

More generally, the spectral efficiency ρ can be defined as the number of transmitted bits per degree of freedom. From (7.79), achieving a large value of spectral efficiency requires making the symbol alphabet large; note that ρ increases only logarithmically with $|\mathcal{A}|$.

Yet another parameter is the energy per bit E_b. Since each symbol contains $\log_2 \mathcal{A}$ bits, E_b is given for both QAM and PAM by

$$E_\mathrm{b} = \frac{E_\mathrm{s}}{\log_2|\mathcal{A}|}. \tag{7.80}$$

One of the most fundamental quantities in communication is the ratio E_b/N_0. Since E_b is the signal energy per bit and N_0 is the noise energy per two degrees of freedom, this provides an important limit on energy consumption. For QAM, we substitute (7.75) and (7.79) into (7.80), yielding

$$\frac{E_\mathrm{b}}{N_0} = \frac{\mathrm{SNR}}{\rho}. \tag{7.81}$$

The same equation is seen to be valid for PAM. This says that achieving a small value for E_b/N_0 requires a small ratio of SNR to ρ. We look at this next in terms of channel capacity.

One of Shannon's most famous results was to develop the concept of the capacity C of an additive WGN communication channel. This is defined as the supremum of the number of bits per second that can be transmitted and received with arbitrarily small error probability. For the WGN channel with a constraint W on the bandwidth and a constraint P on the received signal power, he showed that

$$C = \mathrm{W}\log_2\left(1 + \frac{P}{\mathrm{W}N_0}\right). \tag{7.82}$$

Furthermore, arbitrarily small error probability can be achieved at any rate $R < C$ by using channel coding of arbitrarily large constraint length. He also showed, and later results strengthened the fact, that larger rates would lead to large error probabilities. These results will be demonstrated in Chapter 8; they are widely used as a benchmark for comparison with particular systems. Figure 7.5 shows a sketch of C as a function of W. Note that C increases monotonically with W, reaching a limit of $(P/N_0)\log_2 \mathrm{e}$ as $\mathrm{W} \to \infty$. This is known as the ultimate Shannon limit on achievable rate. Note also that

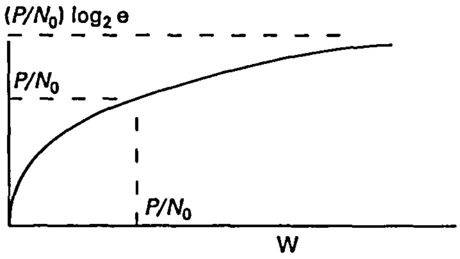


Figure 7.5. Capacity as a function of bandwidth W for fixed P/N_0.

when $W = P/N_0$, i.e. when the bandwidth is large enough for the SNR to reach 1, then C is within $1/\log_2 e$, 69% of the ultimate Shannon limit. This is usually expressed as being within 1.6 dB of the ultimate Shannon limit.

Shannon's result showed that the error probability can be made arbitrarily small for any rate $R < C$. Using (7.81) for C, ρ for R/W, and SNR for $P/(WN_0)$, the inequality $R < C$ becomes

$$\rho < \log_2(1+\text{SNR}). \tag{7.83}$$

If we substitute this into (7.81), we obtain

$$\frac{E_b}{N_0} > \frac{\text{SNR}}{\log_2(1+\text{SNR})}.$$

This is a monotonic increasing function of the single-variable SNR, which in turn is decreasing in W. Thus $(E_b/N_0)_{\min}$ is monotonically decreasing in W. As $W \to \infty$ it reaches the limit $\ln 2 = 0.693$, i.e. -1.59 dB. As W decreases, it grows, reaching 0 dB at SNR = 1, and increasing without bound for yet smaller W. The limiting spectral efficiency, however, is C/W. This is also monotonically decreasing in W, going to 0 as $W \to \infty$. In other words, there is a trade-off between the required E_b/N_0, which is preferably small, and the required spectral efficiency ρ, which is preferably large. This is discussed further in Chapter 8.

7.10 Summary of random processes

The additive noise in physical communication systems is usually best modeled as a random process, i.e. a collection of random variables, one at each real-valued instant of time. A random process can be specified by its joint probability distribution over all finite sets of epochs, but additive noise is most often modeled by the assumption that the rvs are all zero-mean Gaussian and their joint distribution is jointly Gaussian.

These assumptions were motivated partly by the central limit theorem, partly by the simplicity of working with Gaussian processes, partly by custom, and partly by various extremal properties. We found that jointly Gaussian means a great deal more than individually Gaussian, and that the resulting joint densities are determined by the covariance matrix. These densities have ellipsoidal contours of equal probability density whose axes are the eigenfunctions of the covariance matrix.

A sample function $Z(t, \omega)$ of a random process $Z(t)$ can be viewed as a waveform and interpreted as an $\mathcal{L}_2$ vector. For any fixed $\mathcal{L}_2$ function $g(t)$, the inner product $\langle g(t), Z(t, \omega)\rangle$ maps ω into a real number and thus can be viewed over Ω as a random variable. This rv is called a linear function of $Z(t)$ and is denoted by $\int g(t)Z(t)\,dt$.

These linear functionals arise when expanding a random process into an orthonormal expansion and also at each epoch when a random process is passed through a linear filter. For simplicity, these linear functionals and the underlying random processes are not viewed in a measure-theoretic perspective, although the $\mathcal{L}_2$ development in Chapter 4 provides some insight about the mathematical subtleties involved.

Noise processes are usually viewed as being stationary, which effectively means that their statistics do not change in time. This generates two problems: first, the sample functions have infinite energy, and, second, there is no clear way to see whether results are highly sensitive to time regions far outside the region of interest. Both of these problems are treated by defining effective stationarity (or effective wide-sense stationarity) in terms of the behavior of the process over a finite interval. This analysis shows, for example, that Gaussian linear functionals depend only on effective stationarity over the signal space of interest. From a practical standpoint, this means that the simple results arising from the assumption of stationarity can be used without concern for the process statistics outside the time range of interest.

The spectral density of a stationary process can also be used without concern for the process outside the time range of interest. If a process is effectively WSS, it has a single-variable covariance function corresponding to the interval of interest, and this has a Fourier transform which operates as the spectral density over the region of interest. How these results change as the region of interest approaches ∞ is explained in Appendix 7.11.3.

7.11 Appendix: Supplementary topics

7.11.1 Properties of covariance matrices

This appendix summarizes some properties of covariance matrices that are often useful but not absolutely critical to our treatment of random processes. Rather than repeat everything twice, we combine the treatment for real and complex rvs together. On a first reading, however, one might assume everything to be real. Most of the results are the same in each case, although the complex-conjugate signs can be removed in the real case. It is important to realize that the properties developed here apply to non-Gaussian as well as Gaussian rvs. All rvs and rvs here are assumed to be zero-mean.

A square matrix K is a *covariance matrix* if a (real or complex) rv $\boldsymbol{Z}$ exists such that $\mathsf{K} = \mathsf{E}[\boldsymbol{Z}\boldsymbol{Z}^{\mathsf{T}*}]$. The complex conjugate of the transpose, $\boldsymbol{Z}^{\mathsf{T}*}$, is called the *Hermitian transpose* and is denoted by $\boldsymbol{Z}^{\dagger}$. If $\boldsymbol{Z}$ is real, of course, $\boldsymbol{Z}^{\dagger} = \boldsymbol{Z}^{\mathsf{T}}$. Similarly, for a matrix K, the Hermitian conjugate, denoted by $\mathsf{K}^{\dagger}$, is $\mathsf{K}^{\mathsf{T}*}$. A matrix is *Hermitian* if $\mathsf{K} = \mathsf{K}^{\dagger}$. Thus a real Hermitian matrix (a Hermitian matrix containing all real terms) is a symmetric matrix.

An n by n square matrix K with real or complex terms is *nonnegative definite* if it is Hermitian and if, for all $b \in \mathbb{C}^n$, $b^\dagger \mathsf{K} b$ is real and nonnegative. It is *positive definite* if, in addition, $b^\dagger \mathsf{K} b > 0$ for $b \neq 0$. We now list some of the important relationships between nonnegative definite, positive definite, and covariance matrices and state some other useful properties of covariance matrices.

(1) Every covariance matrix K is nonnegative definite. To see this, let Z be a rv such that $\mathsf{K} = \mathsf{E}[ZZ^\dagger]$; K is Hermitian since $\mathsf{E}[Z_k Z_m^*] = \mathsf{E}[Z_m^* Z_k]$ for all k, m. For any $b \in \mathbb{C}^n$, let $X = b^\dagger Z$. Then $0 \le \mathsf{E}[|X|^2] = \mathsf{E}\left[(b^\dagger Z)(b^\dagger Z)^*\right] = \mathsf{E}\left[b^\dagger ZZ^\dagger b\right] = b^\dagger \mathsf{K} b$.

(2) For any complex n by n matrix A, the matrix $\mathsf{K} = \mathsf{A}\mathsf{A}^\dagger$ is a covariance matrix. In fact, let Z have n independent unit-variance elements so that K_Z is the identity matrix I_n. Then $Y = \mathsf{A}Z$ has the covariance matrix $\mathsf{K}_Y = \mathsf{E}[(\mathsf{A}Z)(\mathsf{A}Z)^\dagger] = \mathsf{E}[\mathsf{A}ZZ^\dagger \mathsf{A}^\dagger] = \mathsf{A}\mathsf{A}^\dagger$. Note that if A is real and Z is real, then Y is real and, of course, K_Y is real. It is also possible for A to be real and Z complex, and in this case K_Y is still real but Y is complex.

(3) A covariance matrix K is positive definite if and only if K is nonsingular. To see this, let $\mathsf{K} = \mathsf{E}[ZZ^\dagger]$ and note that, if $b^\dagger \mathsf{K} b = 0$ for some $b \neq 0$, then $X = b^\dagger Z$ has zero variance, and therefore is 0 with probability 1. Thus $\mathsf{E}[XZ^\dagger] = 0$, so $b^\dagger \mathsf{E}[ZZ^\dagger] = 0$. Since $b \neq 0$ and $b^\dagger \mathsf{K} = 0$, K must be singular. Conversely, if K is singular, there is some b such that $\mathsf{K}b = 0$, so $b^\dagger \mathsf{K} b$ is also 0.

(4) A complex number λ is an *eigenvalue* of a square matrix K if $\mathsf{K}q = \lambda q$ for some nonzero vector q; the corresponding q is an *eigenvector* of K. The following results about the eigenvalues and eigenvectors of positive (nonnegative) definite matrices K are standard linear algebra results (see, for example, Strang (1976), sect 5.5).

All eigenvalues of K are positive (nonnegative). If K is real, the eigenvectors can be taken to be real. Eigenvectors of different eigenvalues are orthogonal, and the eigenvectors of any one eigenvalue form a subspace whose dimension is called the *multiplicity* of that eigenvalue. If K is n by n, then n orthonormal eigenvectors $q_1, \ldots, q_n$ can be chosen. The corresponding list of eigenvalues $\lambda_1, \ldots, \lambda_n$ need not be distinct; specifically, the number of repetitions of each eigenvalue equals the multiplicity of that eigenvalue. Finally, $\det(\mathsf{K}) = \prod_{k=1}^n \lambda_k$.

(5) If K is nonnegative definite, let Q be the matrix with the orthonormal columns $q_1, \ldots, q_n$ defined in item (4) above. Then Q satisfies $\mathsf{KQ} = \mathsf{Q}\Lambda$, where $\Lambda = \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$. This is simply the vector version of the eigenvector/eigenvalue relationship above. Since $q_k^\dagger q_m = \delta_{nm}$, Q also satisfies $\mathsf{Q}^\dagger \mathsf{Q} = \mathsf{I}_n$. We then also have $\mathsf{Q}^{-1} = \mathsf{Q}^\dagger$ and thus $\mathsf{Q}\mathsf{Q}^\dagger = \mathsf{I}_n$; this says that the rows of Q are also orthonormal. Finally, by post-multiplying $\mathsf{KQ} = \mathsf{Q}\Lambda$ by $\mathsf{Q}^\dagger$, we see that $\mathsf{K} = \mathsf{Q}\Lambda\mathsf{Q}^\dagger$. The matrix Q is called *unitary* if complex and *orthogonal* if real.

(6) If K is positive definite, then $\mathsf{K}b \neq 0$ for $b \neq 0$. Thus K can have no zero eigenvalues and Λ is nonsingular. It follows that K can be inverted as $\mathsf{K}^{-1} = \mathsf{Q}\Lambda^{-1}\mathsf{Q}^\dagger$. For any n-vector b,

$$b^\dagger \mathsf{K}^{-1} b = \sum_k \lambda_k^{-1} |\langle b, q_k \rangle|^2.$$

To see this, note that $b^{\dagger}\mathsf{K}^{-1}b = b^{\dagger}Q\Lambda^{-1}\mathsf{Q}^{\dagger}b$. Letting $v = \mathsf{Q}^{\dagger}b$ and using the fact that the rows of Q^{T} are the orthonormal vectors q_k, we see that $\langle b, q_k\rangle$ is the kth component of v. We then have $v^{\dagger}\Lambda^{-1}v = \sum_k \lambda_k^{-1}|v_k|^2$, which is equivalent to the desired result. Note that $\langle b, q_k\rangle$ is the projection of b in the direction of q_k.

(7) We have $\det \mathsf{K} = \prod_{k=1}^{n} \lambda_k$, where $\lambda_1, \dots, \lambda_n$ are the eigenvalues of K repeated according to their multiplicity. Thus, if K is positive definite, $\det \mathsf{K} > 0$, and, if K is nonnegative definite, $\det \mathsf{K} \geq 0$.

(8) If K is a positive definite (semi-definite) matrix, then there is a unique positive definite (semi-definite) square root matrix R satisfying $\mathsf{R}^2 = \mathsf{K}$. In particular, R is given by

$$\mathsf{R} = \mathsf{Q}\Lambda^{1/2}\mathsf{Q}^{\dagger}, \quad \text{where } \Lambda^{1/2} = \text{diag}\left(\sqrt{\lambda_1}, \sqrt{\lambda_2}, \dots, \sqrt{\lambda_n}\right). \tag{7.84}$$

(9) If K is nonnegative definite, then K is a covariance matrix. In particular, K is the covariance matrix of $Y = \mathsf{R}Z$, where R is the square root matrix in (7.84) and $\mathsf{K}_Z = \mathsf{I}_m$. This shows that zero-mean jointly Gaussian rvs exist with any desired covariance matrix; the definition of jointly Gaussian here as a linear combination of normal rvs does not limit the possible set of covariance matrices.

For any given covariance matrix K, there are usually many choices for A satisfying $\mathsf{K} = \mathsf{A}\mathsf{A}^{\dagger}$. The square root matrix R is simply a convenient choice. Some of the results in this section are summarized in the following theorem.

Theorem 7.11.1 An n by n matrix K is a covariance matrix if and only if it is nonnegative definite. Also K is a covariance matrix if and only if $\mathsf{K} = \mathsf{A}\mathsf{A}^{\dagger}$ for an n by n matrix A. One choice for A is the square root matrix R in (7.84).

7.11.2 The Fourier series expansion of a truncated random process

Consider a (real zero-mean) random process that is effectively WSS over some interval $[-T_0/2, T_0/2]$ where T_0 is viewed intuitively as being very large. Let $\{Z(t); |t| \leq T_0/2\}$ be this process truncated to the interval $[-T_0/2, T_0/2]$. The objective of this and Section 7.11.3 is to view this truncated process in the frequency domain and discover its relation to the spectral density of an untruncated WSS process. A second objective is to interpret the statistical independence between different frequencies for stationary Gaussian processes in terms of a truncated process.

Initially assume that $\{Z(t); |t| \leq T_0/2\}$ is arbitrary; the effective WSS assumption will be added later. Assume the sample functions of the truncated process are $\mathcal{L}_2$ real functions with probability 1. Each $\mathcal{L}_2$ sample function, say $\{Z(t, \omega); |t| \leq T_0/2\}$ can then be expanded in a Fourier series, as follows:

$$Z(t, \omega) = \sum_{k=-\infty}^{\infty} \hat{Z}_k(\omega) e^{2\pi i k t/T_0}, \qquad |t| \leq \frac{T_0}{2}. \tag{7.85}$$

The orthogonal functions here are complex and the coefficients $\hat{Z}_k(\omega)$ can be similarly complex. Since the sample functions $\{Z(t,\omega);\ |t| \le T_0/2\}$ are real, $\hat{Z}_k(\omega) = \hat{Z}^*_{-k}(\omega)$ for each k. This also implies that $\hat{Z}_0(\omega)$ is real. The inverse Fourier series is given by

$$\hat{Z}_k(\omega) = \frac{1}{T_0}\int_{-T_0/2}^{T_0/2} Z(t,\omega)\mathrm{e}^{-2\pi ikt/T_0}\,\mathrm{d}t. \tag{7.86}$$

For each sample point ω, $\hat{Z}_k(\omega)$ is a complex number, so $\hat{Z}_k$ is a complex rv, i.e. $\Re(\hat{Z}_k)$ and $\Im(\hat{Z}_k)$ are both rvs. Also, $\Re(\hat{Z}_k) = \Re(\hat{Z}_{-k})$ and $\Im(\hat{Z}_k) = -\Im(\hat{Z}_{-k})$ for each k. It follows that the truncated process $\{Z(t);\ |t| \le T_0/2\}$ defined by

$$Z(t) = \sum_{k=-\infty}^{\infty} \hat{Z}_k \mathrm{e}^{2\pi ikt/T_0}, \qquad -\frac{T_0}{2} \le t \le \frac{T_0}{2}, \tag{7.87}$$

is a (real) random process and the complex rvs $\hat{Z}_k$ are complex linear functionals of $Z(t)$ given by

$$\hat{Z}_k = \frac{1}{T_0}\int_{-T_0/2}^{T_0/2} Z(t)\mathrm{e}^{-2\pi ikt/T_0}\,\mathrm{d}t. \tag{7.88}$$

Thus (7.87) and (7.88) are a Fourier series pair between a random process and a sequence of complex rvs. The sample functions satisfy

$$\frac{1}{T_0}\int_{-T_0/2}^{T_0/2} Z^2(t,\omega)\,\mathrm{d}t = \sum_{k\in\mathbb{Z}} |\hat{Z}_k(\omega)|^2,$$

so that

$$\frac{1}{T_0}\mathsf{E}\left[\int_{t=-T_0/2}^{T_0/2} Z^2(t)\,\mathrm{d}t\right] = \sum_{k\in\mathbb{Z}} \mathsf{E}\left[|\hat{Z}_k|^2\right]. \tag{7.89}$$

The assumption that the sample functions are $\mathcal{L}_2$ with probability 1 can be seen to be equivalent to the assumption that

$$\sum_{k\in\mathbb{Z}} S_k < \infty, \qquad \text{where } S_k = \mathsf{E}[|\hat{Z}_k|^2]. \tag{7.90}$$

This is summarized in the following theorem.

Theorem 7.11.2 *If a zero-mean (real) random process is truncated to* $[-T_0/2, T_0/2]$, *and the truncated sample functions are* $\mathcal{L}_2$ *with probability 1, then the truncated process is specified by the joint distribution of the complex Fourier-coefficient random variables* $\{\hat{Z}_k\}$. *Furthermore, any joint distribution of* $\{\hat{Z}_k;\ k \in \mathbb{Z}\}$ *that satisfies (7.90) specifies such a truncated process.*

The covariance function of a truncated process can be calculated from (7.87) as follows:

$$\mathsf{K}_Z(t,\tau) = \mathsf{E}[Z(t)Z^*(\tau)] = \mathsf{E}\left[\sum_k \hat{Z}_k \mathrm{e}^{2\pi ikt/T_0} \sum_m \hat{Z}^*_m \mathrm{e}^{-2\pi im\tau/T_0}\right]$$

$$= \sum_{k,m} \mathsf{E}[\hat{Z}_k\hat{Z}^*_m]\mathrm{e}^{2\pi ikt/T_0}\mathrm{e}^{-2\pi im\tau/T_0}, \qquad \text{for } -\frac{T_0}{2} \le t,\ \tau \le \frac{T_0}{2}. \tag{7.91}$$

Note that if the function on the right of (7.91) is extended over all $t, \tau \in \mathbb{R}$, it becomes periodic in t with period T_0 for each τ, and periodic in τ with period T_0 for each t.

Theorem 7.11.2 suggests that virtually any truncated process can be represented as a Fourier series. Such a representation becomes far more insightful and useful, however, if the Fourier coefficients are uncorrelated. Sections 7.11.3 and 7.11.4 look at this case and then specialize to Gaussian processes, where uncorrelated implies independent.

7.11.3 Uncorrelated coefficients in a Fourier series

Consider the covariance function in (7.91) under the additional assumption that the Fourier coefficients $\{\tilde{Z}_k; k \in \mathbb{Z}\}$ are uncorrelated, i.e. that $\mathsf{E}[\hat{Z}_k \hat{Z}_m^*] = 0$ for all k, m such that $k \neq m$. This assumption also holds for $m = -k \neq 0$, and, since $Z_k = Z_{-k}^*$ for all k, implies both that $\mathsf{E}[(\Re(Z_k))^2] = \mathsf{E}[(\Im(Z_k))^2]$ and $\mathsf{E}[\Re(Z_k)\Im(Z_k)] = 0$ (see Exercise 7.10). Since $\mathsf{E}[\hat{Z}_k \hat{Z}_m^*] = 0$ for $k \neq m$, (7.91) simplifies to

$$\mathsf{K}_Z(t, \tau) = \sum_{k \in \mathbf{Z}} S_k e^{2\pi i k (t-\tau)/T_0}, \qquad \text{for } -\frac{T_0}{2} \leq t,\ \tau \leq \frac{T_0}{2}. \tag{7.92}$$

This says that $\mathsf{K}_Z(t, \tau)$ is a function only of $t - \tau$ over $-T_0/2 \leq t, \tau \leq T_0/2$, i.e. that $\mathsf{K}_Z(t, \tau)$ is effectively WSS over $[-T_0/2, T_0/2]\}$. Thus $\mathsf{K}_Z(t, \tau)$ can be denoted by $\tilde{\mathsf{K}}_Z(t - \tau)$ in this region, and

$$\tilde{\mathsf{K}}_Z(\tau) = \sum_k S_k e^{2\pi i k \tau / T_0}. \tag{7.93}$$

This means that the variances S_k of the sinusoids making up this process are the Fourier series coefficients of the covariance function $\tilde{\mathsf{K}}_Z(r)$.

In summary, the assumption that a truncated (real) random process has uncorrelated Fourier series coefficients over $[-T_0/2, T_0/2]$ implies that the process is WSS over $[-T_0/2, T_0/2]$ and that the variances of those coefficients are the Fourier coefficients of the single-variable covariance. This is intuitively plausible since the sine and cosine components of each of the corresponding sinusoids are uncorrelated and have equal variance.

Note that $\mathsf{K}_Z(t, \tau)$ in the above example is defined for all $t, \tau \in [-T_0/2, T_0/2]$ and thus $t - \tau$ ranges from $-T_0$ to T_0 and $\tilde{\mathsf{K}}_Z(r)$ must satisfy (7.93) for $-T_0 \leq r \leq T_0$. From (7.93), $\tilde{\mathsf{K}}_Z(r)$ is also periodic with period T_0, so the interval $[-T_0, T_0]$ constitutes two periods of $\tilde{\mathsf{K}}_Z(r)$. This means, for example, that $\mathsf{E}[Z(-\varepsilon)Z^*(\varepsilon)] = \mathsf{E}[Z(T_0/2 - \varepsilon)Z^*(-T_0/2 + \varepsilon)]$. More generally, the periodicity of $\tilde{\mathsf{K}}_Z(r)$ is reflected in $\mathsf{K}_Z(t, \tau)$, as illustrated in Figure 7.6.

We have seen that essentially any random process, when truncated to $[-T_0/2, T_0/2]$, has a Fourier series representation, and that, if the Fourier series coefficients are uncorrelated, then the truncated process is WSS over $[-T_0/2, T_0/2]$ and has a covariance function which is periodic with period T_0. This proves the first half of the following theorem.

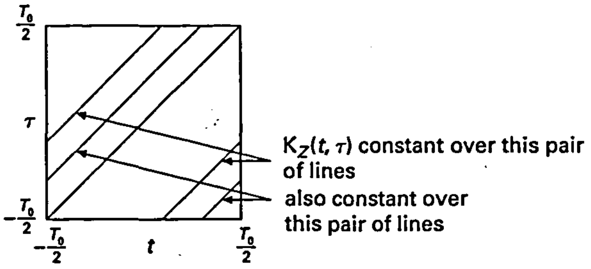

$\frac{T_0}{2}$
τ
$-\frac{T_0}{2}$
$-\frac{T_0}{2}$
t
$\frac{T_0}{2}$
$K_Z(t,\tau)$ constant over this pair of lines
also constant over this pair of lines

Figure 7.6. Constraint on $K_Z(t,\tau)$ imposed by periodicity of $\tilde{K}_Z(t-\tau)$.

Theorem 7.11.3 *Let $\{Z(t);\ t \in [-T_0/2,\ T_0/2]\}$ be a finite-energy zero-mean (real) random process over $[-T_0/2,\ T_0/2]$ and let $\{\hat{Z}_k;\ k \in \mathbb{Z}\}$ be the Fourier series rvs of (7.87) and (7.88).*

- *If $\mathsf{E}[Z_k Z_m^*] = S_k \delta_{k,m}$ for all $k, m \in \mathbb{Z}$, then $\{Z(t); t \in [-T_0/2,\ T_0/2]\}$ is effectively WSS within $[-T_0/2,\ T_0/2]$ and satisfies (7.93).*
- *If $\{Z(t);\ t \in [-T_0/2,\ T_0/2]\}$ is effectively WSS within $[-T_0/2,\ T_0/2]$ and if $\tilde{K}_Z(t-\tau)$ is periodic with period T_0 over $[-T_0,\ T_0]$, then $\mathsf{E}[Z_k Z_m^*] = S_k \delta_{k,m}$ for some choice of $S_k \geq 0$ and for all $k, m \in \mathbb{Z}$.*

Proof To prove the second part of the theorem, note from (7.88) that

$$\mathsf{E}[\hat{Z}_k \hat{Z}_m^*] = \frac{1}{T_0^2} \int_{-T_0/2}^{T_0/2} \int_{-T_0/2}^{T_0/2} \mathsf{K}_Z(t,\tau) e^{-2\pi i k t/T_0} e^{2\pi i m \tau/T_0}\, dt\, d\tau. \tag{7.94}$$

By assumption, $\mathsf{K}_Z(t,\tau) = \tilde{\mathsf{K}}_Z(t-\tau)$ for $t, \tau \in [-T_0/2,\ T_0/2]$ and $\tilde{\mathsf{K}}_Z(t-\tau)$ is periodic with period T_0. Substituting $s = t-\tau$ for t as a variable of integration, (7.94) becomes

$$\mathsf{E}[Z_k Z_m^*] = \frac{1}{T_0^2} \int_{-T_0/2}^{T_0/2} \left(\int_{-T_0/2-\tau}^{T_0/2-\tau} \tilde{\mathsf{K}}_Z(s) e^{-2\pi i k s/T_0}\, ds \right) e^{-2\pi i k \tau/T_0} e^{2\pi i m \tau/T_0}\, d\tau. \tag{7.95}$$

The integration over s does not depend on τ because the interval of integration is one period and $\tilde{\mathsf{K}}_Z$ is periodic. Thus, this integral is only a function of k, which we denote by $T_0 S_k$. Thus

$$\mathsf{E}[Z_k Z_m^*] = \frac{1}{T_0} \int_{-T_0/2}^{T_0/2} S_k e^{-2\pi i (k-m)\tau/T_0}\, d\tau = \begin{cases} S_k & \text{for } m = k; \\ 0 & \text{otherwise.} \end{cases} \tag{7.96}$$

This shows that the Z_k are uncorrelated, completing the proof. □

The next issue is to find the relationship between these processes and processes that are WSS over all time. This can be done most cleanly for the case of Gaussian processes. Consider a WSS (and therefore stationary) zero-mean Gaussian random

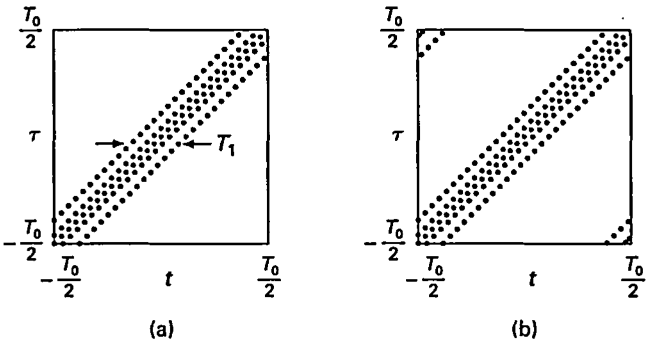


Figure 7.7. (a) $\mathsf{K}_{Z'}(t, \tau)$ over the region $-T_0/2 \le t, \tau \le T_0/2$ for a stationary process Z' satisfying $\tilde{\mathsf{K}}_{Z'}(\tau) = 0$ for $|\tau| > T_1/2$. (b) $\mathsf{K}_Z(t, \tau)$ for an approximating process Z comprising independent sinusoids, spaced by $1/T_0$ and with uniformly distributed phase. Note that the covariance functions are identical except for the anomalous behavior at the corners where t is close to $T_0/2$ and τ is close to $-T_0/2$ or vice versa.

process[26] $\{Z'(t); t \in \mathbb{R}\}$ with covariance function $\tilde{\mathsf{K}}_{Z'}(\tau)$ and assume a limited region of nonzero covariance; i.e.

$$\tilde{\mathsf{K}}_{Z'}(\tau) = 0 \qquad \text{for} \quad |\tau| > \frac{T_1}{2}.$$

Let $S_{Z'}(f) \ge 0$ be the spectral density of Z' and let T_0 satisfy $T_0 > T_1$. The Fourier series coefficients of $\tilde{\mathsf{K}}_{Z'}(\tau)$ over the interval $[-T_0/2, T_0/2]$ are then given by $S_k = S_{Z'}(k/T_0)/T_0$. Suppose this process is approximated over the interval $[-T_0/2, T_0/2]$ by a truncated Gaussian process $\{Z(t);\ t \in [-T_0/2, T_0/2]\}$ composed of independent Fourier coefficients $\hat{Z}_k$, i.e.

$$Z(t) = \sum_k \hat{Z}_k e^{2\pi ikt/T_0}, \qquad -\frac{T_0}{2} \le t \le \frac{T_0}{2},$$

where

$$\mathsf{E}[\hat{Z}_k \hat{Z}_m^*] = S_k \delta_{k,m}, \qquad \text{for all } k, m \in \mathbb{Z}.$$

By Theorem 7.11.3, the covariance function of $Z(t)$ is $\tilde{\mathsf{K}}_Z(\tau) = \sum_k S_k e^{2\pi ikt/T_0}$. This is periodic with period T_0 and for $|\tau| \le T_0/2$, $\tilde{\mathsf{K}}_Z(\tau) = \tilde{\mathsf{K}}_{Z'}(\tau)$. The original process $Z'(t)$ and the approximation $Z(t)$ thus have the same covariance for $|\tau| \le T_0/2$. For $|\tau| > T_0/2$, $\tilde{\mathsf{K}}_{Z'}(\tau) = 0$, whereas $\tilde{\mathsf{K}}_Z(\tau)$ is periodic over all τ. Also, of course, Z' is stationary, whereas Z is effectively stationary within its domain $[-T_0/2, T_0/2]$. The difference between Z and Z' becomes more clear in terms of the two-variable covariance function, illustrated in Figure 7.7.

[26] Equivalently, one can assume that Z' is effectively WSS over some interval much larger than the intervals of interest here.

It is evident from the figure that if Z' is modeled as a Fourier series over $[-T_0/2, T_0/2]$ using independent complex circularly symmetric Gaussian coefficients, then $\mathsf{K}_{Z'}(t, \tau) = \mathsf{K}_Z(t, \tau)$ for $|t|, |\tau| \leq (T_0 - T_1)/2$. Since zero-mean Gaussian processes are completely specified by their covariance functions, this means that Z' and Z are statistically identical over this interval.

In summary, a stationary Gaussian process Z' cannot be perfectly modeled over an interval $[-T_0/2, T_0/2]$ by using a Fourier series over that interval. The anomalous behavior is avoided, however, by using a Fourier series over an interval large enough to include the interval of interest plus the interval over which $\tilde{\mathsf{K}}_{Z'}(\tau) \neq 0$. If this latter interval is unbounded, then the Fourier series model can only be used as an approximation. The following theorem has been established.

Theorem 7.11.4 *Let $Z'(t)$ be a zero-mean stationary Gaussian random process with spectral density $S(f)$ and covariance $\tilde{\mathsf{K}}_{Z'}(\tau) = 0$ for $|\tau| \geq T_1/2$. Then for $T_0 > T_1$, the truncated process $Z(t) = \sum_k Z_k e^{2\pi i k t/T_0}$ for $|t| \leq T_0/2$, where the Z_k are independent and $Z_k \sim \mathcal{CN}(S(k/T_0)/T_0)$ for all $k \in \mathbb{Z}$ is statistically identical to $Z'(t)$ over $[-(T_0 - T_1)/2, (T_0 - T_1)/2]$.*

The above theorem is primarily of conceptual use, rather than as a problem-solving tool. It shows that, aside from the anomalous behavior discussed above, stationarity can be used over the region of interest without concern for how the process behaves far outside the interval of interest. Also, since T_0 can be arbitrarily large, and thus the sinusoids arbitrarily closely spaced, we see that the relationship between stationarity of a Gaussian process and independence of frequency bands is quite robust and more than something valid only in a limiting sense.

7.11.4 The Karhunen–Loeve expansion

There is another approach, called the Karhunen–Loeve expansion, for representing a random process that is truncated to some interval $[-T_0/2, T_0/2]$ by an orthonormal expansion. The objective is to choose a set of orthonormal functions such that the coefficients in the expansion are uncorrelated.

We start with the covariance function $\mathsf{K}(t, \tau)$ defined for $t, \tau \in [-T_0/2, T_0/2]$. The basic facts about these time-limited covariance functions are virtually the same as the facts about covariance matrices in Appendix 7.11.1. That is, $\mathsf{K}(t, \tau)$ is nonnegative definite in the sense that for all $\mathcal{L}_2$ functions $g(t)$,

$$\int_{-T_0/2}^{T_0/2} \int_{-T_0/2}^{T_0/2} g(t) \mathsf{K}_Z(t, \tau) g(\tau) \mathrm{d}t \, \mathrm{d}\tau \geq 0$$

Note that K_Z also has real-valued orthonormal eigenvectors defined over $[-T_0/2, T_0/2]$ and nonnegative eigenvalues. That is,

$$\int_{-T_0/2}^{T_0/2} \mathsf{K}_Z(t, \tau) \phi_m(\tau) \mathrm{d}\tau = \lambda_m \phi_m(t); \quad t \in \left[-\frac{T_0}{2}, \frac{T_0}{2}\right],$$

where $\langle \phi_m, \phi_k \rangle = \delta_{m,k}$. These eigenvectors span the $\mathcal{L}_2$ space of real functions over $[-T_0/2, T_0/2]$. By using these eigenvectors as the orthonormal functions of $Z(t) = \sum_m Z_m \phi_m(t)$, it is easy to show that $\mathsf{E}[Z_m Z_k] = \lambda_m \delta_{m,k}$. In other words, given an arbitrary covariance function over the truncated interval $[-T_0/2, T_0/2]$, we can find a particular set of orthonormal functions so that $Z(t) = \sum_m Z_m \phi_m(t)$ and $\mathsf{E}[Z_m Z_k] = \lambda_m \delta_{m,k}$. This is called the Karhunen–Loeve expansion.

These equations for the eigenvectors and eigenvalues are well known integral equations and can be calculated by computer. Unfortunately, they do not provide a great deal of insight into the frequency domain.

7.12 Exercises

7.1 (a) Let X, Y be iid rvs, each with density $f_x(x) = \alpha \exp(-x^2/2)$. In part (b), we show that α must be $1/\sqrt{2\pi}$ in order for $f_x(x)$ to integrate to 1, but in this part we leave α undetermined. Let $S = X^2 + Y^2$. Find the probability density of S in terms of α. [Hint. Sketch the contours of equal probability density in the X, Y plane.]

(b) Prove from part (a) that α must be $1/\sqrt{2\pi}$ in order for S, and thus X and Y, to be rvs. Show that $\mathsf{E}[X] = 0$ and that $\mathsf{E}[X^2] = 1$.

(c) Find the probability density of $R = \sqrt{S}$ (R is called a *Rayleigh* rv).

7.2 (a) Let $X \sim \mathcal{N}(0, \sigma_X^2)$ and $Y \sim \mathcal{N}(0, \sigma_Y^2)$ be independent zero-mean Gaussian rvs. By convolving their densities, find the density of $X + Y$. [Hint. In performing the integration for the convolution, you should do something called "completing the square" in the exponent. This involves multiplying and dividing by $e^{\alpha y^2/2}$ for some α, and you can be guided in this by knowing what the answer is. This technique is invaluable in working with Gaussian rvs.]

(b) The Fourier transform of a probability density $f_X(x)$ is $\hat{f}_X(\theta) = \int f_X(x) e^{-2\pi i x\theta}\, dx = \mathsf{E}[e^{-2\pi i X\theta}]$. By scaling the basic Gaussian transform in (4.48), show that, for $X \sim \mathcal{N}(0, \sigma_X^2)$,

$$\hat{f}_X(\theta) = \exp\left(-\frac{(2\pi\theta)^2 \sigma_X^2}{2}\right).$$

(c) Now find the density of $X + Y$ by using Fourier transforms of the densities.

(d) Using the same Fourier transform technique, find the density of $V = \sum_{k=1}^{n} \alpha_k W_k$, where $W_1, \ldots, W_k$ are independent normal rvs.

7.3 In this exercise you will construct two rvs that are individually Gaussian but not jointly Gaussian. Consider the nonnegative random variable X with density given by

$$f_X(x) = \sqrt{\frac{2}{\pi}} \exp\left(\frac{-x^2}{2}\right), \qquad \text{for } x \geq 0.$$

Let U be binary, ±1, with $p_U(1) = p_U(-1) = 1/2$.

(a) Find the probability density of $Y_1 = UX$. Sketch the density of Y_1 and find its mean and variance.

(b) Describe two normalized Gaussian rvs, say Y_1 and Y_2, such that the joint density of Y_1, Y_2 is *zero* in the second and fourth quadrants of the plane. It is nonzero in the first and third quadrants where it has the density $(1/\pi)\exp(-y_1^2/2 - y_2^2/2)$. Are Y_1, Y_2 jointly Gaussian? [Hint. Use part (a) for Y_1 and think about how to construct Y_2.]

(c) Find the covariance $\mathsf{E}[Y_1 Y_2]$. [Hint. First find the mean of the rv X above.]

(d) Use a variation of the same idea to construct two normalized Gaussian rvs V_1, V_2 whose probability is concentrated on the diagonal axes $v_1 = v_2$ and $v_1 =-v_2$, i.e. for which $\Pr(V_1 \neq V_2 \,\text{and}\, V_1 \neq -V_2) = 0$. Are V_1, V_2 jointly Gaussian?

7.4 Let $W_1 \sim \mathcal{N}(0,1)$ and $W_2 \sim \mathcal{N}(0,1)$ be independent normal rvs. Let $X = \max(W_1, W_2)$ and $Y = \min(W_1, W_2)$.

(a) Sketch the transformation from sample values of W_1, W_2 to sample values of X, Y. Which sample pairs w_1, w_2 of W_1, W_2 map into a given sample pair x, y of X, Y?

(b) Find the probability density $f_{XY}(x,y)$ of X, Y. Explain your argument briefly but work from your sketch rather than equations.

(c) Find $f_S(s)$, where $S = X + Y$.

(d) Find $f_D(d)$, where $D = X - Y$.

(e) Let U be a random variable taking the values ± 1 with probability 1/2 each and let U be statistically independent of W_1, W_2. Are S and UD jointly Gaussian?

7.5 Let $\phi(t)$ be an $\mathcal{L}_1$ and $\mathcal{L}_2$ function of energy 1 and let $h(t)$ be $\mathcal{L}_1$ and $\mathcal{L}_2$. Show that $\int_{-\infty}^{\infty} \phi(t)h(\tau - t)\mathrm{d}t$ is an $\mathcal{L}_2$ function of τ. [Hint. Consider the Fourier transform of $\phi(t)$ and $h(t)$.]

7.6 (a) Generalize the random process of (7.30) by assuming that the Z_k are arbitrarily correlated. Show that every sample function is still $\mathcal{L}_2$.

(b) For this same case, show that $\iint |\mathsf{K}_Z(t,\tau)|^2 \,\mathrm{d}t\,\mathrm{d}\tau < \infty$.

7.7 (a) Let $Z_1, Z_2, \dots$ be a sequence of independent Gaussian rvs, $Z_k \sim \mathcal{N}(0, \sigma_k^2)$, and let $\{\phi_k(t) : \mathbb{R} \to \mathbb{R}\}$ be a sequence of orthonormal functions. Argue from fundamental definitions that, for each t, $Z(t) = \sum_{k=1}^{n} Z_k \phi_k(t)$ is a Gaussian rv. Find the variance of $Z(t)$ as a function of t.

(b) For any set of epochs $t_1, \dots, t_\ell$, let $Z(t_m) = \sum_{k=1}^{n} Z_k \phi_k(t_m)$ for $1 \le m \le \ell$. Explain carefully from the basic definitions why $\{Z(t_1), \dots, Z(t_\ell)\}$ are jointly Gaussian and specify their covariance matrix. Explain why $\{Z(t);\, t \in \mathbb{R}\}$ is a Gaussian random process.

(c) Now let $n = \infty$ in the definition of $Z(t)$ in part (a) and assume that $\sum_k \sigma_k^2 < \infty$. Also assume that the orthonormal functions are bounded for all k and t in the sense that, for some constant A, $|\phi_k(t)| \le A$ for all k and t. Consider the linear combination of rvs:

$$Z(t) = \sum_k Z_k \phi_k(t) = \lim_{n\to\infty} \sum_{k=1}^{n} Z_k \phi_k(t).$$

Let $Z^{(n)}(t) = \sum_{k=1}^{n} Z_k \phi_k(t)$. For any given t, find the variance of $Z^{(j)}(t) - Z^{(n)}(t)$ for $j > n$. Show that, for all $j > n$, this variance approaches 0 as $n \to \infty$. Explain intuitively why this indicates that $Z(t)$ is a Gaussian rv. Note: $Z(t)$ is, in fact, a Gaussian rv, but proving this rigorously requires considerable background; $Z(t)$ is a limit of a sequence of rvs, and each rv is a function of a sample space – the issue here is the same as that of a sequence of functions going to a limit function, where we had to invoke the Riesz–Fischer theorem.

(d) For the above Gaussian random process $\{Z(t); t \in \mathbb{R}\}$, let $z(t)$ be a sample function of $Z(t)$ and find its energy, i.e. $\|z\|^2$, in terms of the sample values $z_1, z_2, \ldots$ of $Z_1, Z_2, \ldots$ Find the expected energy in the process, $\mathsf{E}[\|\{Z(t); t \in \mathbb{R}\}\|^2]$.

(e) Find an upperbound on $\Pr\{\|\{Z(t); t \in \mathbb{R}\}\|^2 > \alpha\}$ that goes to zero as $\alpha \to \infty$. [Hint. You might find the Markov inequality useful. This says that for a nonnegative rv Y, $\Pr\{Y \geq \alpha\} \leq \mathsf{E}[Y]/\alpha$.] Explain why this shows that the sample functions of $\{Z(t)\}$ are $\mathcal{L}_2$ with probability 1.

7.8 Consider a stochastic process $\{Z(t); t \in \mathbb{R}\}$ for which each sample function is a sequence of rectangular pulses as in Figure 7.8. Analytically, $Z(t) = \sum_{k=-\infty}^{\infty} Z_k \operatorname{rect}(t-k)$, where $\ldots Z_{-1}, Z_0, Z_1, \ldots$ is a sequence of iid normal variables, $Z_k \sim \mathcal{N}(0,1)$.

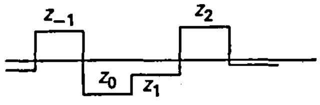


Figure 7.8.

(a) Is $\{Z(t);\ t \in \mathbb{R}\}$ a Gaussian random process? Explain why or why not carefully.

(b) Find the covariance function of $\{Z(t);\ t \in \mathbb{R}\}$.

(c) Is $\{Z(t);\ t \in \mathbb{R}\}$ a stationary random process? Explain carefully.

(d) Now suppose the stochastic process is modified by introducing a random time shift Φ which is uniformly distributed between 0 and 1. Thus, the new process $\{V(t);\ t \in \mathbb{R}\}$ is defined by $V(t) = \sum_{k=-\infty}^{\infty} Z_k \operatorname{rect}(t-k-\Phi)$. Find the conditional distribution of $V(0.5)$ conditional on $V(0) = v$.

(e) Is $\{V(t);\ t \in \mathbb{R}\}$ a Gaussian random process? Explain why or why not carefully.

(f) Find the covariance function of $\{V(t);\ t \in \mathbb{R}\}$.

(g) Is $\{V(t);\ t \in \mathbb{R}\}$ a stationary random process? It is easier to explain this than to write a lot of equations.

7.9 Consider the Gaussian sinc process, $V(t) = \sum_k V_k \operatorname{sinc}[(t-kT)/T]$, where $\{\ldots, V_{-1}, V_0, V_1, \ldots,\}$ is a sequence of iid rvs, $V_k \sim \mathcal{N}(0, \sigma^2)$.

(a) Find the probability density for the linear functional $\int V(t)\,\mathrm{sinc}(t/T)\mathrm{d}t$.

(b) Find the probability density for the linear functional $\int V(t)\,\mathrm{sinc}(\alpha t/T)\mathrm{d}t$ for $\alpha > 1$.

(c) Consider a linear filter with impulse response $h(t) = \mathrm{sinc}(\alpha t/T)$, where $\alpha > 1$. Let $\{Y(t)\}$ be the output of this filter when $V(t)$ is the input. Find the covariance function of the process $\{Y(t)\}$. Explain why the process is Gaussian and why it is stationary.

(d) Find the probability density for the linear functional $Y(\tau) = \int V(t)\,\mathrm{sinc}(\alpha(t-\tau)/T)\mathrm{d}t$ for $\alpha \geq 1$ and arbitrary τ.

(e) Find the spectral density of $\{Y(t);\, t \in \mathbb{R}\}$.

(f) Show that $\{Y(t);\, t \in \mathbb{R}\}$ can be represented as $Y(t) = \sum_k Y_k\,\mathrm{sinc}[(t-kT)/T]$ and characterize the rvs $\{Y_k;\, k \in Z\}$.

(g) Repeat parts (c), (d), and (e) for $\alpha < 1$.

(h) Show that $\{Y(t)\}$ in the $\alpha < 1$ case can be represented as a Gaussian sinc process (like $\{V(t)\}$ but with an appropriately modified value of T).

(i) Show that if any given process $\{Z(t);\, t \in \mathbb{R}\}$ is stationary, then so is the process $\{Y(t);\, t \in \mathbb{R}\}$, where $Y(t) = Z^2(t)$ for all $t \in \mathbb{R}$.

7.10 (Complex random variables)

(a) Suppose the zero-mean complex random variables X_k and X_{-k} satisfy $X^*_{-k} = X_k$ for all k. Show that if $\mathsf{E}[X_k X^*_{-k}] = 0$ then $\mathsf{E}[(\Re(X_k))^2] = \mathsf{E}[(\Im(X_k))^2]$ and $\mathsf{E}[\Re(X_k)\Im(X_{-k})] = 0$.

(b) Use this to show that if $\mathsf{E}[X_k X^*_m] = 0$ then $\mathsf{E}[\Re(X_k)\Re(X_m)] = 0$, $\mathsf{E}[\Re(X_k)\Im(X_m)] = 0$, and $\mathsf{E}[\Im(X_k)\Im(X_m)] = 0$ for all m not equal to either k or $-k$.

7.11 Explain why the integral in (7.58) must be real for $g_1(t)$ and $g_2(t)$ real, but the integrand $\hat{g}_1(f)S_Z(f)\hat{g}^*_2(f)$ need not be real.

7.12 (Filtered white noise) Let $\{Z(t)\}$ be a WGN process of spectral density $N_0/2$.

(a) Let $Y = \int_0^T Z(t)\,\mathrm{d}t$. Find the probability density of Y.

(b) Let $Y(t)$ be the result of passing $Z(t)$ through an ideal baseband filter of bandwidth W whose gain is adjusted so that its impulse response has unit energy. Find the joint distribution of $Y(0)$ and $Y(1/4\mathsf{W})$.

(c) Find the probability density of

$$V = \int_0^\infty \mathrm{e}^{-t} Z(t)\mathrm{d}t.$$

7.13 (Power spectral density)

(a) Let $\{\phi_k(t)\}$ be any set of real orthonormal $\mathcal{L}_2$ waveforms whose transforms are limited to a band B, and let $\{W(t)\}$ be WGN with respect to B with power spectral density $S_W(f) = N_0/2$ for $f \in B$. Let the orthonormal expansion of $W(t)$ with respect to the set $\{\phi_k(t)\}$ be defined by

$$\tilde{W}(t) = \sum_k W_k \phi_k(t),$$

where $W_k = \langle W(t), \phi_k(t)\rangle$. Show that $\{W_k\}$ is an iid Gaussian sequence, and give the probability distribution of each W_k.

(b) Let the band $B = [-1/2T, 1/2T]$, and let $\phi_k(t) = (1/\sqrt{T})\,\mathrm{sinc}[(t-kT)/T]$, $k \in \mathbb{Z}$. Interpret the result of part (a) in this case.

7.14 (Complex Gaussian vectors)

(a) Give an example of a 2D complex rv $\mathbf{Z} = (Z_1, Z_2)$, where $Z_k \sim \mathcal{CN}(0, 1)$ for $k = 1, 2$ and where $\mathbf{Z}$ has the same joint probability distribution as $e^{i\phi}\mathbf{Z}$ for all $\phi \in [0, 2\pi]$, but where $\mathbf{Z}$ is not jointly Gaussian and thus not circularly symmetric Gaussian. [Hint. Extend the idea in part (d) of Exercise 7.3.]

(b) Suppose a complex rv $Z = Z_{\mathrm{re}} + iZ_{\mathrm{im}}$ has the properties that Z_{re} and Z_{im} are individually Gaussian and that Z has the same probability density as $e^{i\phi}Z$ for all $\phi \in [0, 2\pi]$. Show that Z is complex circularly symmetric Gaussian.

8 Detection, coding, and decoding

8.1 Introduction

Chapter 7 showed how to characterize noise as a random process. This chapter uses that characterization to retrieve the signal from the noise-corrupted received waveform. As one might guess, this is not possible without occasional errors when the noise is unusually large. The objective is to retrieve the data while minimizing the effect of these errors. This process of retrieving data from a noise-corrupted version is known as *detection.*

Detection, decision making, hypothesis testing, and decoding are synonyms. The word *detection* refers to the effort to detect whether some phenomenon is present or not on the basis of observations. For example, a radar system uses observations to detect whether or not a target is present; a quality control system attempts to *detect* whether a unit is defective; a medical test *detects* whether a given disease is present. The meaning of detection has been extended in the digital communication field from a yes/no decision to a decision at the receiver between a finite set of possible transmitted signals. Such a decision between a set of possible transmitted signals is also called *decoding*, but here the possible set is usually regarded as the set of codewords in a code rather than the set of signals in a signal set.[1] *Decision making* is, again, the process of deciding between a number of mutually exclusive alternatives. *Hypothesis testing* is the same, but here the mutually exclusive alternatives are called hypotheses. We use the word hypotheses for the possible choices in what follows, since the word conjures up the appropriate intuitive image of making a choice between a set of alternatives, where only one alternative is correct and there is a possibility of erroneous choice.

These problems will be studied initially in a purely probabilistic setting. That is, there is a probability model within which each hypothesis is an event. These events are mutually exclusive and collectively exhaustive; i.e., the sample outcome of the experiment lies in one and only one of these events, which means that in each performance of the experiment, one and only one hypothesis is correct. Assume there are M hypotheses,[2] labeled $a_0, \dots, a_{M-1}$. The sample outcome of the experiment will

[1] As explained more fully later, there is no fundamental difference between a code and a signal set.

[2] The principles here apply essentially without change for a countably infinite set of hypotheses; for an uncountably infinite set of hypotheses, the process of choosing a hypothesis from an observation is called *estimation*. Typically, the probability of choosing correctly in this case is 0, and the emphasis is on making an estimate that is close in some sense to the correct hypothesis.

be one of these M events, and this defines a random symbol U which, for each m, takes the value a_m when event a_m occurs. The marginal probability $p_U(a_m)$ of hypothesis a_m is denoted by p_m and is called the *a-priori probability* of a_m. There is also a random variable (rv) V, called the observation. A sample value v of V is observed, and on the basis of that observation the detector selects one of the possible M hypotheses. The observation could equally well be a complex random variable, a random vector, a random process, or a random symbol; these generalizations are discussed in what follows.

Before discussing how to make decisions, it is important to understand when and why decisions must be made. For a binary example, assume that the conditional probability of hypothesis a_0 given the observation is 2/3 and that of hypothesis a_1 is 1/3. Simply deciding on hypothesis a_0 and forgetting about the probabilities throws away the information about the probability that the decision is correct. However, actual decisions sometimes must be made. In a communication system, the user usually wants to receive the message (even partly garbled) rather than a set of probabilities. In a control system, the controls must occasionally take action. Similarly, managers must occasionally choose between courses of action, between products, and between people to hire. In a sense, it is by making decisions that we return from the world of mathematical probability models to the world being modeled.

There are a number of possible criteria to use in making decisions. Initially assume that the criterion is to maximize the probability of correct choice. That is, when the experiment is performed, the resulting experimental outcome maps into both a sample value a_m for U and a sample value v for V. The decision maker observes v (but not a_m) and maps v into a decision $\tilde{u}(v)$. The decision is correct if $\tilde{u}(v) = a_m$. In principle, maximizing the probability of correct choice is almost trivially simple. Given v, calculate $p_{U|V}(a_m|v)$ for each possible hypothesis a_m. This *is* the probability that a_m is the correct hypothesis conditional on v. Thus the rule for maximizing the probability of being correct is to choose $\tilde{u}(v)$ to be that a_m for which $p_{U|V}(a_m|v)$ is maximized. For each possible observation v, this is denoted by

$$\tilde{u}(v) = \arg\max_m [p_{U|V}(a_m|v)] \qquad \text{(MAP rule)}, \qquad (8.1)$$

where $\arg\max_m$ means the argument m that maximizes the function. If the maximum is not unique, the probability of being correct is the same no matter which maximizing m is chosen, so, to be explicit, the smallest such m will be chosen.[3] Since the rule (8.1) applies to each possible sample output v of the random variable V, (8.1) also defines the selected hypothesis as a random symbol $\tilde{U}(V)$. The conditional probability $p_{U|V}$ is called an *a-posteriori probability*. This is in contrast to the *a-priori probability* p_U of the hypothesis before the observation of V. The decision rule in (8.1) is thus called the maximum a-posteriori probability (MAP) rule.

[3] As discussed in Appendix 8.10, it is sometimes desirable to choose randomly among the maximum a-posteriori choices when the maximum in (8.1) is not unique. There are often situations (such as with discrete coding and decoding) where nonuniqueness occurs with positive probability.

An important consequence of (8.1) is that the MAP rule depends only on the conditional probability $p_{U|V}$ and thus is completely determined by the joint distribution of U and V. Everything else in the probability space is irrelevant to making a MAP decision.

When distinguishing between different decision rules, the MAP decision rule in (8.1) will be denoted by $\tilde{u}_{\text{MAP}}(v)$. Since the MAP rule maximizes the probability of correct decision for each sample value v, it also maximizes the probability of correct decision averaged over all v. To see this analytically, let $\tilde{u}_D(v)$ be an arbitrary decision rule. Since $\tilde{u}_{\text{MAP}}$ maximizes $p_{U|V}(m \mid v)$ over m,

$$p_{U|V}(\tilde{u}_{\text{MAP}}(v)|v) - p_{U|V}(\tilde{u}_D(v)|v) \geq 0; \qquad \text{for each rule } D \text{ and observation } v. \tag{8.2}$$

Taking the expected value of the first term on the left over the observation V, we get the probability of correct decision using the MAP decision rule. The expected value of the second term on the left for any given D is the probability of correct decision using that rule. Thus, taking the expected value of (8.2) over V shows that the MAP rule maximizes the probability of correct decision over the observation space. The above results are very simple, but also important and fundamental. They are summarized in the following theorem.

Theorem 8.1.1 *The MAP rule, given in (8.1), maximizes the probability of a correct decision, both for each observed sample value v and as an average over V. The MAP rule is determined solely by the joint distribution of U and V.*

Before discussing the implications and use of the MAP rule, the above assumptions are reviewed. First, a probability model was assumed in which all probabilities are known, and in which, for each performance of the experiment, one and only one hypothesis is correct. This conforms very well to the communication model in which a transmitter sends one of a set of possible signals and the receiver, given signal plus noise, makes a decision on the signal actually sent. It does not always conform well to a scientific experiment attempting to verify the existence of some new phenomenon; in such situations, there is often no sensible way to model a-priori probabilities. Detection in the absence of known a-priori probabilities is discussed in Appendix 8.10.

The next assumption was that maximizing the probability of correct decision is an appropriate decision criterion. In many situations, the cost of a wrong decision is highly asymmetric. For example, when testing for a treatable but deadly disease, making an error when the disease is present is far more costly than making an error when the disease is not present. As shown in Exercise 8.1, it is easy to extend the theory to account for relative costs of errors..

With the present assumptions, the detection problem can be stated concisely in the following probabilistic terms. There is an underlying sample space Ω, a probability measure, and two rvs U and V of interest. The corresponding experiment is performed, an observer sees the sample value v of rv V, but does not observe anything else, particularly not the sample value of U, say a_m. The observer uses a detection rule, $\tilde{u}(v)$, which is a function mapping each possible value of v to a possible value of U.

If $\tilde{v}(v) = a_m$, the detection is correct; otherwise an error has been made. The above MAP rule maximizes the probability of correct detection conditional on each v and also maximizes the unconditional probability of correct detection. Obviously, the observer must know the conditional probability assignment $p_{U|V}$ in order to use the MAP rule.

Sections 8.2 and 8.3 are restricted to the case of binary hypotheses where $M = 2$. This allows us to understand most of the important ideas, but simplifies the notation considerably. This is then generalized to an arbitrary number of hypotheses; fortunately, this extension is almost trivial.

8.2 Binary detection

Assume a probability model in which the correct hypothesis U is a binary random variable with possible values $\{a_0, a_1\}$ and a-priori probabilities p_0 and p_1. In the communication context, the a-priori probabilities are usually modeled as equiprobable, but occasionally there are multistage detection processes in which the result of the first stage can be summarized by a new set of a-priori probabilities. Thus let p_0 and $p_1 = 1 - p_0$ be arbitrary. Let V be a rv with a conditional probability density $f_{V|U}(v \mid a_m)$ that is finite and nonzero for all $v \in \mathbb{R}$ and $m \in \{0, 1\}$. The modifications for zero densities, discrete V, complex V, or vector V are relatively straightforward and are discussed later.

The conditional densities $f_{V|U}(v \mid a_m)$, $m \in \{0, 1\}$, are called *likelihoods* in the jargon of hypothesis testing. The marginal density of V is given by $f_V(v) = p_0 f_{V|U}(v \mid a_0) + p_1 f_{V|U}(v \mid a_1)$. The a-posteriori probability of U, for $m = 0$ or 1, is given by

$$p_{U|V}(a_m|v) = \frac{p_m f_{V|U}(v \mid a_m)}{f_V(v)}. \tag{8.3}$$

Writing out (8.1) explicitly for this case, we obtain

$$\frac{p_0 f_{V|U}(v \mid a_0)}{f_V(v)} \begin{array}{c} \geq^{\tilde{U}=a_0} \\ <_{\tilde{U}=a_1} \end{array} \frac{p_1 f_{V|U}(v \mid a_1)}{f_V(v)}. \tag{8.4}$$

This "equation" indicates that the MAP decision is a_0 if the left side is greater than or equal to the right, and is a_1 if the left side is less than the right. Choosing the decision $\tilde{U} = a_0$ when equality holds in (8.4) is an arbitrary choice and does not affect the probability of being correct. Canceling $f_V(v)$ and rearranging, we obtain

$$\Lambda(v) = \frac{f_{V|U}(v \mid a_0)}{f_{V|U}(v \mid a_1)} \begin{array}{c} \geq^{\tilde{U}=a_0} \\ <_{\tilde{U}=a_1} \end{array} \frac{p_1}{p_0} = \eta. \tag{8.5}$$

The ratio $\Lambda(v) = f_{V|U}(v \mid a_0)/f_{V|U}(v \mid a_1)$ is called the *likelihood ratio*, and is a function only of v. The ratio $\eta = p_1/p_0$ is called the *threshold* and depends only on the a-priori probabilities. The binary MAP rule (or MAP test, as it is usually called) then compares the likelihood ratio to the threshold, and decides on hypothesis a_0 if the threshold is reached, and on hypothesis a_1 otherwise. Note that if the a-priori probability p_0 is increased, the threshold decreases, and the set of v for which hypothesis a_0 is chosen increases; this corresponds to our intuition – the more certain we are initially that U is 0, the stronger the evidence required to make us change our minds. As shown in Exercise 8.1, the only effect of minimizing over costs rather than error probability is to change the threshold η in (8.5).

An important special case of (8.5) is that in which $p_0 = p_1$. In this case $\eta = 1$, and the rule chooses $\tilde{U}(v) = a_0$ for $f_{V|U}(v|a_0) \geq f_{V|U}(v|a_1)$ and chooses $\tilde{U}(v) = 1$ otherwise. This is called a *maximum likelihood (ML) rule* or *test*. In the communication case, as mentioned above, the a-priori probabilities are usually equal, so MAP then reduces to ML. The maximum likelihood test is also often used when p_0 and p_1 are unknown.

The *probability of error*, i.e. one minus the probability of choosing correctly, is now derived for MAP detection. First we find the probability of error conditional on each hypothesis, $\Pr\{e|U = a_1\}$ and $\Pr\{e|U = a_0\}$. The overall probability of error is then given by

$$\Pr\{e\} = p_0 \Pr\{e|U = a_0\} + p_1 \Pr\{e|U = a_1\}.$$

In the radar field, $\Pr\{e|U = a_0\}$ is called the probability of false alarm, and $\Pr\{e|U = a_1\}$ is called the probability of a miss. Also $1 - \Pr\{e|U = a_1\}$ is called the probability of detection. In statistics, $\Pr\{e|U = a_1\}$ is called the probability of error of the second kind, and $\Pr\{e|U = a_0\}$ is the probability of error of the first kind. These terms are not used here.

Note that (8.5) partitions the space of observed sample values into two regions: $R_0 = \{v : \Lambda(v) \geq \eta\}$ is the region for which $\tilde{U} = a_0$ and $R_1 = \{v : \Lambda(v) < \eta\}$ is the region for which $\tilde{U} = a_1$. For $U = a_1$, an error occurs if and only if v is in R_0, and for $U = a_0$ an error occurs if and only if v is in R_1. Thus,

$$\Pr\{e|U = a_0\} = \int_{R_1} f_{V|U}(v|a_0)\mathrm{d}v; \tag{8.6}$$

$$\Pr\{e|U = a_1\} = \int_{R_0} f_{V|U}(v|a_1)\mathrm{d}v. \tag{8.7}$$

Another, often simpler, approach is to work directly with the likelihood ratio. Since $\Lambda(v)$ is a function of the observed sample value v, the random variable, $\Lambda(V)$, also called a likelihood ratio, is defined as follows: for every sample point ω, $V(\omega)$ is the corresponding sample value v, and $\Lambda(V)$ is then shorthand for $\Lambda(V(\omega))$. In the same way, $\tilde{U}(V)$ (or more briefly $\tilde{U}$) is the decision random variable. In these terms, (8.5) states that

$$\tilde{U} = a_0 \quad \text{if and only if} \quad \Lambda(V) \geq \eta. \tag{8.8}$$

Thus, for MAP detection with a threshold η,

$$\Pr\{e|U = a_0\} = \Pr\{\tilde{U} = a_1|U = a_0\} = \Pr\{\Lambda(V) < \eta|U = a_0\}; \tag{8.9}$$

$$\Pr\{e|U = a_1\} = \Pr\{\tilde{U} = a_0|U = a_1\} = \Pr\{\Lambda(V) \geq \eta|U = a_1\}. \tag{8.10}$$

A *sufficient statistic* is defined as any function of the observation v from which the likelihood ratio can be calculated. As examples, v itself, $\Lambda(v)$, and any one-to-one function of $\Lambda(v)$ are sufficient statistics. Note that $\Lambda(v)$, and functions of $\Lambda(v)$, are often simpler to work with than v, since $\Lambda(v)$ is simply a real number, whereas v could be a vector or a waveform.

We have seen that the MAP rule (and thus also the ML rule) is a threshold test on the likelihood ratio. Similarly the min-cost rule (see Exercise 8.1) and the Neyman–Pearson test (which, as shown in Appendix 8.10, makes no assumptions about a-priori probabilities) are threshold tests on the likelihood ratio. Not only are all these binary decision rules based only on threshold tests on the likelihood ratio, but the properties of these rules, such as the conditional error probabilities in (8.9) and (8.10), are based only on $\Lambda(V)$ and η. In fact, it is difficult to imagine any sensible binary decision procedure, especially in the digital communication context, that is not a threshold test on the likelihood ratio. Thus, once a sufficient statistic has been calculated from the observed vector, that observed vector has no further value in any decision rule of interest here.

The log likelihood ratio, $\mathrm{LLR}(V) = \ln[\Lambda(V)]$, is an important sufficient statistic which is often easier to work with than the likelihood ratio itself. As seen in Section 8.3, the LLR is particularly convenient for use with Gaussian noise statistics.

8.3 Binary signals in white Gaussian noise

This section first treats standard 2-PAM, then 2-PAM with an offset, then binary signals with vector observations, and finally binary signals with waveform observations.

8.3.1 Detection for PAM antipodal signals

Consider PAM antipodal modulation (i.e. 2-PAM), as illustrated in Figure 8.1. The correct hypothesis U is either $a_0 = a$ or $a_1 = -a$. Let $Z \sim \mathcal{N}(0, N_0/2)$ be a Gaussian noise rv of mean 0 and variance $N_0/2$, independent of U. That is,

$$f_z(z) = \frac{1}{\sqrt{2\pi N_0/2}} \exp\left(\frac{-z^2}{N_0}\right).$$

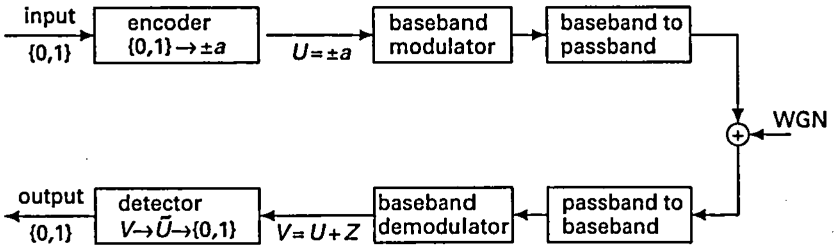


Figure 8.1. The source produces a binary digit, which is mapped into $U = \pm a$. This is modulated into a waveform, WGN is added, and the resultant waveform is demodulated and sampled, resulting in a noisy received value $V = U + Z$. From Section 7.8, $Z \sim \mathcal{N}(0, N_0/2)$. This is explained more fully later. Based on this observation, the receiver makes a decision $\tilde{U}$ and maps this back to the binary output, which is the hypothesized version of the binary input.

Assume that 2-PAM is simplified by sending only a single binary symbol (rather than a sequence over time) and by observing only the single sample value v corresponding to that input. As seen later, these simplifications are unnecessary, but they permit the problem to be viewed in the simplest possible context. The observation V (i.e. the channel output prior to detection) is $a+Z$ or $-a+Z$, depending on whether $U=a$ or $-a$. Thus, conditional on $U=a$, $V\sim\mathcal{N}(a,N_0/2)$ and, conditional on $U=-a$, $V\sim\mathcal{N}(-a,N_0/2)$:

$$f_{V|U}(v|a)=\frac{1}{\sqrt{\pi N_0}}\exp\left(\frac{-(v-a)^2}{N_0}\right); \qquad f_{V|U}(v|-a)=\frac{1}{\sqrt{\pi N_0}}\exp\left(\frac{-(v+a)^2}{N_0}\right).$$

The likelihood ratio is the ratio of these likelihoods, and is given by

$$\Lambda(v)=\exp\left(\frac{-(v-a)^2+(v+a)^2}{N_0}\right)=\exp\left(\frac{4av}{N_0}\right). \tag{8.11}$$

Substituting this into (8.5), we obtain

$$\exp\left(\frac{4av}{N_0}\right)\ \mathop{\substack{\geq^{\tilde{U}=a}\\ <_{\tilde{U}=-a}}}\ \frac{p_1}{p_0}=\eta. \tag{8.12}$$

This is further simplified by taking the logarithm, yielding

$$\mathrm{LLR}(v)=\frac{4av}{N_0}\ \mathop{\substack{\geq^{\tilde{U}=a}\\ <_{\tilde{U}=-a}}}\ \ln\eta; \tag{8.13}$$

$$v\ \mathop{\substack{\geq^{\tilde{U}=a}\\ <_{\tilde{U}=-a}}}\ \frac{N_0\ln\eta}{4a}. \tag{8.14}$$

Figure 8.2 interprets this decision rule.

The probability of error, given $U=-a$, is seen to be the probability that the noise value is greater than $a+N_0\ln\eta/4a$. Since the noise has variance $N_0/2$, this is the probability that the normalized Gaussian rv $Z/\sqrt{N_0/2}$ exceeds $a/\sqrt{N_0/2}+\sqrt{N_0/2}\,\ln(\eta)/2a$. Thus,

$$\Pr\{e\,|\,U=-a\}=Q\left(\frac{a}{\sqrt{N_0/2}}+\frac{\sqrt{N_0/2}\ln\eta}{2a}\right). \tag{8.15}$$

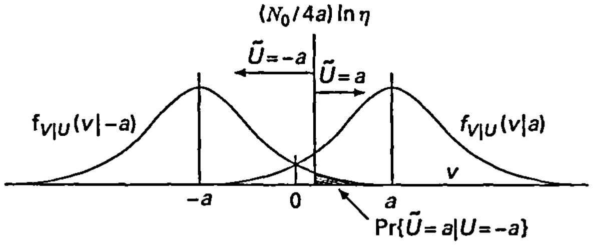


Figure 8.2. Binary hypothesis testing for antipodal signals, $0\to a$, $1\to -a$. The a-priori probabilities are p_0 and p_1, the threshold is $\eta=p_0/p_1$, and the noise is $\mathcal{N}(0,N_0/2)$.

where $Q(x)$, the complementary distribution function of $\mathcal{N}(0,1)$, is given by

$$Q(x) = \int_x^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left(\frac{-z^2}{2}\right) dz.$$

The probability of error given $U = a$ is calculated the same way, but is the probability that $-Z$ is greater than or equal to $a - N_0 \ln \eta / 4a$. Since $-Z$ has the same distribution as Z,

$$\Pr\{e | U = a\} = Q\left(\frac{a}{\sqrt{N_0/2}} - \frac{\sqrt{N_0/2} \ln \eta}{2a}\right). \tag{8.16}$$

It is more insightful to express $a/\sqrt{N_0/2}$ as $\sqrt{2a^2/N_0}$. As seen before, a^2 can be viewed as the energy per bit, E_b, so that (8.15) and (8.16) become

$$\Pr\{e | U = -a\} = Q\left(\sqrt{\frac{2E_b}{N_0}} + \frac{\ln \eta}{2\sqrt{2E_b/N_0}}\right), \tag{8.17}$$

$$\Pr\{e | U = a\} = Q\left(\sqrt{\frac{2E_b}{N_0}} - \frac{\ln \eta}{2\sqrt{2E_b/N_0}}\right). \tag{8.18}$$

Note that these formulas involve only the ratio E_b/N_0 rather than E_b or N_0 separately. If the signal, observation, and noise had been measured on a different scale, then both E_b and N_0 would change by the same factor, helping to explain why only the ratio is relevant. In fact, the scale could be normalized so that either the noise has variance 1 or the signal has variance 1.

The hypotheses in these communication problems are usually modeled as equiprobable, $p_0 = p_1 = 1/2$. In this case, $\ln \eta = 0$ and the MAP rule is equivalent to the ML rule. Equations (8.17) and (8.18) then simplify to the following:

$$\Pr\{e\} = \Pr\{e | U = -a\} = \Pr\{e | U = a\} = Q\left(\sqrt{\frac{2E_b}{N_0}}\right). \tag{8.19}$$

In terms of Figure 8.2, this is the tail of either Gaussian distribution from the point 0 where they cross. This equation keeps reappearing in different guises, and it will soon seem like a completely obvious result for a variety of Gaussian detection problems.

8.3.2 Detection for binary nonantipodal signals

Next consider the slightly more complex case illustrated in Figure 8.3. Instead of mapping 0 to $+a$ and 1 to $-a$, 0 is mapped to an arbitrary number b_0 and 1 to an arbitrary number b_1. To analyze this, let c be the midpoint between b_0 and b_1, $c = (b_0 + b_1)/2$. Assuming $b_1 < b_0$, let $a = b_0 - c = c - b_1$. Conditional on $U = b_0$, the observation is $V = c + a + Z$; conditional on $U = b_1$, it is $V = c - a + Z$. In other words, this more general case is simply the result of shifting the previous signals by the constant c.

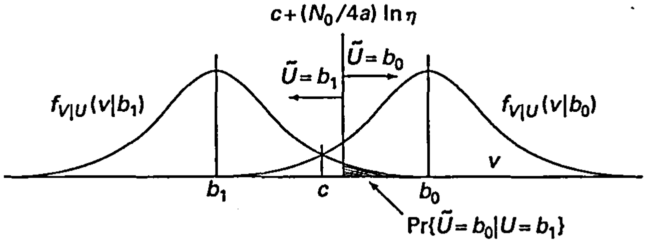


Figure 8.3. Binary hypothesis testing for arbitrary signals, $0 \to b_0$, $1 \to b_1$, for $b_0 > b_1$. With $c=(b_0+b_1)/2$ and $a=|b_0-b_1|/2$, this is the same as Figure 8.2 shifted by c. For $b_0 < b_1$, the picture must be reversed, but the answer is the same.

Define $\tilde{V} = V - c$ as the result of shifting the observation by $-c$; $\tilde{V}$ is a sufficient statistic and $\tilde{V} = \pm a + Z$. This is the same as the antipodal signal case in Section 8.3.1, so the error probability is again given by (8.15) and (8.16).

The energy used in achieving this error probability has changed from the antipodal case. The energy per bit (assuming equal a-priori probabilities) is now $(b_0^2 + b_1{}^2)/2 = a^2 + c^2$. A center value c is frequently used as a "pilot tone" in communication for tracking the channel. We see that E_b is then the sum of the energy used for the actual binary transmission (a^2) plus the energy used for the pilot tone (c^2). The fraction of energy E_b used for the signal is $\gamma = a^2/(a^2+c^2)$, and (8.17) and (8.18) are changed as follows:

$$\Pr\{e|U=b_1\} = Q\left(\sqrt{\frac{2\gamma E_b}{N_0}} + \frac{\ln\eta}{2\sqrt{2\gamma E_b/N_0}}\right), \tag{8.20}$$

$$\Pr\{e|U=b_0\} = Q\left(\sqrt{\frac{2\gamma E_b}{N_0}} - \frac{\ln\eta}{2\sqrt{2\gamma E_b/N_0}}\right). \tag{8.21}$$

For example, a common binary communication technique called *on-off keying* uses the binary signals 0 and $2a$. In this case, $\gamma = 1/2$ and there is an energy loss of 3 dB from the antipodal case. For the ML rule, the probability of error then becomes $Q(\sqrt{E_b/N_0})$.

8.3.3 Detection for binary real vectors in WGN

Next consider the vector version of the Gaussian detection problem. Suppose the observation is a random n-vector $\boldsymbol{V} = \boldsymbol{U} + \boldsymbol{Z}$. The noise $\boldsymbol{Z}$ is a random n-vector $(Z_1, Z_2, \dots, Z_n)^{\mathsf{T}}$, independent of $\boldsymbol{U}$, with iid components given by $Z_k \sim \mathcal{N}(0, N_0/2)$. The input $\boldsymbol{U}$ is a random n-vector with M possible values (hypotheses). The mth hypothesis, $0 \le m \le M-1$, is denoted by $\boldsymbol{a}_m = (a_{m1}, a_{m2}, \dots, a_{mn})^{\mathsf{T}}$. A sample value $\boldsymbol{v}$ of $\boldsymbol{V}$ is observed, and the problem is to make a MAP decision, denoted by $\tilde{\boldsymbol{U}}$, about $\boldsymbol{U}$.

Initially assume the binary antipodal case, where $\boldsymbol{a}_1 = -\boldsymbol{a}_0$. For notational simplicity, let $\boldsymbol{a}_0$ be denoted by $\boldsymbol{a} = (a_1, a_2, \dots, a_n)^{\mathsf{T}}$. Thus the two hypotheses are $\boldsymbol{U} = \boldsymbol{a}$

and $U = -a$, and the observation is either $a+Z$ or $-a+Z$. The likelihoods are then given by

$$f_{V|U}(v|a) = \frac{1}{(\pi N_0)^{n/2}} \exp \sum_{k=1}^{n} \frac{-(v_k - a_k)^2}{N_0} = \frac{1}{(\pi N_0)^{n/2}} \exp\left(\frac{-\|v-a\|^2}{N_0}\right),$$

$$f_{V|U}(v|-a) = \frac{1}{(\pi N_0)^{n/2}} \exp \sum_{k=1}^{n} \frac{-(v_k + a_k)^2}{N_0} = \frac{1}{(\pi N_0)^{n/2}} \exp\left(\frac{-\|v+a\|^2}{N_0}\right).$$

The log likelihood ratio is thus given by

$$\text{LLR}(v) = \frac{-\|v-a\|^2 + \|v+a\|^2}{N_0} = \frac{4\langle v, a\rangle}{N_0}, \tag{8.22}$$

and the MAP test is

$$\text{LLR}(v) = \frac{4\langle v, a\rangle}{N_0} \underset{<_{\tilde{U}=-a}}{\geq^{\tilde{U}=a}} \ln \frac{p_1}{p_0} = \ln \eta.$$

This can be restated as follows:

$$\frac{\langle v, a\rangle}{\|a\|} \underset{<_{\tilde{U}=-a}}{\geq^{\tilde{U}=a}} \frac{N_0 \ln \eta}{4\|a\|}. \tag{8.23}$$

The projection of the observation v onto the signal a is $(\langle v, a\rangle/\|a\|)(a/\|a\|)$. Thus the left side of (8.23) is the component of v in the direction of a, showing that the decision is based solely on that component of v. This result is rather natural; the noise is independent in different orthogonal directions, and only the noise in the direction of the signal should be relevant in detecting the signal.

The geometry of the situation is particularly clear in the ML case (see Figure 8.4). The noise is spherically symmetric around the origin, and the likelihoods depend only

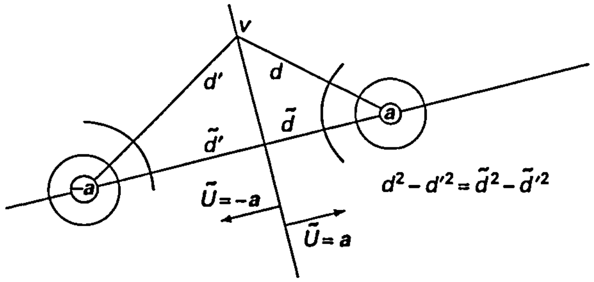


Figure 8.4. ML decision regions for binary signals in WGN. A vector v on the threshold boundary is shown. The distance from v to a is $d = \|v-a\|$. Similarly, the distance to $-a$ is $d' = \|v+a\|$. As shown algebraically in (8.22), any point at which $d^2 - d'^2 = 0$ is a point at which $\langle v, a\rangle = 0$, and thus at which the LLR is 0. Geometrically, from the Pythagorean theorem, however, $d^2 - d'^2 = \tilde{d}^2 - \tilde{d}'^2$, where $\tilde{d}$ and $\tilde{d}'$ are the distances from a and $-a$ to the projection of v on the straight line generated by a. This demonstrates geometrically why it is only the projection of v onto a that is relevant.

on the distance from the origin. The ML detection rule is then equivalent to choosing the hypothesis closest to the received point. The set of points equidistant from the two hypotheses, as illustrated in Figure 8.4, is the perpendicular bisector between them; this bisector is the set of $\boldsymbol{v}$ satisfying $\langle \boldsymbol{v}, \boldsymbol{a}\rangle = 0$. The set of points closer to $\boldsymbol{a}$ is on the $\boldsymbol{a}$ side of this perpendicular bisector; it is determined by $\langle \boldsymbol{v}, \boldsymbol{a}\rangle > 0$ and is mapped into $\boldsymbol{a}$ by the ML rule. Similarly, the set of points closer to $-\boldsymbol{a}$ is determined by $\langle \boldsymbol{v}, \boldsymbol{a}\rangle < 0$, and is mapped into $-\boldsymbol{a}$. In the general MAP case, the region mapped into $\boldsymbol{a}$ is again separated from the region mapped into $-\boldsymbol{a}$ by a perpendicular to $\boldsymbol{a}$, but in this case it is the perpendicular defined by $\langle \boldsymbol{v}, \boldsymbol{a}\rangle = N_0 \ln(\eta)/4$.

Another way of interpreting (8.23) is to view it in a different coordinate system. That is, choose $\boldsymbol{\phi}_1 = \boldsymbol{a}/\|\boldsymbol{a}\|$ as one element of an orthonormal basis for the n-vectors, and choose another $n-1$ orthonormal vectors by the Gram–Schmidt procedure. In this new coordinate system, $\boldsymbol{v}$ can be expressed as $(v'_1, v'_2, \dots, v'_n)^{\mathsf{T}}$, where, for $1 \le k \le n$, $v'_k = \langle \boldsymbol{v}, \boldsymbol{\phi}_k\rangle$. Since $\langle \boldsymbol{v}, \boldsymbol{a}\rangle = \|\boldsymbol{a}\|\langle \boldsymbol{v}, \boldsymbol{\phi}_1\rangle = \|\boldsymbol{a}\| v'_1$, the left side of (8.23) is simply v'_1, i.e. the size of the projection of $\boldsymbol{v}$ onto $\boldsymbol{a}$. Thus (8.23) becomes

$$v'_1 \mathop{\gtrless}_{<_{\tilde{U}=1}}^{\geq^{\tilde{U}=0}} \frac{N_0 \ln \eta}{4\|\boldsymbol{a}\|}.$$

This is the same as the scalar MAP test in (8.14). In other words, the vector problem is the same as the scalar problem when the appropriate coordinate system is chosen. Actually, the derivation of (8.23) has shown something more, namely that v'_1 is a sufficient statistic. The components $v'_2, \dots, v'_n$, which contain only noise, cancel out in (8.22) if (8.22) is expressed in the new coordinate system. The fact that the coordinates of $\boldsymbol{v}$ in directions orthogonal to the signal do not affect the LLR is sometimes called the *theorem of irrelevance*. A generalized form of this theorem is stated later as Theorem 8.4.2.

Some additional insight into (8.23) (in the original coordinate system) can be gained by writing $\langle \boldsymbol{v}, \boldsymbol{a}\rangle$ as $\sum_k v_k a_k$. This says that the MAP test weights each coordinate linearly by the amount of signal in that coordinate. This is not surprising, since the two hypotheses are separated more by the larger components of $\boldsymbol{a}$ than by the smaller.

Next consider the error probability conditional on $U = -\boldsymbol{a}$. Given $U = -\boldsymbol{a}$, $\boldsymbol{V} = -\boldsymbol{a} + \boldsymbol{Z}$, and thus

$$\frac{\langle \boldsymbol{V}, \boldsymbol{a}\rangle}{\|\boldsymbol{a}\|} = -\|\boldsymbol{a}\| + \langle \boldsymbol{Z}, \boldsymbol{\phi}_1\rangle.$$

Conditional on $U = -\boldsymbol{a}$, the mean and variance of $\langle \boldsymbol{V}, \boldsymbol{a}\rangle/\|\boldsymbol{a}\|$ are $-\|\boldsymbol{a}\|$ and $N_0/2$, respectively. Thus, $\langle \boldsymbol{V}, \boldsymbol{a}\rangle/\|\boldsymbol{a}\| \sim \mathcal{N}(-\|\boldsymbol{a}\|, N_0/2)$. From (8.23), the probability of error, given $U = -\boldsymbol{a}$, is the probability that $\mathcal{N}(-\|\boldsymbol{a}\|, N_0/2)$ exceeds $N_0 \ln(\eta)/4\,\|\boldsymbol{a}\|$. This is the probability that Z is greater than $\|\boldsymbol{a}\| + N_0 \ln(\eta)/(4\,\|\boldsymbol{a}\|)$. Normalizing as in Section 8.3.1, we obtain

$$\Pr\{e \mid U = -\boldsymbol{a}\} = Q\left(\sqrt{\frac{2\|\boldsymbol{a}\|^2}{N_0}} + \frac{\ln \eta}{2\sqrt{2\|\boldsymbol{a}\|^2/N_0}}\right). \tag{8.24}$$

By the same argument,

$$\Pr\{e|U=\boldsymbol{a}\} = Q\left(\sqrt{\frac{2\|\boldsymbol{a}\|^2}{N_0}} - \frac{\ln\eta}{2\sqrt{2\|\boldsymbol{a}\|^2/N_0}}\right). \tag{8.25}$$

It can be seen that this is the same answer as given by (8.15) and (8.16) when the problem is converted to a coordinate system where $\boldsymbol{a}$ is collinear with a coordinate vector. The energy per bit is $E_\mathrm{b} = \|\boldsymbol{a}\|^2$, so that (8.17) and (8.18) follow as before. This is not surprising, of course, since this vector decision problem is identical to the scalar problem when the appropriate basis is used.

For most communication problems, the a-priori probabilities are assumed to be equal, so that $\eta = 1$. Thus, as in (8.19),

$$\Pr\{e\} = Q\left(\sqrt{\frac{2E_\mathrm{b}}{N_0}}\right). \tag{8.26}$$

This gives us a useful sanity check – the probability of error does not depend on the orthonormal coordinate basis.

Now suppose that the binary hypotheses correspond to nonantipodal vector signals, say $\boldsymbol{b}_0$ and $\boldsymbol{b}_1$. We analyze this in the same way as the scalar case. Namely, let $\boldsymbol{c} = (\boldsymbol{b}_0 + \boldsymbol{b}_1)/2$ and $\boldsymbol{a} = \boldsymbol{b}_0 - \boldsymbol{c}$. Then the two signals are $\boldsymbol{b}_0 = \boldsymbol{a} + \boldsymbol{c}$ and $\boldsymbol{b}_1 = -\boldsymbol{a} + \boldsymbol{c}$. As before, converting the observation $\boldsymbol{V}$ to $\tilde{\boldsymbol{V}} = \boldsymbol{V} - \boldsymbol{c}$ shifts the midpoint and converts the problem back to the antipodal case. The error probability depends only on the distance $2\|\boldsymbol{a}\|$ between the signals, but the energy per bit, assuming equiprobable inputs, is $E_\mathrm{b} = \|\boldsymbol{a}\|^2 + \|\boldsymbol{c}\|^2$. Thus the center point $\boldsymbol{c}$ contributes to the energy, but not to the error probability. In the important special case where $\boldsymbol{b}_1$ and $\boldsymbol{b}_0$ are orthogonal and have equal energy, $\|\boldsymbol{a}\| = \|\boldsymbol{c}\|$ and

$$\Pr(e) = Q(\sqrt{E_\mathrm{b}/N_0}). \tag{8.27}$$

It is often more convenient, especially when dealing with $M > 2$ hypotheses, to express the LLR for the nonantipodal case directly in terms of $\boldsymbol{b}_0$ and $\boldsymbol{v}_1$. Using (8.22) for the shifted vector $\tilde{\boldsymbol{V}}$, the LLR can be expressed as follows:

$$\mathrm{LLR}(\boldsymbol{v}) = \frac{-\|\boldsymbol{v} - \boldsymbol{b}_0\|^2 + \|\boldsymbol{v} - \boldsymbol{b}_1^2\|^2}{N_0}. \tag{8.28}$$

For ML detection, this is simply the minimum-distance rule, and for MAP the interpretation is the same as for the antipodal case.

8.3.4 Detection for binary complex vectors in WGN

Next consider the complex vector version of the same problem. Assume the observation is a complex random n-vector $\boldsymbol{V} = \boldsymbol{U} + \boldsymbol{Z}$. The noise, $\boldsymbol{Z} = (Z_1, \ldots, Z_n)^\mathsf{T}$, is a complex random vector of n zero-mean complex iid Gaussian rvs with iid real and imaginary parts, each $\mathcal{N}(0, N_0/2)$. Thus each Z_k is circularly symmetric and denoted by $\mathcal{CN}(0, N_0)$. The input U is independent of Z and binary, taking on value a with probability p_0 and $-a$ with probability p_1, where $a = (a_1, \dots, a_n)^\mathsf{T}$ is an arbitrary complex n-vector.

This problem can be reduced to that of Section 8.3.3 by letting Z' be the $2n$-dimensional real random vector with components $\Re(Z_k)$ and $\Im(Z_k)$ for $1 \le k \le n$. Similarly, let a' be the $2n$-dimensional real vector with components $\Re(a_k)$ and $\Im(a_k)$ for $1 \le k \le n$ and let U' be the real random vector that takes on values a' or $-a'$. Finally, let $V' = U' + Z'$.

Recalling that probability densities for complex random variables or vectors are equal to the joint probability densities for the real and imaginary parts, we have

$$f_{V|U}(v|a) = f_{V'|U'}(v'|a') = \frac{1}{(\pi N_0)^n} \exp \sum_{k=1}^{n} \frac{-\Re(v_k - a_k)^2 - \Im(v_k - a_k)^2}{N_0};$$

$$f_{V|U}(v|-a) = f_{V'|U'}(v'|-a') = \frac{1}{(\pi N_0)^n} \exp \sum_{k=1}^{n} \frac{-\Re(v_k + a_k)^2 - \Im(v_k + a_k)^2}{N_0}.$$

The LLR is then given by

$$\mathrm{LLR}(v) = \frac{-\|v - a\|^2 + \|v + a\|^2}{N_0}. \tag{8.29}$$

Note that

$$\|v - a\|^2 = \|v\|^2 - \langle v, a\rangle - \langle a, v\rangle + \|a\|^2 = \|v\|^2 - 2\Re(\langle v, a\rangle) + \|a\|^2.$$

Using this and the analagous expression for $\|v + a\|^2$, (8.29) becomes

$$\mathrm{LLR}(v) = \frac{4\Re(\langle v, a\rangle)}{N_0}. \tag{8.30}$$

The MAP test can now be stated as follows:

$$\frac{\Re(\langle v, a\rangle)}{\|a\|} \underset{<_{\tilde{U}=-a}}{\geq^{\tilde{U}=a}} \frac{N_0 \ln \eta}{4\|a\|}. \tag{8.31}$$

Note that the value of the LLR and the form of the MAP test are the same as the real vector case except for the real part of $\langle v, a\rangle$. The significance of this real-part operation is now discussed.

In the n-dimensional complex vector space, $\langle v, a\rangle/\|a\|$ is the complex value of the projection of v in the direction of a. In order to understand this projection better, consider an orthonormal basis in which $a = (1, 0, 0, \dots, 0)^\mathsf{T}$. Then $\langle v, a\rangle/\|a\| = v_1$. Thus $\Re(v_1) = \pm 1 + \Re(z_1)$ and $\Im(v_1) = \Im(z_1)$. Clearly, only $\Re(v_1)$ is relevant to the binary decision. Using $\Re(\langle v, a\rangle/\|a\|)$ in (8.31) is simply the general way of stating this elementary idea. If the complex plane is viewed as a 2D real space, then taking the real part of $\langle v, a\rangle$ is equivalent to taking the further projection of this 2D real vector in the direction of a (see Exercise 8.12).

The other results and interpretations of Section 8.3.3 remain unchanged. In particular, since $\|a'\| = \|a\|$, the error probability results are given by

$$\Pr\{e \,|\, U = -a\} = Q\left(\sqrt{\frac{2\|a\|^2}{N_0}} + \frac{\ln \eta}{2\sqrt{2\|a\|^2/N_0}}\right); \tag{8.32}$$

$$\Pr\{e \,|\, U = a\} = Q\left(\sqrt{\frac{2\|a\|^2}{N_0}} - \frac{\ln \eta}{2\sqrt{2\|a\|^2/N_0}}\right). \tag{8.33}$$

For the ML case, recognizing that $\|a\|^2 = E_{\mathrm{b}}$, we have the following familiar result:

$$\Pr\{e\} = Q\left(\sqrt{\frac{2E_{\mathrm{b}}}{N_0}}\right). \tag{8.34}$$

Finally, for the nonantipodal case with hypotheses b_0 and b_1, the LLR is again given by (8.28).

8.3.5 Detection of binary antipodal waveforms in WGN

This section extends the vector case of Sections 8.3.3 and 8.3.4 to the waveform case. It will be instructive to do this simultaneously for both passband real random processes and baseband complex random processes. Let $U(t)$ be the baseband modulated waveform. As before, the situation is simplified by transmitting a single bit rather than a sequence of bits, so, for some arbitrary, perhaps complex, baseband waveform $a(t)$, the binary input 0 is mapped into $U(t) = a(t)$ and 1 is mapped into $U(t) = -a(t)$; the a-priori probabilities are denoted by p_0 and p_1. Let $\{\theta_k(t); k \in \mathbb{Z}\}$ be a complex orthonormal expansion covering the baseband region of interest, and let $a(t) = \sum_k a_k \theta_k(t)$.

Assume $U(t) = \pm a(t)$ is modulated onto a carrier f_c larger than the baseband bandwidth. The resulting bandpass waveform is denoted by $X(t) = \pm b(t)$, where, from Section 7.8, the modulated form of $a(t)$, denoted by $b(t)$, can be represented as

$$b(t) = \sum_k b_{k,1}\psi_{k,1}(t) + b_{k,2}\psi_{k,2}(t),$$

where

$$b_{k,1} = \Re(a_k); \qquad \psi_{k,1}(t) = \Re\{2\theta_k(t)\exp[2\pi i f_c t]\};$$
$$b_{k,2} = \Im(a_k); \qquad \psi_{k,2}(t) = -\Im\{2\theta_k(t)\exp[2\pi i f_c t]\}.$$

From Theorem 6.6.1, the set of waveforms $\{\psi_{k,j}(t); k \in \mathbb{Z}, j \in \{1,2\}\}$ are orthogonal, each with energy 2. Let $\{\phi_m(t); m \in \mathbb{Z}\}$ be a set of orthogonal functions, each of energy 2 and each orthogonal to each of the $\psi_{k,j}(t)$. Assume that $\{\phi_m(t); m \in \mathbb{Z}\}$, together with the $\psi_{k,j}(t)$, span $\mathcal{L}_2$.

The noise $W(t)$, by assumption, is WGN. It can be represented as

$$W(t) = \sum_k \left(Z_{k,1}\psi_{k,1}(t) + Z_{k,2}\psi_{k,2}(t)\right) + \sum_m W_m\phi_m(t),$$

where $\{Z_{k,m}; k \in \mathbb{Z}, m \in \{1,2\}\}$ is the set of scaled linear functionals of the noise in the $\mathcal{L}_2$ vector space spanned by the $\psi_{k,m}(t)$, and $\{W_m; m \in \mathbb{Z}\}$ is the set of linear functionals of the noise in the orthogonal complement of the space. As will be seen shortly, the joint distribution of the W_m makes no difference in choosing between $a(t)$ and $-a(t)$, so long as the W_m are independent of the $Z_{k,j}$ and the transmitted binary digit. The observed random process at passband is then $Y(t) = X(t) + W(t)$:

$$Y(t) = \sum_k \left[Y_{k,1}\psi_{k,1}(t) + Y_{k,2}\psi_{k,2}(t)\right] + \sum_m W_m\phi_m(t),$$

where

$$Y_{k,1} = \left(\pm b_{k,1} + Z_{k,1}\right); \qquad Y_{k,2} = \left(\pm b_{k,2} + Z_{k,2}\right).$$

First assume that a finite number n of orthonormal functions are used to represent $a(t)$. This is no loss of generality, since the single function $a(t)/\|a(t)\|$ would be sufficient. Suppose also, initially, that only a finite number, say $W_1, \dots, W_\ell$, of the orthogonal noise functionals are observed. Assume also that the noise variables, $Z_{k,j}$ and W_m, are independent and are each[4] $\mathcal{N}(0, N_0/2)$. Then the likelihoods are given by

$$f_{Y|X}(y \mid b) = \frac{1}{(\pi N_0)^n}\exp\left(\sum_{k=1}^{n}\sum_{j=1}^{2}\frac{-(y_{k,j} - b_{k,j})^2}{N_0} + \sum_{m=1}^{\ell}\frac{-w_m^2}{N_0}\right),$$

$$f_{Y|X}(y \mid -b) = \frac{1}{(\pi N_0)^n}\exp\left(\sum_{k=1}^{n}\sum_{j=1}^{2}\frac{-(y_{k,j} + b_{k,j})^2}{N_0} + \sum_{m=1}^{\ell}\frac{-w_m^2}{N_0}\right).$$

The log likelihood ratio is thus given by

$$\text{LLR}(y) = \sum_{k=1}^{n}\sum_{j=1}^{2}\frac{-(y_{k,j} - b_{k,j})^2 + (y_{k,j} + b_{k,j})^2}{N_0}$$

$$= \frac{-\|y - b\|^2 + \|y + b\|^2}{N_0} \tag{8.35}$$

$$= \sum_{k=1}^{n}\sum_{j=1}^{2}\frac{4y_{k,j}b_{k,j}}{N_0} = \frac{4\langle y, b\rangle}{N_0}, \tag{8.36}$$

and the MAP test is

$$\frac{\langle y, b\rangle}{\|b\|} \mathop{\gtrless}_{\tilde{x}=-b}^{\tilde{x}=b} \frac{N_0 \ln \eta}{4\|b\|}.$$

[4] Recall that $N_0/2$ is the noise variance using the same scale as used for the input, and the use of orthogonal functions of energy 2 at passband corresponds to orthonormal functions at baseband. Thus, since the input energy is measured at baseband, the noise is also; at passband, then, both the signal energy and the spectral density of the noise are multiplied by 2.

This is the same as the real vector case analyzed in Section 8.3.3. In fact, the only difference is that the observation here includes noise in the degrees of freedom orthogonal to the range of interest, and the derivation of the LLR shows clearly why these noise variables do not appear in the LLR. In fact, the number ℓ of rvs W_m can be taken to be arbitrarily large, and they can have any joint density. So long as they are independent of the $Z_{k,j}$ (and of $X(t)$), they cancel out in the LLR. In other words, WGN should be viewed as noise that is iid Gaussian over a large enough space to represent the signal, and is independent of the signal and noise elsewhere.

The preceding argument leading to (8.35) and (8.36) is not entirely satisfying mathematically, since it is based on the slightly vague notion of the signal space of interest, but in fact it is just this feature that makes it useful in practice, since physical noise characteristics do change over large changes in time and frequency.

The inner product in (8.36) is the inner product over the $\mathcal{L}_2$ space of real sequences. Since these sequences are coefficients in an orthogonal (rather than orthonormal) expansion, the conversion to an inner product over the corresponding functions (see Exercise 8.5) is given by

$$\sum_{k,j} y_{k,j} b_{k,j} = \frac{1}{2} \int y(t)b(t)\mathrm{d}t. \tag{8.37}$$

This shows that the LLR is independent of the basis, and that this waveform problem reduces to the single-dimensional problem if $b(t)$ is a multiple of one of the basis functions. Also, if a countably infinite basis for the signal space of interest is used, (8.37) is still valid.

Next consider what happens when $Y(t) = \pm b(t) + W(t)$ is demodulated to the baseband waveform $V(t)$. The component $\sum_m W_m(t)$ of $Y(t)$ extends to frequencies outside the passband, and thus $Y(t)$ is filtered before demodulation, preventing an aliasing-like effect between $\sum_m W_m(t)$ and the signal part of $Y(t)$ (see Exercise 6.11). Assuming that this filtering does not affect $b(t)$, $b(t)$ maps back into $a(t) = \sum_k a_k \theta_k(t)$, where $a_k = b_{k,1} + \mathrm{i}b_{k,2}$. Similarly, $W(t)$ maps into

$$Z(t) = \sum_k Z_k \theta_k(t) + Z_\perp(t),$$

where $Z_k = Z_{k,1} + \mathrm{i}Z_{k,2}$ and $Z_\perp(t)$ is the result of filtering and frequency demodulation on $\sum_m W_m \phi_m(t)$. The received baseband complex process is then given by

$$V(t) = \sum_k V_k \theta_k(t) + Z_\perp(t), \qquad \text{where} \quad V_k = \pm a_k + Z_k. \tag{8.38}$$

By the filtering assumption above, the sample functions of $Z_\perp(t)$ are orthogonal to the space spanned by the $\theta_k(t)$, and thus the sequence $\{V_k;\, k \in \mathbb{Z}\}$ is determined from $V(t)$. Since $V_k = Y_{k,1} + \mathrm{i}Y_{k,2}$, the sample value LLR($y$) in (8.36) is determined as follows by the sample values of $\{v_k;\, k \in \mathbb{Z}\}$:

$$\mathrm{LLR}(y) = \frac{4\langle y, b\rangle}{N_0} = \frac{4\Re(\langle v, a\rangle)}{N_0}. \tag{8.39}$$

Thus $\{v_k; k \in \mathbb{Z}\}$ is a sufficient statistic for $y(t)$, and thus the MAP test based on $y(t)$ can be performed using $v(t)$. Now an implementation that first finds the sample function $v(t)$ from $y(t)$ and then does a MAP test on $v(t)$ is simply a particular kind of test on $y(t)$, and thus cannot achieve a smaller error probability than the MAP test on y. Finally, since $\{v_k; k \in \mathbb{Z}\}$ is a sufficient statistic for $y(t)$, it is also a sufficient statistic for $v(t)$ and thus the orthogonal noise $Z_{\perp}(t)$ is irrelevant.

Note that the LLR in (8.39) is the same as the complex vector result in (8.30). One could repeat the argument there, adding in an orthogonal expansion for $Z_{\perp}(t)$ to verify the argument that $Z_{\perp}(t)$ is irrelevant. Since $Z_{\perp}(t)$ could take on virtually any form, the argument above, based on the fact that $Z_{\perp}(t)$ is a function of $\sum_m W_m \phi_m(t)$, which is independent of the signal and noise in the signal space, is more insightful.

To summarize this subsection, the detection of a single bit, sent by generating antipodal signals at baseband and modulating to passband, has been analyzed. After adding WGN, the received waveform is demodulated to baseband and then the single bit is detected. The MAP detector at passband is a threshold test on $\int y(t)b(t)\mathrm{d}t$. This is equivalent to a threshold test at baseband on $\Re(\int v(t)a^*(t)\mathrm{d}t)$. This shows that no loss of optimality occurs by demodulating to baseband and also shows that detection can be done either at passband or at baseband. In the passband case, the result is an immediate extension of binary detection for real vectors, and at baseband it is an immediate extension of binary detection of complex vectors.

The results of this section can now be interpreted in terms of PAM and QAM, while still assuming a "one-shot" system in which only one binary digit is actually sent. Recall that for both PAM and QAM modulation, the modulation pulse $p(t)$ is orthogonal to its T-spaced time shifts if $|\hat{p}(f)|^2$ satisfies the Nyquist criterion. Thus, if the corresponding received baseband waveform is passed through a matched filter (a filter with impulse response $p^*(t)$) and sampled at times kT, the received samples will have no intersymbol interference. For a single bit transmitted at discrete time 0, $u(t) = \pm a(t) = \pm ap(t)$. The output of the matched filter at receiver time 0 is then given by

$$\int v(t)p^*(t)\mathrm{d}t = \frac{\Re[\langle v, a\rangle]}{a},$$

which is a scaled version of the LLR. Thus the receiver from Chapter 6 that avoids intersymbol interference also calculates the LLR, from which a threshold test yields the MAP detection.

Section 8.4 shows that this continues to provide MAP tests on successive signals. It should be noted also that sampling the output of the matched filter at time 0 yields the MAP test whether or not $p(t)$ has been chosen to avoid intersymbol interference.

It is important to note that the performance of binary antipodal communication in WGN depends only on the energy of the transmitted waveform. With ML detection, the error probability is the familiar expression $Q(\sqrt{2E_{\mathrm{b}}/N_0})$, where $E_{\mathrm{b}} = \int |a(t)|^2\,\mathrm{d}t$ and the variance of the noise in each real degree of freedom in the region of interest is $N_0/2$.

This completes the analysis of binary detection in WGN, including the relationship between the vector case and waveform case and that between complex waveforms or vectors at baseband and real waveforms or vectors at passband.

The following sections analyze M-ary detection. The relationships between vector and waveform and between real and complex is the same as above, so the following sections each assume whichever of these cases is most instructive without further discussion of these relationships.

8.4 *M*-ary detection and sequence detection

The analysis in Section 8.3 was limited in several ways. First, only binary signal sets were considered, and second only the "one-shot" problem where a single bit rather than a sequence of bits was considered. In this section, M-ary signal sets for arbitrary M will be considered, and this will then be used to study the transmission of a sequence of signals and to study arbitrary modulation schemes.

8.4.1 *M*-ary detection

Going from binary to M-ary hypothesis testing is a simple extension. To be specific, this will be analyzed for the complex random vector case. Let the observation be a complex random n-vector $\boldsymbol{V}$ and let the complex random n-vector $\boldsymbol{U}$ to be detected take on a value from the set $\{\boldsymbol{a}_0, \dots, \boldsymbol{a}_{M-1}\}$ with a-priori probabilities $p_0, \dots, p_{M-1}$. Denote the a-posteriori probabilities by $p_{\boldsymbol{U}|\boldsymbol{V}}(\boldsymbol{a}_m|\boldsymbol{v})$. The MAP rule (see Section 8.1) then chooses $\tilde{\boldsymbol{U}}(\boldsymbol{v}) = \arg\max_m p_{\boldsymbol{U}|\boldsymbol{V}}(\boldsymbol{a}_m|\boldsymbol{v})$. Assuming that the likelihoods can be represented as probability densities $f_{\boldsymbol{V}|\boldsymbol{U}}$, the MAP rule can be expressed as

$$\tilde{\boldsymbol{U}}(\boldsymbol{v}) = \arg\max_m p_m f_{\boldsymbol{V}|\boldsymbol{U}}(\boldsymbol{v}|\boldsymbol{a}_m).$$

Usually, the simplest approach to this M-ary rule is to consider multiple binary hypothesis testing problems. That is, $\tilde{\boldsymbol{U}}(v)$ is that $\boldsymbol{a}_m$ for which

$$\Lambda_{m,m'}(\boldsymbol{v}) = \frac{f_{\boldsymbol{V}|\boldsymbol{U}}(\boldsymbol{v}|\boldsymbol{a}_m)}{f_{\boldsymbol{V}|\boldsymbol{U}}(\boldsymbol{v}|\boldsymbol{a}_{m'})} \geq \frac{p_{m'}}{p_m}$$

for all m'. In the case of ties, it makes no difference which of the maximizing hypotheses are chosen.

For the complex vector additive WGN case, the observation is $\boldsymbol{V} = \boldsymbol{U} + \boldsymbol{Z}$, where $\boldsymbol{Z}$ is complex Gaussian noise with iid real and imaginary components. As derived in (8.29), the log likelihood ratio (LLR) between each pair of hypotheses $\boldsymbol{a}_m$ and $\boldsymbol{a}_{m'}$ is given by

$$\mathrm{LLR}_{m,m'}(\boldsymbol{v}) = \frac{-\|\boldsymbol{v}-\boldsymbol{a}_m\|^2 + \|\boldsymbol{v}-\boldsymbol{a}_{m'}\|^2}{N_0}. \tag{8.40}$$

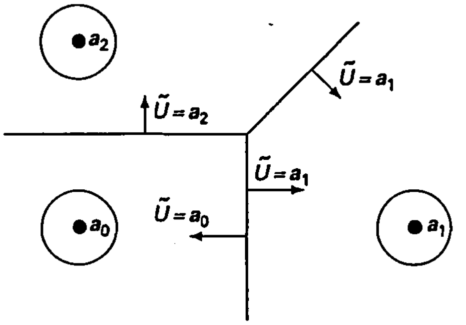


Figure 8.5. Decision regions for a 3-ary alphabet of vector signals in iid Gaussian noise. For ML detection, the decision regions are Voronoi regions, i.e. regions separated by perpendicular bisectors between the signal points.

Thus, each binary test separates the observation space[5] into two regions separated by the perpendicular bisector between the two points. With M hypotheses, the space is separated into the Voronoi regions of points closest to each of the signals (hypotheses) (see Figure 8.5). If the a-priori probabilities are unequal, then these perpendicular bisectors are shifted, remaining perpendicular to the axis joining the two signals, but no longer being bisectors.

The probability that noise carries the observation across one of these perpendicular bisectors is given in (8.31). The only new problem that arises with M-ary hypothesis testing is that the error probability, given $U = m$, is the union of $M-1$ events, namely crossing the corresponding perpendicular to each other point. This can be found exactly by integrating over the n-dimensional vector space, but is usually upperbounded and approximated by the union bound, where the probability of crossing each perpendicular is summed over the $M-1$ incorrect hypotheses. This is usually a good approximation (if M is not too large), because the Gaussian density decreases so rapidly with distance; thus, in the ML case, most errors are made when observations occur roughly halfway between the transmitted and the detected signal points.

8.4.2 Successive transmissions of QAM signals in WGN

This subsection extends the "one-shot" analysis of detection for QAM and PAM in the presence of WGN to the case in which an n-tuple of successive independent symbols are transmitted. We shall find that, under many conditions, both the detection rule and the corresponding probability of symbol error can be analyzed by looking at one symbol at a time.

First consider a QAM modulation system using a modulation pulse $p(t)$. Assume that $p(t)$ has unit energy and is orthonormal to its T-spaced shifts $\{p(t-kT);\ k \in \mathbb{Z}\}$,

[5] For an n-dimensional complex vector space, it is simplest to view the observation space as the corresponding $2n$-dimensional real vector space.

i.e. that $\{p(t-kT);\ k\in\mathbb{Z}\}$ is a set of orthonormal functions. Let $\mathcal{A}=\{a_0,\dots,a_{M-1}\}$ be the alphabet of complex input signals and denote the input waveform over an arbitrary n-tuple of successive input signals by

$$u(t)=\sum_{k=1}^{n}u_k p(t-kT),$$

where each u_k is a selection from the input alphabet $\mathcal{A}$.

Let $\{\phi_k(t);\ k\geq 1\}$ be an orthonormal basis of complex $\mathcal{L}_2$ waveforms such that the first n waveforms in that basis are given by $\phi_k(t)=p(t-kT)$, $1\leq k\leq n$. The received baseband waveform is then given by

$$V(t)=\sum_{k=1}^{\infty}V_k\phi_k(t)=\sum_{k=1}^{n}(u_k+Z_k)p(t-kT)+\sum_{k>n}Z_k\phi_k(t). \tag{8.41}$$

We now compare two different detection schemes. In the first, a single ML decision between the M^n hypotheses for all possible joint values of $U_1,\dots,U_n$ is made based on $V(t)$. In the second scheme, for each k, $1\leq k\leq n$, an ML decision between the M possible hypotheses $a_0,\dots,a_{M-1}$ is made for input U_k based on the observation $V(t)$. Thus in this scheme, n separate M-ary decisions are made, one for each of the n successive inputs.

For the first alternative, each hypothesis corresponds to an n-dimensional vector of inputs, $\boldsymbol{u}=(u_1,\dots,u_n)^{\mathsf{T}}$. As in Section 8.3.5, the sample value $v(t)=\sum_k v_k\phi_k(t)$ of the received waveform can be taken as an ℓ-tuple $\boldsymbol{v}=(v_1,v_2,\dots,v_\ell)^{\mathsf{T}}$ with $\ell\geq n$. The likelihood of $\boldsymbol{v}$ conditional on $\boldsymbol{u}$ is then given by

$$f_{\boldsymbol{V}|\boldsymbol{U}}(\boldsymbol{v}|\boldsymbol{u})=\prod_{k=1}^{n}f_Z(v_k-u_k)\prod_{k=n+1}^{\ell}f_Z(v_k).$$

For any two hypotheses, say $\boldsymbol{u}=(u_1,\dots,u_n)^{\mathsf{T}}$ and $\boldsymbol{u}'=(u_1',\dots,u_n')^{\mathsf{T}}$, the likelihood ratio and LLR are given by

$$\Lambda_{\boldsymbol{u},\boldsymbol{u}'}(\boldsymbol{v})=\prod_{k=1}^{n}\frac{f_Z(v_k-u_k)}{f_Z(v_k-u_k')}, \tag{8.42}$$

$$\mathrm{LLR}_{\boldsymbol{u},\boldsymbol{u}'}(\boldsymbol{v})=\frac{-\|\boldsymbol{v}-\boldsymbol{u}\|^2+\|\boldsymbol{v}-\boldsymbol{u}'\|^2}{N_0}. \tag{8.43}$$

Note that for each $k>n$, v_k does not appear in this likelihood ratio. Thus this likelihood ratio is still valid[6] in the limit $\ell\to\infty$, but the only relevant terms in the decision are $v_1,\dots,v_n$. Therefore, let $\boldsymbol{v}=(v_1,\dots,v_n)^{\mathsf{T}}$ in what follows. From (8.43), this likelihood ratio is positive if and only if $\|\boldsymbol{v}-\boldsymbol{u}\|<\|\boldsymbol{v}-\boldsymbol{u}'\|$. The conclusion is that, for M^n-ary detection, done jointly on $u_1,\dots,u_n$, the ML decision is the vector $\boldsymbol{u}$ that minimizes the distance $\|\boldsymbol{v}-\boldsymbol{u}\|$.

[6] In fact, these final $\ell-n$ components do not have to be independent or equally distributed, they must simply be independent of the signals and noise for $1\leq k\leq n$.

Consider how to minimize $\|\mathbf{v}-\mathbf{u}\|$. Note that

$$\|\mathbf{v}-\mathbf{u}\|^2 = \sum_{k=1}^{n}(v_k - u_k)^2. \tag{8.44}$$

Suppose that $\tilde{\mathbf{u}} = (\tilde{u}_1, \dots, \tilde{u}_n)^{\mathsf{T}}$ minimizes this sum. Then for each k, $\tilde{u}_k$ minimizes $(v_k - u_k)^2$ over the M choices for u_k (otherwise some $a_m \neq \tilde{u}_k$ could be substituted for $\tilde{u}_k$ to reduce $(v_k - u_k)^2$ and therefore reduce the sum in (8.44)). Thus the ML sequence detector with M^n hypotheses detects each U_k by minimizing $(v_k - u_k)^2$ over the M hypotheses for that U_k.

Next consider the second alternative above. For a given sample observation $\mathbf{v} = (v_1, \dots, v_\ell)^{\mathsf{T}}$ and a given k, $1 \le k \le n$, the likelihood of $\mathbf{v}$ conditional on $U_k = u_k$ is given by

$$f_{\mathbf{V}|U_k}(\mathbf{v}|u_k) = f_Z(v_k - u_k) \prod_{j\neq k, 1\le j\le n} f_{V_j}(v_j) \prod_{j=n+1}^{\ell} f_Z(v_j),$$

where $f_{V_j}(v_j) = \sum_m p_m f_{V_j|U_j}(v_j|a_m)$ is the marginal probability of V_j. The likelihood ratio of $\mathbf{v}$ between the hypotheses $U_k = a_m$ and $U_k = a_{m'}$ is then

$$\Lambda^{(k)}_{m,m'}(\mathbf{v}) = \frac{f_Z(v_k - a_m)}{f_Z(v_k - a_{m'})}.$$

This is the familiar 1D nonantipodal Gaussian detection problem, and the ML decision is to choose $\tilde{u}_k$ as the a_m closest to u_k. Thus, given the sample observation $v(t)$, the vector $(\tilde{u}_1, \dots, \tilde{u}_n)^{\mathsf{T}}$ of individual M-ary ML detectors for each U_k is the same as the M^n-ary ML sequence detector for the sequence $\mathbf{U} = (U_1, \dots, U_n)^{\mathsf{T}}$. Moreover, each such detector is equivalent to a vector of ML decisions on each U_k based solely on the observation V_k.

Summarizing, we have proved the following theorem.

Theorem 8.4.1 *Let $U(t) = \sum_{k=1}^{n} U_k p(t-kT)$ be a QAM (or PAM) baseband input to a WGN channel and assume that $\{p(t-kT); 1 \le k \le n\}$ is an orthonormal sequence. Then the M^n-ary ML decision on $\mathbf{U} = (U_1, \dots, U_n)^{\mathsf{T}}$ is equivalent to making separate M-ary ML decisions on each U_k, $1 \le k \le n$, where the decision on each U_k can be based either on the observation $v(t)$ or the observation of v_k.*

Note that the theorem states that the same decision is made for both sequence detection and separate detection for each signal. It *does not* say that the probability of an error within the sequence is the same as the error for a single signal. Letting P be the probability of error for a single signal, the probability of error for the sequence is $1-(1-P)^n$.

Theorem 8.4.1 makes no assumptions about the probabilities of the successive inputs, although the use of ML detection would not minimize the probability of error if the inputs were not independent and equally likely. If coding is used between the n input signals, then not all of these M^n n-tuples are possible. In this case, ML detection on the *possible encoded* sequences (as opposed to all M^n sequences) is different from separate detection on each signal. As an example, if the transmitter always repeats each signal, with $u_1 = u_2$, $u_3 = u_4$, etc., then the detection of u_1 should be based on both v_1 and v_2. Similarly, the detection of u_3 should be based on v_3 and v_4, etc.

When coding is used, it is possible to make ML decisions on each signal separately, and then to use the coding constraints to correct errors in the detected sequence. These individual signal decisions are then called *hard decisions*. It is also possible, for each k, to save a sufficient statistic (such as v_k) for the decision on U_k. This is called a *soft decision* since it saves all the relevant information needed for an ML decision between the set of possible codewords. Since the ML decision between possible encoded sequences minimizes the error probability (assuming equiprobable codewords), soft decisions yield smaller error probabilities than hard decisions.

Theorem 8.4.1 can be extended to MAP detection if the input signals are statistically independent of each other (see Exercise 8.15). One can see this intuitively by drawing the decision boundaries for the 2D real case; these decision boundaries are then horizontal and vertical lines.

A nice way to interpret Theorem 8.4.1 is to observe that the detection of each signal U_k depends only on the corresponding received signal V_k; all other components of the received vector are irrelevant to the decision on U_k. Section 8.4.3 generalizes from QAM to arbitrary modulation schemes and also generalizes this notion of irrelevance.

8.4.3 Detection with arbitrary modulation schemes

The previous sections have concentrated on detection of PAM and QAM systems, using real hypotheses $\mathcal{A} = \{a_0, \dots, a_{M-1}\}$ for PAM and complex hypotheses $\mathcal{A} = \{a_0, \dots, a_{M-1}\}$ for QAM. In each case, a sequence $\{u_k; k \in \mathbb{Z}\}$ of signals from $\mathcal{A}$ is modulated into a baseband waveform $u(t) = \sum_k u_k p(t - kT)$. The PAM waveform is then either transmitted or first modulated to passband. The complex QAM waveform is necessarily modulated to a real passband waveform.

This is now generalized by considering a signal set $\mathcal{A}$ to be an M-ary alphabet, $\{\boldsymbol{a}_0, \dots, \boldsymbol{a}_{M-1}\}$, of real n-tuples. Thus each $\boldsymbol{a}_m$ is an element of $\mathbb{R}^n$. The n components of the mth signal vector are denoted by $\boldsymbol{a}_m = (a_{m,1}, \dots, a_{m,n})^{\mathsf{T}}$. The selected signal vector $\boldsymbol{a}_m$ is then modulated into a signal waveform $b_m(t) = \sum_{k=1}^{n} a_{m,k}\phi_k(t)$, where $\{\phi_1(t), \dots, \phi_n(t)\}$ is a set of n orthonormal waveforms.

The above provides a general scenario for mapping the symbols 0 to $M - 1$ into a set of signal waveforms $b_0(t)$ to $b_{M-1}(t)$. A provision must also be made for transmitting a sequence of such M-ary symbols. If these symbols are to be transmitted at T-spaced intervals, the most straightforward way of accomplishing this is to choose the orthonormal waveforms $\phi_1(t), \dots, \phi_n(t)$ in such a way that $\phi_k(t - \ell T)$ and $\phi_j(t - \ell' T)$ are orthonormal for all j, k, $1 \le j, k \le n$, and all integers ℓ, ℓ'. In this case, a sequence of symbols, say $s_1, s_2, \dots$, each drawn from the alphabet $\{0, \dots, M-1\}$, could be mapped into a sequence of waveforms $b_{s_1}(t), b_{s_2}(t - T), \dots$ The transmitted waveform would then be $\sum_\ell b_{s_\ell}(t - \ell T)$.

Note that PAM is a special case of this scenario where the dimension n is 1. The function $\phi_1(t)$ in this case is the real modulation pulse $p(t)$ for baseband transmission and $\sqrt{2}p(t)\cos(2\pi f_c t)$ for passband transmission; QAM is another special case where n is 2 at passband. In this case, the complex signals a_m are viewed as 2D real signals. The orthonormal waveforms (assuming real $p(t)$) are $\phi_1(t) = \sqrt{2}p(t)\cos(2\pi f_c t)$ and $\phi_2(t) = \sqrt{2}p(t)\sin(2\pi f_c t)$.

More generally, it is not necessary to start at baseband and shift to passband,[7] and it is not necessary for successive signals to be transmitted as time shifts of a basic waveform set. For example, in frequency-hopping systems, successive n-dimensional signals can be modulated to different carrier frequencies. What is important is that the successive transmitted signal waveforms are all orthogonal to each other.

Let $X(t)$ be the first signal waveform in such a sequence of successive waveforms. Then $X(t)$ is a choice from the set of M waveforms, $b_0(t), \dots, b_{M-1}(t)$. We can represent $X(t)$ as $\sum_{k=1}^{n} X_k \phi_k(t)$, where, under hypothesis m, $X_k = a_{m,k}$ for $1 \le k \le n$. Let $\phi_{n+1}(t), \phi_{n+2}(t), \dots$ be an additional set of orthonormal functions such that the entire set $\{\phi_k(t); k \ge 1\}$ spans the space of real $\mathcal{L}_2$ waveforms. The subsequence $\phi_{n+1}(t), \phi_{n+2}(t), \dots$ might include the successive time shifts of $\phi_1(t), \dots, \phi_n(t)$ for the example above, but in general can be arbitrary. We do assume, however, that successive signal waveforms are orthogonal to $\phi_1(t), \dots, \phi_n(t)$, and thus that they can be expanded in terms of $\phi_{n+1}(t), \phi_{n+2}(t), \dots$ The received random waveform $Y(t)$ is assumed to be the sum of $X(t)$, the WGN $Z(t)$, and contributions of signal waveforms other than X. These other waveforms could include successive signals from the given channel input and also signals from other users. This sum can be expanded over an arbitrarily large number, say ℓ, of these orthonormal functions:

$$Y(t) = \sum_{k=1}^{\ell} Y_k \phi_k(t) = \sum_{k=1}^{n} (X_k + Z_k)\phi_k(t) + \sum_{k=n+1}^{\ell} Y_k \phi_k(t). \tag{8.45}$$

Note that in (8.45) the random process $\{Y(t); t \in \mathbb{R}\}$ specifies the random variables $Y_1, \dots, Y_\ell$. Assuming that the sample waveforms of $Y(t)$ are $\mathcal{L}_2$, it also follows that the limit as $\ell \to \infty$ of $Y_1, \dots, Y_\ell$ specifies $Y(t)$ in the $\mathcal{L}_2$ sense. Thus we consider $Y_1, \dots, Y_\ell$ to be the observation at the channel output. It is convenient to separate these terms into two vectors, $Y = (Y_1, \dots, Y_n)^{\mathsf{T}}$ and $Y' = (Y_{n+1}, \dots, Y_\ell)^{\mathsf{T}}$.

Similarly, the WGN $Z(t) = \sum_k Z_k \phi_k(t)$ can be represented by $Z = (Z_1, \dots, Z_n)^{\mathsf{T}}$ and $Z' = (Z_{n+1}, \dots, Z_\ell)^{\mathsf{T}}$, and $X(t)$ can be represented as $X = (X_1, \dots, X_n)^{\mathsf{T}}$. Finally, let $V(t) = \sum_{k>n} V_k \phi_k(t)$ be the contributions from other users and successive signals from the given user. Since these terms are orthogonal to $\phi_1(t), \dots, \phi_n(t)$, $V(t)$ can be represented by $V' = (V_{n+1}, \dots, V_\ell)^{\mathsf{T}}$. With these changes, (8.45) becomes

$$Y = X + Z; \qquad Y' = Z' + V'. \tag{8.46}$$

[7] It seems strange at first that the real vector and real waveform case here is more general than the complex case, but the complex case is used for notational and conceptual simplifications at baseband, where the baseband waveform will be modulated to passsband and converted to a real waveform.

The observation is a sample value of (Y, Y'), and the detector must choose the MAP value of X. Assuming that X, Z, Z', and V' are statistically independent, the likelihoods can be expressed as follows:

$$f_{YY'|X}(yy'|a_m) = f_Z(y-a_m)f_{Y'}(y').$$

The likelihood ratio between hypotheses a_m and $a_{m'}$ is then given by

$$\Lambda_{m,m'}(y) = \frac{f_Z(y-a_m)}{f_Z(y-a_{m'})}. \tag{8.47}$$

The important thing here is that all the likelihood ratios (for $0 \le m, m' \le M-1$) depend only on Y, and thus Y is a sufficient statistic for a MAP decision on X. Note that Y' is irrelevant to the decision, and thus its probability density is irrelevant (other than the need to assume that Y' is statistically independent of (Z, X)). This also shows that the size of ℓ is irrelevant. This is summarized (and slightly generalized by dropping the Gaussian noise assumption) in the following theorem.

Theorem 8.4.2 (Theorem of irrelevance) *Let* $\{\phi_k(t); k \ge 1\}$ *be a set of real orthonormal functions. Let* $X(t) = \sum_{k=1}^{n} X_k\phi_k(t)$ *and* $Z(t) = \sum_{k=1}^{n} Z_k\phi_k(t)$ *be the input to a channel and the corresponding noise, respectively, where* $X = (X_1, \dots, X_n)^{\mathsf{T}}$ *and* $Z = (Z_1, \dots, Z_n)^{\mathsf{T}}$ *are random vectors. Let* $Y'(t) = \sum_{k>n} Y_k\phi_k(t)$*, where, for each* $\ell > n$, $Y' = (Y_{n+1}, \dots, Y_\ell)^{\mathsf{T}}$ *is a random vector that is statistically independent of the pair* X, Z. *Let* $Y = X + Z$. *Then the LLR and the MAP detection of* X *from the observation of* Y, Y' *depends only on* Y. *That is, the observed sample value of* Y' *is irrelevant.*

The orthonormal set $\{\phi_1(t), \dots, \phi_n(t)\}$ chosen above appears to have a more central importance than it really has. What is important is the existence of an n-dimensional subspace of real $\mathcal{L}_2$ that includes the signal set and has the property that the noise and signals orthogonal to this subspace are independent of the noise and signal within the subspace. In the usual case, we choose this subspace to be the space spanned by the signal set, but there are also cases where the subspace must be somewhat larger to provide for the independence between the subspace and its complement.

The irrelevance theorem does not specify how to do MAP detection based on the observed waveform, but rather shows how to reduce the problem to a finite-dimensional problem. Since the likelihood ratios specify both the decision regions and the error probability for MAP detection, it is clear that the choice of orthonormal set cannot influence either the error probability or the mapping of received waveforms to hypotheses.

One important constraint in the above analysis is that both the noise and the interference (from successive transmissions and from other users) are additive. The other important constraint is that the interference is both orthogonal to the signal $X(t)$ and also statistically independent of $X(t)$. The orthogonality is why $Y = X + Z$, with no contribution from the interference. The statistical independence is what makes Y' irrelevant.

If the interference is orthogonal but not independent, then a MAP decision could still be made based on Y alone; the error probability would be the same as if Y, Y' were independent, but using the dependence at the decoder could lead to a smaller error probability.

On the other hand, if the interference is nonorthogonal but independent, then Y would include both noise and a contribution from the interference, and the error probability would typically be larger, but never smaller, than in the orthogonal case. As a rule of thumb, then, nonorthogonal interference tends to increase error probability, whereas dependence (if the receiver makes use of it) tends to reduce error probability.

If successive statistically independent signals, $X_1, X_2, \dots,$ are modulated onto distinct sets of orthonormal waveforms (i.e. if X_1 is modulated onto the orthonormal waveforms $\phi_1(t)$ to $\phi_n(t)$, X_2 is modulated onto $\phi_{n+1}(t)$ to $\phi_{2n}(t)$, etc.), then it also follows, as in Section 8.4.2, that ML detection on a sequence $X_1, \dots, X_\ell$ is equivalent to separate ML decisions on each input signal X_j, $1 \leq j \leq \ell$. The details are omitted since the only new feature in this extension is the more complicated notation.

The higher-dimensional mappings allowed in this subsection are sometimes called *channel codes*, and are sometimes simply viewed as more complex forms of modulation. The coding field is very large, but the following sections provide an introduction.

8.5 Orthogonal signal sets and simple channel coding

An orthogonal signal set is a set $a_0, \dots, a_{M-1}$ of M real orthogonal M-vectors, each with the same energy E. Without loss of generality we choose a basis for $\mathbb{R}^M$ in which the mth basis vector is $a_m/\sqrt{E}$. In this basis, $a_0 = (\sqrt{E}, 0, 0, \dots, 0)^{\mathsf{T}}$, $a_1 = (0, \sqrt{E}, 0, \dots, 0)^{\mathsf{T}}$, etc. Modulation onto an orthonormal set $\{\phi_m(t)\}$ of waveforms then maps hypothesis a_m $(0 \leq m \leq M-1)$ into the waveform $\sqrt{E}\phi_m(t)$. After addition of WGN, the sufficient statistic for detection is a sample value y of $Y = A + Z$, where A takes on the values $a_0, \dots, a_{M-1}$ with equal probability and $Z = (Z_0, \dots, Z_{M-1})^{\mathsf{T}}$ has iid components $\mathcal{N}(0, N_0/2)$. It can be seen that the ML decision is to decide on that m for which y_m is largest.

The major case of interest for orthogonal signals is where M is a power of 2, say $M = 2^b$. Thus the signal set can be used to transmit b binary digits, so the energy per bit is $E_{\mathrm{b}} = E/b$. The number of required degrees of freedom for the signal set, however, is $M = 2^b$, so the spectral efficiency ρ (the number of bits per pair of degrees of freedom) is then $\rho = b/2^{b-1}$. As b gets large, ρ gets small at almost an exponential rate. It will be shown, however, that for large enough E_{b}, as b gets large holding E_{b} constant, the ML error probabiliity goes to 0. In particular, for any $E_{\mathrm{b}}/N_0 < \ln 2 = 0.693$, the error probability goes to 0 exponentially as $b \to \infty$. Recall that $\ln 2 = 0.693$, i.e. -1.59 dB, is the Shannon limit for reliable communication on a WGN channel with unlimited bandwidth. Thus the derivation to follow will establish the Shannon theorem for WGN and unlimited bandwidth. Before doing that, however, two closely related types of signal sets are discussed.

8.5.1 Simplex signal sets

Consider the random vector A with orthogonal equiprobable sample values $a_0, \dots, a_{M-1}$ as described above. The mean value of A is then given by

$$\overline{A} = \left(\frac{\sqrt{E}}{M}, \frac{\sqrt{E}}{M}, \dots, \frac{\sqrt{E}}{M} \right)^{\mathsf{T}}.$$

We have seen that if a signal set is shifted by a constant vector, the Voronoi detection regions are also shifted and the error probability remains the same. However, such a shift can change the expected energy of the random signal vector. In particular, if the signals are shifted to remove the mean, then the signal energy is reduced by the energy (squared norm) of the mean. In this case, the energy of the mean is E/M. A *simplex signal set* is an orthogonal signal set with the mean removed. That is,

$$S = A - \overline{A}; \quad s_m = a_m - \overline{A}; \quad 0 \leq m \leq M-1.$$

In other words, the mth component of s_m is $\sqrt{E}\,(M-1)/M$ and each other component is $-\sqrt{E}/M$. Each simplex signal has energy $E(M-1)/M$, so the simplex set has the same error probability as the related orthogonal set, but requires less energy by a factor of $(M-1)/M$. The simplex set of size M has dimensionality $M-1$, as can be seen from the fact that the sum of all the signals is 0, so the signals are linearly dependent. Figure 8.6 illustrates the orthogonal and simplex sets for $M = 2$ and 3.

For small M, the simplex set is a substantial improvement over the orthogonal set. For example, for $M = 2$, it has a 3 dB energy advantage (it is simply the antipodal 1D set). Also, it uses one fewer dimension than the orthogonal set. For large M, however, the improvement becomes almost negligible.

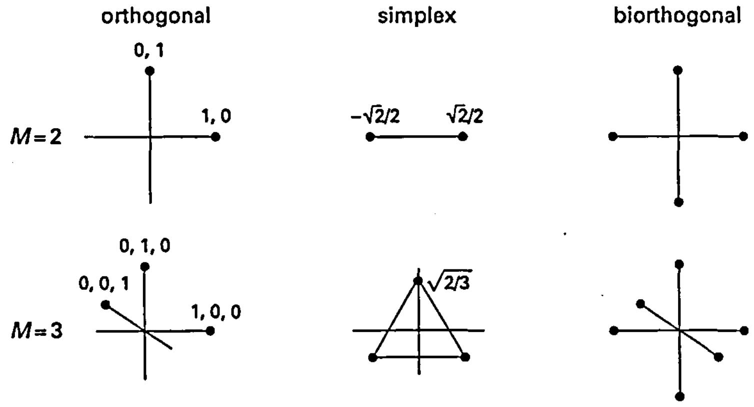


Figure 8.6. Orthogonal, simplex, and biorthogonal signal constellations.

8.5.2 Biorthogonal signal sets

If $a_0, \dots, a_{M-1}$ is a set of orthogonal signals, we call the set of $2M$ signals consisting of $\pm a_0, \dots, \pm a_{M-1}$ a *biorthogonal signal set*; 2D and 3D examples of biorthogonal signals sets are given in Figure 8.6.

It can be seen by the same argument used for orthogonal signal sets that the ML detection rule for such a set is to first choose the dimension m for which $|y_m|$ is largest, and then choose a_m or $-a_m$ depending on whether y_m is positive or negative. Orthogonal signal sets and simplex signal sets each have the property that each signal is equidistant from every other signal. For biorthogonal sets, each signal is equidistant from all but one of the other signals. The exception, for the signal a_m, is the signal $-a_m$.

The biorthogonal signal set of M dimensions contains twice as many signals as the orthogonal set (thus sending one extra bit per signal), but has the same minimum distance between signals. It is hard to imagine[8] a situation where we would prefer an orthogonal signal set to a biorthogonal set, since one extra bit per signal is achieved at essentially no cost. However, for the limiting argument to follow, an orthogonal set is used in order to simplify the notation. As M gets very large, the advantage of biorthogonal signals becomes smaller, which is why, asymptotically, the two are equivalent.

8.5.3 Error probability for orthogonal signal sets

Since the signals differ only by the ordering of the coordinates, the probability of error does not depend on which signal is sent; thus $\Pr(e) = \Pr(e|A = a_0)$. Conditional on $A = a_0$, Y_0 is $\mathcal{N}(\sqrt{E}, N_0/2)$ and Y_m is $\mathcal{N}(0, N_0/2)$ for $1 \leq m \leq M-1$. Note that if $A = a_0$ and $Y_0 = y_0$, then an error is made if $Y_m \geq y_0$ for any m, $1 \leq m \leq M-1$. Thus

$$\Pr(e) = \int_{-\infty}^{\infty} f_{Y_0|A}(y_0 \,|\, a_0) \Pr\left(\bigcup_{m=1}^{M-1} (Y_m \geq y_0 \,|\, A = a_0) \right) \mathrm{d}y_0. \tag{8.48}$$

The rest of the derivation of $\Pr(e)$, and its asymptotic behavior as M gets large, is simplified if we normalize the outputs to $W_m = Y_m\sqrt{2/N_0}$. Then, conditional on signal a_0 being sent, W_0 is $\mathcal{N}(\sqrt{2E/N_0}, 1) = \mathcal{N}(\alpha, 1)$, where α is an abbreviation for $\sqrt{2E/N_0}$. Also, conditional on $A = a_0$, W_m is $\mathcal{N}(0, 1)$ for $1 \leq m \leq M-1$. It follows that

$$\Pr(e) = \int_{-\infty}^{\infty} f_{W_0|A}(w_0 \,|\, a_0) \Pr\left(\bigcup_{m=1}^{M-1} (W_m \geq w_0 \,|\, A = a_0) \right) \mathrm{d}w_0. \tag{8.49}$$

Using the union bound on the union above,

$$\Pr\left(\bigcup_{m=1}^{M-1} (W_m \geq w_0 | A = a_0) \right) \leq (M-1) Q(w_0). \tag{8.50}$$

[8] One possibility is that at passband a phase error of π can turn a_m into $-a_m$. Thus with biorthogonal signals it is necessary to track phase or use differential phase.

The union bound is quite tight when applied to independent quantitities that have small aggregate probability. Thus, this bound will be quite tight when w_0 is large and M is not too large. When w_0 is small, however, the bound becomes loose. For example, for $w_0 = 0$, $Q(w_0) = 1/2$ and the bound in (8.50) is $(M-1)/2$, much larger than the obvious bound of 1 for any probability. Thus, in the analysis to follow, the left side of (8.50) will be upperbounded by 1 for small w_0 and by $(M-1)Q(w_0)$ for large w_0. Since both 1 and $(M-1)Q(w_0)$ are valid upperbounds for all w_0, the dividing point γ between small and large can be chosen arbitrarily. It is chosen in what follows to satisfy

$$\exp(-\gamma^2/2) = 1/M; \qquad \gamma = \sqrt{2\ln M}. \tag{8.51}$$

It might seem more natural in light of (8.50) to replace γ above by the γ_1 that satisfies $(M-1)Q(\gamma_1) = 1$, and that turns out to be the natural choice in the lowerbound to $\Pr(e)$ developed in Exercise 8.10. It is not hard to see, however, that γ/γ_1 goes to 1 as $M \to \infty$, so the difference is not of major importance. Splitting the integral in (8.49) into $w_0 \le \gamma$ and $w_0 > \gamma$, we obtain

$$\Pr(e) \le \int_{-\infty}^{\gamma} f_{W_0|A}(w_0 \mid a_0)\mathrm{d}w_0 + \int_{\gamma}^{\infty} f_{W_0|A}(w_0 \mid a_0)(M-1)Q(w_0)\mathrm{d}w_0 \tag{8.52}$$

$$\le Q(\alpha-\gamma) + \int_{\gamma}^{\infty} f_{W_0|A}(w_0 \mid a_0)(M-1)Q(\gamma)\exp\left(\frac{\gamma^2}{2} - \frac{w_0^2}{2}\right)\mathrm{d}w_0 \tag{8.53}$$

$$\le Q(\alpha-\gamma) + \int_{\gamma}^{\infty} \frac{1}{\sqrt{2\pi}}\exp\left(\frac{-(w_0-\alpha)^2 + \gamma^2 - w_0^2}{2}\right)\mathrm{d}w_0 \tag{8.54}$$

$$= Q(\alpha-\gamma) + \int_{\gamma}^{\infty} \frac{1}{\sqrt{2\pi}}\exp\left(\frac{-2(w_0-\alpha/2)^2 + \gamma^2 - \alpha^2/2}{2}\right)\mathrm{d}w_0 \tag{8.55}$$

$$= Q(\alpha-\gamma) + \frac{1}{\sqrt{2}}Q\left(\sqrt{2}\left(\gamma - \frac{\alpha}{2}\right)\right)\exp\left(\frac{\gamma^2}{2} - \frac{\alpha^2}{4}\right). \tag{8.56}$$

The first term on the right side of (8.52) is the lower tail of the distribution of W_0, and is the probability that the *negative* of the fluctuation of W_0 exceeds $\alpha - \gamma$, i.e. $Q(\alpha - \gamma)$. In the second term, $Q(w_0)$ is upperbounded using Exercise 8.7(c), thus resulting in (8.53). This is simplified by $(M-1)Q(\gamma) \le M\exp(-\gamma^2/2) = 1$, resulting in (8.54). The exponent is then manipulated to "complete the square" in (8.55), leading to an integral of a Gaussian density, as given in (8.56).

The analysis now breaks into three special cases: the first where $\alpha \le \gamma$; the second where $\alpha/2 \le \gamma < \alpha$; and the third where $\gamma \le \alpha/2$. We explain the significance of these cases after completing the bounds.

Case (1) ($\alpha \le \gamma$) The argument of the first Q function in (8.55) is less than or equal to 0, so its value lies between 1/2 and 1. This means that $\Pr(e) \le 1$, which is a useless result. As seen later, this is the case where the rate is greater than or equal to capacity. It is also shown in Exercise 8.10 that the error probability must be large in this case.

Case (2) ($\alpha/2 \le \gamma < \alpha$) Each Q function in (8.56) has a non-negative argument, so the bound $Q(x) \le (1/2)\exp(-x^2/2)$ applies (see Exercise 8.7(b)):

$$\Pr(e) \le \frac{1}{2}\exp\left(\frac{-(\alpha-\gamma)^2}{2}\right) + \frac{1}{2\sqrt{2}}\exp\left(\frac{-\alpha^2}{4} + \frac{\gamma^2}{2} - (\gamma - \alpha/2)^2\right) \quad (8.57)$$

$$\le \left(\frac{1}{2} + \frac{1}{2\sqrt{2}}\right)\exp\left(\frac{-(\alpha-\gamma)^2}{2}\right) \le \exp\left(\frac{-(\alpha-\gamma)^2}{2}\right). \quad (8.58)$$

Note that (8.58) follows (8.57) from combining the terms in the exponent of the second term. The fact that the exponents are equal is not too surprising, since γ was chosen to equalize approximately the integrands in (8.52) at $w_0 = \gamma$.

Case (3) ($\gamma \le \alpha/2$) The argument of the second Q function in (8.56) is less than or equal to 0, so its value lies between 1/2 and 1 and is upperbounded by 1, yielding

$$\Pr(e) \le \frac{1}{2}\exp\left(\frac{-(\alpha-\gamma)^2}{2}\right) + \frac{1}{2\sqrt{2}}\exp\left(\frac{-\alpha^2}{4} + \frac{\gamma^2}{2}\right) \quad (8.59)$$

$$\le \exp\left(\frac{-\alpha^2}{4} + \frac{\gamma^2}{2}\right). \quad (8.60)$$

Since the two exponents in (8.57) are equal, the first exponent in (8.59) must be smaller than the second, leading to (8.60). This is essentially the union bound derived in Exercise 8.8.

The lowerbound in Exercise 8.10 shows that these bounds are quite tight, but the sense in which they are tight will be explained later.

We now explore what α and γ are in terms of the number of codewords M and the energy per bit E_b. Recall that $\alpha = \sqrt{2E/N_0}$. Also $\log_2 M = b$, where b is the number of bits per signal. Thus $\alpha = \sqrt{2bE_b/N_0}$. From (8.51), $\gamma^2 = 2\ln M = 2b\ln 2$. Thus,

$$\alpha - \gamma = \sqrt{2b}\left[\sqrt{E_b/N_0} - \sqrt{\ln 2}\right].$$

Substituting these values into (8.58) and (8.60), we obtain

$$\Pr(e) \le \exp\left[-b\left(\sqrt{E_b/N_0} - \sqrt{\ln 2}\right)^2\right] \qquad \text{for } \frac{E_b}{4N_0} \le \ln 2 < \frac{E_b}{N_0}; \quad (8.61)$$

$$\Pr(e) \le \exp\left[-b\left(\frac{E_b}{2N_0} - \ln 2\right)\right] \qquad \text{for } \ln 2 < \frac{E_b}{4N_0}. \quad (8.62)$$

We see from this that for fixed $E_b/N_0 > \ln 2$, $\Pr(e) \to 0$ as $b \to \infty$.

Recall that in (7.82) we stated that the capacity (in bits per second) of a WGN channel of bandwidth W, noise spectral density $N_0/2$, and power P is given by

$$C = W\log\left(1 + \frac{P}{WN_0}\right). \quad (8.63)$$

With no bandwidth constraint, i.e. in the limit W → ∞, the ultimate capacity is $C = P/(N_0 \ln 2)$. This means that, according to Shannon's theorem, for any rate $R < C = P/(N_0 \ln 2)$, there are codes of rate R bits per second for which the error probability is arbitrarily close to 0. Now $P/R = E_b$, so Shannon says that if $E_b/(N_0 \ln 2) > 1$, then codes exist with arbitrarily small probability of error.

The orthogonal codes provide a concrete proof of this ultimate capacity result, since (8.61) shows that $\Pr(e)$ can be made arbitrarily small (by increasing b) so long as $E_b/(N_0 \ln 2) > 1$. Shannon's theorem also says that the error probability cannot be made small if $E_b/(N_0 \ln 2) < 1$. We have not quite proven that here, although Exercise 8.10 shows that the error probability cannot be made arbitrarily small for an orthogonal code[9] if $E_b/(N_0 \ln 2) < 1$.

The limiting operation here is slightly unconventional. As b increases, E_b is held constant. This means that the energy E in the signal increases linearly with b, but the size of the constellation increases exponentially with b. Thus the bandwidth required for this scheme is infinite in the limit, and going to infinity very rapidly. This means that this is not a practical scheme for approaching capacity, although sets of 64 or even 256 biorthogonal waveforms are used in practice.

The point of the analysis, then, is first to show that this infinite bandwidth capacity can be approached, but second to show also that using large but finite sets of orthogonal (or biorthogonal or simplex) waveforms does decrease error probability for fixed signal-to-noise ratio, and decreases it as much as desired (for rates below capacity) if enough bandwidth is used.

The different forms of solution in (8.61) and (8.62) are interesting, and not simply consequences of the upperbounds used. For case (2), which leads to (8.61), the typical way that errors occur is when $w_0 \approx \gamma$. In this situation, the union bound is on the order of 1, which indicates that, conditional on $y_0 \approx \gamma$, it is quite likely that an error will occur. In other words, the typical error event involves an unusually large negative value for w_0 rather than any unusual values for the other noise terms. In case (3), which leads to (8.62), the typical way for errors to occur is when $w_0 \approx \alpha/2$ and when some other noise term is also at about $\alpha/2$. In this case, an unusual event is needed both in the signal direction and in some other direction.

A more intuitive way to look at this distinction is to visualize what happens when E/N_0 is held fixed and M is varied. Case 3 corresponds to small M, case 2 to larger M, and case 1 to very large M. For small M, one can visualize the Voronoi region around the transmitted signal point. Errors occur when the noise carries the signal point outside the Voronoi region, and that is most likely to occur at the points in the Voronoi surface closest to the transmitted signal, i.e. at points halfway between the transmitted point and some other signal point. As M increases, the number of these

[9] Since a simplex code has the same error probability as the corresponding orthogonal code, but differs in energy from the orthogonal code by a vanishingly small amount as $M \to \infty$, the error probability for simplex codes also cannot be made arbitrarily small for any given $E_b/(N_0 \ln 2) < 1$. It is widely believed, but never proven, that simplex codes are optimal in terms of ML error probability whenever the error probability is small. There is a known example (Steiner, 1994), for all $M \geq 7$, where the simplex is nonoptimal, but in this example the signal-to-noise ratio is very small and the error probability is very large.

midway points increases until one of them is almost certain to cause an error when the noise in the signal direction becomes too large.

8.6 Block coding

This section provides a brief introduction to the subject of coding for error correction on noisy channels. Coding is a major topic in modern digital communication, certainly far more important than suggested by the length of this introduction. In fact, coding is a topic that deserves its own text and its own academic subject in any serious communication curriculum. Suggested texts are Forney (2005) and Lin and Costello (2004). Our purpose here is to provide enough background and examples to understand the role of coding in digital communication, rather than to prepare the student for coding research. We start by viewing orthogonal codes as block codes using a binary alphabet. This is followed by the Reed–Muller codes, which provide considerable insight into coding for the WGN channel. This then leads into Shannon's celebrated noisy-channel coding theorem.

A *block code* is a code for which the incoming sequence of binary digits is segmented into blocks of some given length m and then these binary m-tuples are mapped into *codewords*. There are thus 2^m codewords in the code; these codewords might be binary n-tuples of some *block length* $n > m$, or they might be vectors of signals, or waveforms. Successive codewords then pass through the remaining stages of modulation before transmission. There is no fundamental difference between coding and modulation; for example, the orthogonal code above with $M = 2^m$ codewords can be viewed either as modulation with a large signal set or coding using binary m-tuples as input.

8.6.1 Binary orthogonal codes and Hadamard matrices

When orthogonal codewords are used on a WGN channel, any orthogonal set is equally good from the standpoint of error probability. One possibility, for example, is the use of orthogonal sine waves. From an implementation standpoint, however, there are simpler choices than orthogonal sine waves. Conceptually, also, it is helpful to see that orthogonal codewords can be constructed from binary codewords. This digital approach will turn out to be conceptually important in the study of fading channels and diversity in Chapter 9. It also helps in implementation, since it postpones the point at which digital hardware gives way to analog waveforms.

One digital approach to generating a large set of orthogonal waveforms comes from first generating a set of M binary codewords, each of length M and each distinct pair differing in exactly $M/2$ places. Each binary digit can then be mapped into an antipodal signal, $0 \to +a$ and $1 \to -a$. This yields an M-tuple of real-valued antipodal signals, $s_1, \ldots, s_M$, which is then mapped into the waveform $\sum_j s_j \phi_j(t)$, where $\{\phi_j(t); 1 \leq j \leq M\}$ is an orthonormal set (such as sinc functions or Nyquist pulses). Since each pair of binary codewords differs in $M/2$ places, the corresponding pair of waveforms are orthogonal and each waveform has equal energy. A binary code with the above properties is called a *binary orthogonal code*.

There are many ways to generate binary orthogonal codes. Probably the simplest is from a *Hadamard matrix*. For each integer $m \geq 1$, there is a 2^m by 2^m Hadamard matrix H_m. Each distinct pair of rows in the Hadamard matrix H_m differs in exactly 2^{m-1} places, so the 2^m rows of H_m constitute a binary orthogonal code with 2^m codewords.

It turns out that there is a simple algorithm for generating the Hadamard matrices. The Hadamard matrix H_1 is defined to have the rows 00 and 01, which trivially satisfy the condition that each pair of distinct rows differ in half the positions. For any integer $m > 1$, the Hadamard matrix H_{m+1} of order 2^{m+1} can be expressed as four 2^m by 2^m submatrices. Each of the upper two submatrices is H_m, and the lower two submatrices are H_m and $\overline{H}_m$, where $\overline{H}_m$ is the complement of H_m. This is illustrated in Figure 8.7.

Note that each row of each matrix in Figure 8.7, other than the all-zero row, contains half 0s and half 1s. To see that this remains true for all larger values of m, we can use induction. Thus assume, for given m, that H_m contains a single row of all 0s and $2^m - 1$ rows, each with exactly half 1s. To prove the same for H_{m+1}, first consider the first 2^m rows of H_{m+1}. Each row has twice the length and twice the number of 1s as the corresponding row in H_m. Next consider the final 2^m rows. Note that $\overline{H}_m$ has all 1s in the first row and 2^{m-1} 1s in each other row. Thus the first row in the second set of 2^m rows of H_{m+1} has no 1s in the first 2^m positions and 2^m 1s in the final 2^m positions, yielding 2^m 1s in 2^{m+1} positions. Each remaining row has 2^{m-1} 1s in the first 2^m positions and 2^{m-1} 1s in the final 2^m positions, totaling 2^m 1s as required.

By a similar inductive argument (see Exercise 8.18), the mod-2 sum[10] of any two rows of H_m is another row of H_m. Since the mod-2 sum of two rows gives the positions in which the rows differ, and only the mod-2 sum of a codeword with itself gives the all-zero codeword, this means that the set of rows is a binary orthogonal set.

The fact that the mod-2 sum of any two rows is another row makes the corresponding code a special kind of binary code called a *linear code*, *parity-check code*, or *group code* (these are all synonyms). Binary M-tuples can be regarded as vectors in a vector space over the binary scalar field. It is not necessary here to be precise about what a field is; so far it has been sufficient to consider vector spaces defined over the real or complex fields. However, the binary numbers, using mod-2 addition and ordinary

$$\begin{array}{|c|c|} \hline 0 & 0 \\ \hline 0 & 1 \\ \hline \end{array} \qquad \begin{array}{|c|c||c|c|} \hline 0 & 0 & 0 & 0 \\ \hline 0 & 1 & 0 & 1 \\ \hline\hline 0 & 0 & 1 & 1 \\ \hline 0 & 1 & 1 & 0 \\ \hline \end{array} \qquad \begin{array}{|c|} \hline 0000\ 0000 \\ 0101\ 0101 \\ 0011\ 0011 \\ 0110\ 0110 \\ 0000\ 1111 \\ 0101\ 1010 \\ 0011\ 1100 \\ 0110\ 1001 \\ \hline \end{array}$$

$m=1$ $\qquad$ $m=2$ $\qquad$ $m=3$

Figure 8.7. Hadamard matrices.

[10] The mod-2 sum of two binary numbers is defined by $0 \oplus 0 = 0$, $0 \oplus 1 = 1$, $1 \oplus 0 = 1$, and $1 \oplus 1 = 0$. The mod-2 sum of two rows (or vectors) or binary numbers is the component-wise row (or vector) of mod-2 sums.

multiplication, also form the field called $\mathbb{F}_2$, and the familiar properties of vector spaces, using $\{0, 1\}$ as scalars, apply here also.

Since the set of codewords in a linear code is closed under mod-2 sums (and also closed under scalar multiplication by 1 or 0), a linear code is a binary vector subspace of the binary vector space of binary M-tuples. An important property of such a subspace, and thus of a linear code, is that the set of positions in which two codewords differ is the set of positions in which the mod-2 sum of those codewords contains the binary digit 1. This means that the distance between two codewords (i.e. the number of positions in which they differ) is equal to the weight (the number of positions containing the binary digit 1) of their mod-2 sum. This means, in turn, that, for a linear code, the minimum distance $d_{\min}$ taken between all distinct pairs of codewords is equal to the minimum weight (minimum number of 1s) of any nonzero codeword.

Another important property of a linear code (other than the trivial code consisting of all binary M-tuples) is that some components x_k of each codeword $x = (x_1, \dots, x_M)^{\mathsf{T}}$ can be represented as mod-2 sums of other components. For example, in the $m = 3$ case of Figure 8.7, $x_4 = x_2 \oplus x_3$, $x_6 = x_2 \oplus x_5$, $x_7 = x_3 \oplus x_5$, $x_8 = x_4 \oplus x_5$, and $x_1 = 0$. Thus only three of the components can be independently chosen, leading to a 3D binary subspace. Since each component is binary, such a 3D subspace contains $2^3 = 8$ vectors. The components that are mod-2 combinations of previous components are called "parity checks" and often play an important role in decoding. The first component, x_1, can be viewed as a parity check since it cannot be chosen independently, but its only role in the code is to help achieve the orthogonality property. It is irrelevant in decoding.

It is easy to modify a binary orthogonal code generated by a Hadamard matrix to generate a binary simplex code, i.e. a binary code which, after the mapping $0 \to a, 1 \to -a$, forms a simplex in Euclidean space. The first component of each binary codeword is dropped, turning the code into M codewords over $M - 1$ dimensions. Note that in terms of the antipodal signals generated by the binary digits, dropping the first component converts the signal $+a$ (corresponding to the first binary component 0) into the signal 0 (which corresponds neither to the binary 0 or 1). The generation of the binary biorthogonal code is equally simple; the rows of H_m yield half of the codewords and the rows of $\overline{H}_m$ yield the other half. Both the simplex and the biorthogonal code, as expressed in binary form here, are linear binary block codes.

Two things have been accomplished with this representation of orthogonal codes. First, orthogonal codes can be generated from a binary sequence mapped into an antipodal sequence; second, an example has been given where modulation over a large alphabet can be viewed as a binary block code followed by modulation over a binary or very small alphabet.

8.6.2 Reed–Muller codes

Orthogonal codes (and the corresponding simplex and biorthgonal codes) use enormous bandwidth for large M. The Reed–Muller codes constitute a class of binary linear block codes that include large bandwidth codes (in fact, they include the binary biorthogonal codes), but also allow for much smaller bandwidth expansion, i.e. they allow for binary codes with M codewords, where $\log M$ is much closer to the number of dimensions used by the code.

The Reed–Muller codes are specified by two integer parameters, $m \geq 1$ and $0 \leq r \leq m$; a binary linear block code, denoted by RM(r, m), exists for each such choice. The parameter m specifies the block length to be $n = 2^m$. The minimum distance $d_{\min}(r, m)$ of the code and the number of binary information digits $k(r, m)$ required to specify a codeword are given by

$$d_{\min}(r,m) = 2^{m-r}; \qquad k(r,m) = \sum_{j=0}^{r} \binom{m}{j}, \tag{8.64}$$

where $\binom{m}{j} = \frac{m!}{j!(m-j)!}$. Thus these codes, like the binary orthogonal codes, exist only for block lengths equal to a power of 2. While there is only one binary orthogonal code (as defined through H_m) for each m, there is a range of RM codes for each m, ranging from large $d_{\min}$ and small k to small $d_{\min}$ and large k as r increases.

For each m, these codes are trivial for $r = 0$ and $r = m$. For $r = 0$ the code consists of two codewords selected by a single bit, so $k(0, m) = 1$; one codeword is all 0s and the other is all 1s, leading to $d_{\min}(0, m) = 2^m$. For $r = m$, the code is the set of all binary 2^m-tuples, leading to $d_{\min}(m, m) = 1$ and $k(m, m) = 2^m$. For $m = 1$, then, there are two RM codes: RM$(0, 1)$ consists of the two codewords (0,0) and (1,1), and RM$(1, 1)$ consists of the four codewords (0,0), (0,1), (1,0), and (1,1).

For $m > 1$ and intermediate values of r, there is a simple algorithm, much like that for Hadamard matrices, that specifies the set of codewords. The algorithm is recursive, and, for each $m > 1$ and $0 < r < m$, specifies RM(r, m) in terms of RM$(r, m-1)$ and RM$(r-1, m-1)$. Specifically, $x \in$ RM(r, m) if x is the concatenation of u and $u \oplus v$, denoted by $x = (u, u \oplus v)$, for some $u \in RM(r, m-1,)$ and $v \in RM(r-1, m-1)$. More formally, for $0 < r < m$,

$$\text{RM}(r,m) = \{(u, u \oplus v) \mid u \in \text{RM}(r, m-1), v \in \text{RM}(r-1, m-1)\}. \tag{8.65}$$

The analogy with Hadamard matrices is that x is a row of H_m if u is a row of H_{m-1} and v is either all 1s or all 0s.

The first thing to observe about this definition is that if RM$(r, m - 1)$ and RM$(r - 1, m - 1)$ are linear codes, then RM(r, m) is also. To see this, let $x = (u, u \oplus v)$ and $x' = (u', u' \oplus v')$. Then

$$x \oplus x' = (u \oplus u', u \oplus u' \oplus v \oplus v') = (u'', u'' \oplus v''),$$

where $u'' = u \oplus u' \in$ RM$(r, m - 1)$ and $v'' = v \oplus v' \in$ RM$(r - 1, m - 1)$. This shows that $x \oplus x' \in$ RM(r, m), and it follows that RM(r, m) is a linear code if RM$(r, m - 1)$ and RM$(r - 1, m - 1)$ are. Since both RM$(0, m)$ and RM(m, m) are linear for all $m \geq 1$, it follows by induction on m that all the Reed–Muller codes are linear.

Another observation is that different choices of the pair u and v cannot lead to the same value of $x = (u, u \oplus v)$. To see this, let $x' = (u', v')$. Then, if $u \neq u'$, it follows that the first half of x differs from that of x'. Similarly, if $u = u'$ and $v \neq v'$, then the second half of x differs from that of x'. Thus, $x = x'$ only if both $u = u'$ and $v = v'$. As a consequence of this, the number of information bits required to specify a codeword in RM(r, m), denoted by $k(r, m)$, is equal to the number required to specify a codeword in RM$(r, m-1)$ plus that to specify a codeword in RM$(r-1, m-1)$, i.e., for $0 < r < m$,

$$k(r, m) = k(r, m-1) + k(r-1, m-1).$$

Exercise 8.19 shows that this relationship implies the explicit form for $k(r, m)$ given in (8.64). Finally, Exercise 8.20 verifies the explicit form for $d_{\min}(r, m)$ in (8.64).

The RM$(1, m)$ codes are the binary biorthogonal codes, and one can view the construction in (8.65) as being equivalent to the Hadamard matrix algorithm by replacing the M by M matrix H_m in the Hadamard algorithm by the $2M$ by M matrix $\begin{bmatrix} H_m \\ G_m \end{bmatrix}$, where $G_m = \overline{H}_m$.

Another interesting case is the RM$(m-2, m)$ codes. These have $d_{\min}(m-2, m) = 4$ and $k(m-2, m) = 2^m - m - 1$ information bits. In other words, they have $m+1$ parity checks. As explained below, these codes are called *extended Hamming codes.*

A property of all RM codes is that all codewords have an even number[11] of 1s and thus the last component in each codeword can be viewed as an overall parity check which is chosen to ensure that the codeword contains an even number of 1s. If this final parity check is omitted from RM$(m-2, m)$ for any given m, the resulting code is still linear and must have a minimum distance of at least 3, since only one component has been omitted. This code is called the Hamming code of block length $2^m - 1$ with m parity checks. It has the remarkable property that every binary $(2^m - 1)$-tuple is either a codeword in this code or distance 1 from a codeword.[12]

The Hamming codes are not particularly useful in practice for the following reasons. If one uses a Hamming code at the input to a modulator and then makes hard decisions on the individual bits before decoding, then a block decoding error is made whenever two or more bit errors occur. This is a small improvement in reliability at a very substantial cost in transmission rate. On the other hand, if soft decisions are made, the use of the extended Hamming code (i.e. RM$(m-2, m)$) extends $d_{\min}$ from 3 to 4, significantly decreasing the error probability with a marginal cost in added redundant bits.

8.7 Noisy-channel coding theorem

Sections 8.5 and 8.6 provided a brief introduction to coding. Several examples were discussed showing that the use of binary codes could accomplish the same thing, for

[11] This property can be easily verified by induction.

[12] To see this, note that there are 2^{2^m-1-m} codewords, and each codeword has $2^m - 1$ neighbors; these are distinct from the neighbors of other codewords since $d_{\min}$ is at least 3. Adding the codewords and the neighbors, we get the entire set of 2^{2^m-1} vectors. This also shows that the minimum distance is exactly 3.

example as the use of large sets of orthogonal, simplex, or biorthogonal waveforms. There was an ad hoc nature to the development, however, illustrating a number of schemes with various interesting properties, but little in the way of general results.

The earlier results on $\Pr(e)$ for orthogonal codes were more fundamental, showing that $\Pr(e)$ could be made arbitrarily small for a WGN channel with no bandwidth constraint if E_b/N_0 is greater than ln 2. This constituted a special case of the noisy-channel coding theorem, saying that arbitrarily small $\Pr(e)$ can be achieved for that very special channel and set of constraints.

8.7.1 Discrete memoryless channels

This section states and proves the noisy-channel coding theorem for another special case, that of discrete memoryless channels (DMCs). This may seem a little peculiar after all the emphasis in this and the preceding chapter on WGN. There are two major reasons for this choice. The first is that the argument is particularly clear in the DMC case, particularly after studying the AEP for discrete memoryless sources. The second is that the argument can be generalized easily, as will be discussed later. A DMC has a discrete input sequence $X = X_1, \dots, X_k, \dots$ At each discrete time k, the input to the channel belongs to a finite alphabet $\mathcal{X}$ of symbols. For example, in Section 8.6, the input alphabet could be viewed as the signals $\pm a$. The question of interest would then be whether it is possible to communicate reliably over a channel when the decision to use the alphabet $\mathcal{X} = \{a, -a\}$ has already been made. The channel would then be regarded as the part of the channel from signal selection to an output sequence from which detection would be done. In a more general case, the signal set could be an arbitrary QAM set.

A DMC is also defined to have a discrete output sequence $Y = Y_1, \dots, Y_k, \dots,$ where each output Y_k in the output sequence is a selection from a finite alphabet $\mathcal{Y}$ and is a probabilistic function of the input and noise in a way to be described shortly. In the example above, the output alphabet could be chosen as $\mathcal{Y} = \{a, -a\}$, corresponding to the case in which hard decisions are made on each signal at the receiver. The channel would then include the modulation and detection as an internal part, and the question of interest would be whether coding at the input and decoding from the single-letter hard decisions at the output could yield reliable communication.

Another choice would be to use the pre-decision outputs, first quantized to satisfy the finite alphabet constraint. Another almost identical choice would be a detector that produced a quantized LLR as opposed to a decision.

In summary, the choice of discrete memoryless channel alphabets depends on what part of the overall communication problem is being addressed.

In general, a channel is described not only by the input and output alphabets, but also by the probabilistic description of the outputs conditional on the inputs (the probabilistic description of the inputs is selected by the channel user). Let $X^n = (X_1, X_2, \dots X_n)^{\mathsf{T}}$ be the channel input, here viewed either over the lifetime of the channel or any time greater than or equal to the duration of interest. Similarly, the output is denoted by

$Y^n = (Y_1, \dots, Y_n)^\mathsf{T}$. For a DMC, the probability of the output n-tuple, conditional on the input n-tuple, is defined to satisfy

$$p_{Y^n|X^n}(y_1, \dots, y_n \mid x_1, \dots, x_n) = \prod_{k=1}^{n} p_{Y_k|X_k}(y_k|x_k), \tag{8.66}$$

where $p_{Y_k|X_k}(y_k = j|x_k = i)$, for each $j \in \mathcal{Y}$ and $i \in \mathcal{X}$, is a function only of i and j and not of the time k. Thus, conditional on the inputs, the outputs are independent and have the same conditional distribution at all times. This conditional distribution is denoted by $P_{i,j}$ for all $i \in \mathcal{X}$ and $j \in \mathcal{Y}$, i.e. $p_{Y_k|X_k}(y_k = j|x_k = i) = P_{i,j}$. Thus the channel is completely described by the input alphabet, the output alphabet, and the conditional distribution function $P_{i,j}$. The conditional distribution function is usually called the *transition function* or *transition matrix.*

The most intensely studied DMC over the past 60 years has been the binary symmetric channel (BSC), which has $\mathcal{X} = \{0, 1\}$, $\mathcal{Y} = \{0, 1\}$, and satisfies $P_{0,1} = P_{1,0}$. The single number $P_{0,1}$ thus specifies the BSC. The WGN channel with antipodal inputs and ML hard decisions at the output is an example of the BSC. Despite the intense study of the BSC and its inherent simplicity, the question of optimal codes of long block length (optimal in the sense of minimum error probability) is largely unanswered. Thus, the noisy-channel coding theorem, which describes various properties of the achievable error probability through coding plays a particularly important role in coding.

8.7.2 Capacity

This section defines the capacity C of a DMC. Section 8.7.3, after defining the rate R at which information enters the modulator, shows that reliable communication is impossible on a channel if $R > C$. This is known as the converse to the noisy-channel coding theorem, and is in contrast to Section 8.7.4, which shows that arbitrarily reliable communication is possible for any $R < C$. As in the analysis of orthogonal codes, communication at rates below capacity can be made increasingly reliable with increasing block length, while this is not possible for $R > C$.

The capacity is defined in terms of various entropies. For a given DMC and given sequence length n, let $p_{Y^n|X^n}(y^n|x^n)$ be given by (8.66) and let $p_{X^n}(x^n)$ denote an arbitrary probability mass function chosen by the user on the input $X_1, \dots, X_n$. This leads to a joint entropy $\mathsf{H}[X^nY^n]$. From (2.37), this can be broken up as follows:

$$\mathsf{H}[X^nY^n] = \mathsf{H}[X^n] + \mathsf{H}[Y^n|X^n], \tag{8.67}$$

where $\mathsf{H}[Y^n|X^n] = \mathsf{E}[-\log p_{Y^n|X^n}(Y^n|X^n)]$. Note that because $\mathsf{H}[Y^n|X^n]$ is defined as an expectation over both X^n and Y^n, $\mathsf{H}[Y^n|X^n]$ depends on the distribution of X^n as well as the conditional distribution of Y^n given X^n. The joint entropy $\mathsf{H}[X^nY^n]$ can also be broken up the opposite way as follows:

$$\mathsf{H}[X^nY^n] = \mathsf{H}[Y^n] + \mathsf{H}[X^n|Y^n]. \tag{8.68}$$

Combining (8.67) and (8.68), it is seen that $\mathsf{H}[X^n]-\mathsf{H}[X^n|Y^n]=\mathsf{H}[Y^n]-\mathsf{H}[Y^n|X^n]$. This difference of entropies is called the *mutual information* between X^n and Y^n and is denoted by $\mathsf{I}[X^n;Y^n]$, so

$$\mathsf{I}[X^n;Y^n]=\mathsf{H}[X^n]-\mathsf{H}[X^n|Y^n]=\mathsf{H}[Y^n]-\mathsf{H}[Y^n|X^n]. \tag{8.69}$$

The first expression for $\mathsf{I}[X^n;Y^n]$ has a nice intuitive interpretation. From source coding, $\mathsf{H}[X^n]$ represents the number of bits required to represent the channel input. If we look at a particular sample value y^n of the output, $\mathsf{H}[X^n|Y^n=y^n]$ can be interpreted as the number of bits required to represent X^n after observing the output sample value y^n. Note that $\mathsf{H}[X^n|Y^n]$ is the expected value of this over Y^n. Thus $\mathsf{I}[X^n;Y^n]$ can be interpreted as the reduction in uncertainty, or number of required bits for specification, after passing through the channel. This intuition will lead to the converse to the noisy-channel coding theorem in Section 8.7.3.

The second expression for $\mathsf{I}[X^n;Y^n]$ is the one most easily manipulated. Taking the log of the expression in (8.66), we obtain

$$\mathsf{H}[Y^n|X^n]=\sum_{k=1}^{n}\mathsf{H}[Y_k|X_k]. \tag{8.70}$$

Since the entropy of a sequence of random symbols is upperbounded by the sum of the corresponding terms (see Exercise 2.19),

$$\mathsf{H}[Y^n]\leq\sum_{k=1}^{n}\mathsf{H}[Y_k]. \tag{8.71}$$

Substituting this and (8.70) in (8.69) yields

$$\mathsf{I}[X^n;Y^n]\leq\sum_{k=1}^{n}\mathsf{I}[X_k;Y_k]. \tag{8.72}$$

If the inputs are independent, then the outputs are also, and (8.71) and (8.72) are satisfied with equality. The mutual information $\mathsf{I}[X_k;Y_k]$ at each time k is a function only of the pmf for X_k, since the output probabilities conditional on the input are determined by the channel. Thus, each mutual information term in (8.72) is upperbounded by the maximum of the mutual information over the input distribution. This maximum is defined as the *capacity* of the DMC, given by

$$C=\max_{p}\sum_{i\in\mathcal{X}}\sum_{j\in\mathcal{Y}}p_iP_{i,j}\log\frac{P_{i,j}}{\sum_{\ell\in\mathcal{X}}p_\ell P_{\ell,j}}, \tag{8.73}$$

where $p=\{p_0,p_1,\ldots,p_{|\mathcal{X}|-1}\}$ is the set (over the alphabet $\mathcal{X}$) of input probabilities. The maximum is over this set of input probabilities, subject to $p_i\geq0$ for each $i\in\mathcal{X}$ and $\sum_{i\in\mathcal{X}}p_i=1$. The above function is concave in p, and thus the maximization is straightforward. For the BSC, for example, the maximum is at $p_0=p_1=1/2$ and

$C = 1 + P_{0,1}\log P_{0,1} + P_{0,0}\log P_{0,0}$. Since C upperbounds $\mathsf{I}[X_k; Y_k]$ for each k, with equality if the distribution for X_k is the maximizing distribution,

$$\mathsf{I}[X^n; Y^n] \le nC \tag{8.74}$$

with equality if all inputs are independent and chosen with the maximizing probabilities in (8.73).

8.7.3 Converse to the noisy-channel coding theorem

Define the rate R for the DMC above as the number of iid equiprobable binary source digits that enter the channel per channel use. More specifically, assume that nR bits enter the source and are transmitted over the n channel uses under discussion. Assume also that these bits are mapped into the channel input X^n in a one-to-one way. Thus $\mathsf{H}[X^n] = nR$ and X^n can take on $M = 2^{nR}$ equiprobable values. The following theorem now bounds $\Pr(e)$ away from 0 if $R > C$.

Theorem 8.7.1 *Consider a DMC with capacity C. Assume that the rate R satisfies R > C. Then, for any block length n, the ML probability of error, i.e. the probability that the decoded n-tuple* $\tilde{X}^n$ *is unequal to the transmitted n-tuple* X^n*, is lowerbounded by*

$$R - C \le H_b(\Pr(e)) + R\Pr(e), \tag{8.75}$$

where $H_b(\alpha)$ *is the binary entropy,* $-\alpha\log\alpha - (1-\alpha)\log(1-\alpha)$.

Remark Note that the right side of (8.75) is 0 at $\Pr(e) = 0$ and is increasing for $\Pr(e) \le 1/2$, so (8.75) provides a lowerbound to $\Pr(e)$ that depends only on C and R.

Proof Note that $\mathsf{H}[X^n] = nR$ and, from (8.72) and (8.69), $\mathsf{H}(X^n) - \mathsf{H}(X^n|Y^n) \le nC$. Thus

$$\mathsf{H}(X^n|Y^n) \ge nR - nC. \tag{8.76}$$

For each sample value y^n of Y^n, $\mathsf{H}(X^n\,|\,Y^n = y^n)$ is an ordinary entropy. The received y^n is decoded into some $\tilde{x}^n$, and the corresonding probability of error is $\Pr(X^n \ne \tilde{x}^n\,|\,Y^n = y^n)$. The Fano inequality (see Exercise 2.20) states that the entropy $\mathsf{H}(X^n\,|\,Y^n = y^n)$ can be upperbounded as the sum of two terms: first the binary entropy of whether or not $X^n = \tilde{x}^n$, and second the entropy of all $M-1$ possible errors in the case $X^n \ne \tilde{x}^n$, i.e.

$$\mathsf{H}(X^n\,|\,Y^n = y^n) \le H_b(\Pr(e|y^n)) + \Pr(e|y^n)\log(M-1).$$

Upperbounding $\log(M-1)$ by $\log M = nR$ and averaging over Y^n yields

$$\mathsf{H}(X^n|Y^n) \le H_b(\Pr(e)) + nR\Pr(e). \tag{8.77}$$

Combining (8.76) and (8.77), we obtain

$$R - C \leq \frac{H_b(\Pr(e))}{n} + R\Pr(e),$$

and upperbounding $1/n$ by 1 yields (8.75). □

Theorem 8.7.1 is not entirely satisfactory, since it shows that the probability of block error cannot be made negligible at rates above capacity, but it does not rule out the possibility that each block error causes only one bit error, say, and thus the probability of bit error might go to 0 as $n \to \infty$. As shown in Gallager (1968, theorem 4.3.4), this cannot happen, but the proof does not add much insight and will be omitted here.

8.7.4 Noisy-channel coding theorem, forward part

There are two critical ideas in the forward part of the coding theorem. The first is to use the AEP on the joint ensemble X^nY^n. The second, however, is what shows the true genius of Shannon. His approach, rather than an effort to find and analyze good codes, was simply to choose each codeword of a code randomly, choosing each letter in each codeword to be iid with the capacity yielding input distribution.

One would think initially that the codewords should be chosen to be maximally different in some sense, but Shannon's intuition said that independence would be enough. Some initial sense of why this might be true comes from looking at the binary orthogonal codes. Here each codeword of length n differs from each other codeword in $n/2$ positions, which is equal to the average number of differences with random choice. Another initial intuition comes from the fact that mutual information between input and output n-tuples is maximized by iid inputs. Truly independent inputs do not allow for coding constraints, but choosing a limited number of codewords using an iid distribution is at least a plausible approach. In any case, the following theorem proves that this approach works.

It clearly makes no sense for the encoder to choose codewords randomly if the decoder does not know what those codewords are, so we visualize the designer of the modem as choosing these codewords and building them into both transmitter and receiver. Presumably the designer is smart enough to test a code before shipping a million copies around the world, but we won't worry about that. We simply average the performance over all random choices. Thus the probability space consists of M independent iid codewords of block length n, followed by a randomly chosen message m, $0 \leq m \leq M-1$, that enters the encoder. The corresponding sample value x_m^n of the mth randomly chosen codeword is transmitted and combined with noise to yield a received sample sequence y^n. The decoder then compares y^n with the M possible randomly chosen messages (the decoder knows $x_0^n, \ldots, x_{M-1}^n$, but not m) and chooses the most likely of them. It appears that a simple problem has been replaced by a complex problem, but since there is so much independence between all the random symbols, the new problem is surprisingly simple.

These randomly chosen codewords and channel outputs are now analyzed with the help of the AEP. For this particular problem, however, it is simpler to use a slightly different form of AEP, called the *strong AEP*, than that of Chapter 2. The strong AEP was analyzed in Exercise 2.28 and is reviewed here. Let $U^n = U_1, \dots, U_n$ be an n-tuple of iid discrete random symbols with alphabet $\mathcal{U}$ and letter probabilities p_j for each $j \in \mathcal{U}$. Then, for any $\varepsilon > 0$, the strongly typical set $S_\varepsilon(U^n)$ of sample n-tuples is defined as follows:

$$S_\varepsilon(U^n) = \left\{ u^n : p_j(1-\varepsilon) < \frac{N_j(u^n)}{n} < p_j(1+\varepsilon); \quad \text{for all } j \in \mathcal{U} \right\}, \tag{8.78}$$

where $N_j(u^n)$ is the number of appearances of letter j in the n-tuple u^n. The double inequality in (8.78) will be abbreviated as $N_j(u^n) \in np_j(1 \pm \varepsilon)$, so (8.78) becomes

$$S_\varepsilon(U^n) = \{u^n : N_j(u^n) \in np_j(1 \pm \varepsilon); \quad \text{for all } j \in \mathcal{U}\}. \tag{8.79}$$

Thus, the strongly typical set is the set of n-tuples for which each letter appears with approximately the right relative frequency. For any given ε, the law of large numbers says that $\lim_{n\to\infty} \Pr(N_j(U^n) \in p_j(1 \pm \varepsilon)) = 1$ for each j. Thus (see Exercise 2.28),

$$\lim_{n\to\infty} \Pr(U^n \in S_\varepsilon(U^n)) = 1. \tag{8.80}$$

Next consider the probability of n-tuples in $S_\varepsilon(U^n)$. Note that $p_{U^n}(u^n) = \prod_j p_j^{N_j(u^n)}$. Taking the log of this, we see that

$$\begin{aligned} \log p_{U^n}(u^n) &= \sum_j N_j(u^n) \log p_j \\ &\in \sum_j p_j(1 \pm \varepsilon) \log p_j; \\ \log p_{U^n}(u^n) &\in -n\mathsf{H}(U)(1 \pm \varepsilon), \qquad \text{for } u^n \in S_\varepsilon(U^n). \end{aligned} \tag{8.81}$$

Thus the strongly typical set has the same basic properties as the typical set defined in Chapter 2. Because of the requirement that each letter has a typical number of appearances, however, it has additional properties that are useful in the coding theorem that follows.

Consider an n-tuple of channel input/output pairs, $X^nY^n = (X_1Y_1), (X_2Y_2), \dots, (X_nY_n)$, where successive pairs are iid. For each pair XY, let X have the pmf $\{p_i; i \in \mathcal{X}\}$, which achieves capacity in (8.73). Let the pair XY have the pmf $\{p_iP_{i,j}; i \in \mathcal{X}, j \in \mathcal{Y}\}$, where $P_{i,j}$ is the channel transition probability from input i to output j. This is the joint pmf for the randomly chosen codeword that is transmitted and the corresponding received sequence.

The strongly typical set $S_\varepsilon(X^nY^n)$ is then given by (8.79) as follows:

$$S_\varepsilon(X^nY^n) = \{x^ny^n : N_{ij}(x^ny^n) \in n\,p_iP_{i,j}(1 \pm \varepsilon); \quad \text{for all } i \in \mathcal{X}, j \in \mathcal{Y}\}, \tag{8.82}$$

where $N_{ij}(x^ny^n)$ is the number of xy pairs in $((x_1y_1), (x_2y_2), \dots, (x_ny_n))$ for which $x = i$ and $y = j$. Using the same argument as in (8.80), the transmitted codeword X^n and the received n-tuple Y^n jointly satisfy

$$\lim_{n\to\infty} \Pr[(X^nY^n) \in S_\varepsilon(X^nY^n)] = 1. \tag{8.83}$$

Applying the same argument as in (8.81) to the pair x^ny^n, we obtain

$$\log p_{X^nY^n}(x^ny^n) \in -n\mathsf{H}(XY)(1 \pm \varepsilon), \qquad \text{for } (x^ny^n) \in S_\varepsilon(X^nY^n). \tag{8.84}$$

The nice feature about strong typicality is that if x^ny^n is in the set $S_\varepsilon(X^nY^n)$ for a given pair x^ny^n, then the given x^n must be in $S_\varepsilon(X^n)$ and the given Y must be in $S_\varepsilon(Y^n)$. To see this, assume that $(x^ny^n) \in S_\varepsilon(X^nY^n)$. Then, by definition, $N_{ij}(x^ny^n) \in np_iP_{ij}(1 \pm \varepsilon)$ for all i, j. Thus,

$$\begin{aligned} N_i(x^n) &= \sum_j N_{ij}(x^ny^n) \\ &\in \sum_j np_iP_{ij}(1 \pm \varepsilon) = np_i(1 \pm \varepsilon), \qquad \text{for all } i. \end{aligned}$$

Thus $x^n \in S_\varepsilon(X^n)$. By the same argument, $y^n \in S_\varepsilon(Y^n)$.

Theorem 8.7.2 *Consider a DMC with capacity C and let R be any fixed rate $R < C$. Then, for any $\delta > 0$ and all sufficiently large block lengths n, there exist block codes with $M \geq 2^{nR}$ equiprobable codewords such that the ML error probability satisfies* $\Pr(e) \leq \delta$.

Proof As suggested in the preceding discussion, we consider the error probability averaged over the random selection of codes defined above, where, for given block length n and rate R, the number of codewords will be $M = \lceil 2^{nR} \rceil$. Since at least one code must be as good as the average, the theorem can be proved by showing that $\Pr(e) \leq \delta$.

The decoding rule to be used will be different from maximum likelihood, but since the ML rule is optimum for equiprobable messages, proving that $\Pr(e) \leq \delta$ for any decoding rule will prove the theorem. The rule to be used is strong typicality. That is, the decoder, after observing the received sequence y^n, chooses a codeword x^n_m for which $x^n_m y^n$ is jointly typical, i.e. for which $x^n_m y^n \in S_\varepsilon(X^n Y^n)$ for some ε to be determined later. An error is said to be made if either $x^n_m \notin S_\varepsilon(X^n Y^n)$ for the message m actually transmitted or if $x^n_{m'} y^n \in S_\varepsilon(X^n Y^n)$ for some $m' \neq m$. The probability of error given message m is then upperbounded by two terms: $\Pr[X^n Y^n \notin S_\varepsilon(X^n Y^n)]$, where $X^n Y^n$ is the transmitted/received pair, and the probability that some other codeword is jointly typical with Y^n. The other codewords are independent of Y and each are chosen with iid symbols using the same pmf as the transmitted codeword. Let $\overline{X}^n$ be any one of these codewords. Using the union bound,

$$\Pr(e) \leq \Pr[(X^nY^n) \notin S_\varepsilon(X^nY^n)] + (M-1)\Pr\left[(\overline{X}^nY^n) \in S_\varepsilon(X^nY^n)\right]. \tag{8.85}$$

For any large enough n, (8.83) shows that the first term is at most $\delta/2$. Also $M - 1 \leq 2^{nR}$. Thus

$$\Pr(e) \leq \frac{\delta}{2} + 2^{nR}\Pr\left[(\overline{X}^nY^n) \in S_\varepsilon(X^nY^n)\right]. \tag{8.86}$$

To analyze (8.86), define $F(y^n)$ as the set of input sequences x^n that are jointly typical with any given y^n. This set is empty if $y^n \notin S_\varepsilon(Y^n)$. Note that, for $y^n \in S_\varepsilon(Y^n)$,

$$p_{Y^n}(y^n) \geq \sum_{x^n \in F(y^n)} p_{X^nY^n}(x^ny^n) \geq \sum_{x^n \in F(y^n)} 2^{-n\mathsf{H}(XY)(1+\varepsilon)},$$

where the final inequality comes from (8.84). Since $p_{Y^n}(y^n) \leq 2^{-n\mathsf{H}(Y)(1-\varepsilon)}$ for $y^n \in S_\varepsilon(Y^n)$, the conclusion is that the number of n-tuples in $F(y^n)$ for any typical y^n satisfies

$$|F(y^n)| \leq 2^{n[\mathsf{H}(XY)(1+\varepsilon)-\mathsf{H}(Y)(1-\varepsilon)]}. \tag{8.87}$$

This means that the probability that $\overline{X}^n$ lies in $F(y^n)$ is at most the size $|F(y^n)|$ times the maximum probability of a typical $\overline{X}^n$ (recall that $\overline{X}^n$ is independent of Y^n but has the same marginal distribution as X^n). Thus,

$$\begin{aligned}\Pr\left[(\overline{X}^nY^n) \in S_\varepsilon(X^nY^n)\right] &\leq 2^{-n[\mathsf{H}(X)(1-\varepsilon)+\mathsf{H}(Y)(1-\varepsilon)-\mathsf{H}(XY)(1+\varepsilon)]}\\ &= 2^{-n\{C-\varepsilon[\mathsf{H}(X)+\mathsf{H}(Y)+\mathsf{H}(XY)]\}},\end{aligned}$$

where we have used the fact that $C = \mathsf{H}(X) - \mathsf{H}(X|Y) = \mathsf{H}(X) + \mathsf{H}(Y) - \mathsf{H}(XY)$. Substituting this into (8.86) yields

$$\Pr(e) \leq \frac{\delta}{2} + 2^{n(R-C+\varepsilon\alpha)},$$

where $\alpha = \mathsf{H}(X) + \mathsf{H}(Y) + \mathsf{H}(XY)$. Finally, choosing $\varepsilon = (C-R)/(2\alpha)$,

$$\Pr(e) \leq \frac{\delta}{2} + 2^{-n(C-R)/2} \leq \delta$$

for sufficiently large n. □

The above proof is essentially the original proof given by Shannon, with a little added explanation of details. It will be instructive to explain the essence of the proof without any of the epsilons or deltas. The transmitted and received n-tuple pair (X^nY^n) is typical with high probability and the typical pairs essentially have probability $2^{-n\mathsf{H}(XY)}$ (including both the random choice of X^n and the random noise). Each typical output y^n essentially has a marginal probability $2^{-n\mathsf{H}(Y)}$. For each typical y^n, there are essentially $2^{n\mathsf{H}(X|Y)}$ input n-tuples that are jointly typical with y^n (this is the nub of the proof). An error occurs if any of these are selected to be codewords (other than the actual transmitted codeword). Since there are about $2^{n\mathsf{H}(X)}$ typical input n-tuples altogether, a fraction $2^{-n\mathsf{I}[X;Y]} = 2^{-nC}$ of them are jointly typical with the given received y^n.

More recent proofs of the noisy-channel coding theorem also provide much better upperbounds on error probability. These bounds are exponentially decreasing with n, with a rate of decrease that typically becomes vanishingly small as $R \to C$.

The error probability in the theorem is the block error probability averaged over the codewords. This clearly upperbounds the error probability per transmitted binary digit. The theorem can also be easily modified to apply uniformly to each codeword in the code. One simply starts with twice as many codewords as required and eliminates the ones with greatest error probability. The ε and δ in the theorem can still be made arbitrarily small. Usually encoders contain a scrambling operation between input and output to provide privacy for the user, so a uniform bound on error probability is usually unimportant.

8.7.5 The noisy-channel coding theorem for WGN

The coding theorem for DMCs can be easily extended to discrete-time channels with arbitrary real or complex input and output alphabets, but doing this with mathematical generality and precision is difficult with our present tools.

This extension is carried out for the discrete-time Gaussian channel, which will make clear the conditions under which this generalization is easy. Let X_k and Y_k be the input and output to the channel at time k, and assume that $Y_k = X_k + Z_k$, where $Z_k \sim \mathcal{N}(0, N_0/2)$ is independent of X_k and independent of the signal and noise at all other times. Assume the input is constrained in second moment to $\mathsf{E}[X_k^2] \le E$, so $\mathsf{E}[Y^2] \le E + N_0/2$.

From Exercise 3.8, the differential entropy of Y is then upperbounded by

$$\mathsf{h}(Y) \le \frac{1}{2}\log[2\pi e(E+N_0/2)]. \tag{8.88}$$

This is satisfied with equality if Y is $\mathcal{N}(0, E+N_0/2)$, and thus if X is $\mathcal{N}(0, E)$. For any given input x, $\mathsf{h}(Y|X=x) = (1/2)\log(2\pi e N_0/2)$, so averaging over the input space yields

$$\mathsf{h}(Y|X) = \frac{1}{2}\log(2\pi e N_0/2). \tag{8.89}$$

By analogy with the DMC case, let the capacity C (in bits per channel use) be defined as the maximum of $\mathsf{h}(Y) - \mathsf{h}(Y|X)$ subject to the second moment constraint E. Thus, combining (8.88) and (8.89), we have

$$C = \frac{1}{2}\log\left(1+\frac{2E}{N_0}\right). \tag{8.90}$$

Theorem 8.7.2 applies quite simply to this case. For any given rate R in bits per channel use such that $R < C$, one can quantize the channel input and output space finely enough such that the corresponding discrete capacity is arbitrarily close to C and in particular larger than R. Then Theorem 8.7.2 applies, so rates arbitrarily close to C can be transmitted with arbitrarily high reliability. The converse to the coding theorem can be extended in a similar way.

For a discrete-time WGN channel using $2W$ degrees of freedom per second and a power constraint P, the second moment constraint on each degree of freedom[13]

[13] We were careless in not specifying whether the constraint must be satisfied for each degree of freedom or overall as a time average. It is not hard to show, however, that the mutual information is maximized when the same energy is used in each degree of freedom.

becomes $E = P/2W$ and the capacity C_t in bits per second becomes Shannon's famous formula:

$$C_t = W \log\left(1 + \frac{P}{WN_0}\right). \tag{8.91}$$

This is then the capacity of a WGN channel with input power constrained to P and degrees of freedom per second constrained to $2W$.

With some careful interpretation, this is also the capacity of a continuous-time channel constrained in bandwidth to W and in power to P. The problem here is that if the input is strictly constrained in bandwidth, no information at all can be transmitted. That is, if a single bit is introduced into the channel at time 0, the difference in the waveform generated by symbol 1 and that generated by symbol 0 must be zero before time 0, and thus, by the Paley–Wiener theorem, cannot be nonzero and strictly bandlimited. From an engineering perspective, this does not seem to make sense, but the waveforms used in all engineering systems have negligible but nonzero energy outside the nominal bandwidth.

Thus, to use (8.91) for a bandlimited input, it is necessary to start with the constraint that, for any given $\eta > 0$, at least a fraction $(1 - \eta)$ of the energy must lie within a bandwidth W. Then reliable communication is possible at all rates R_t in bits per second less than C_t as given in (8.91). Since this is true for all $\eta > 0$, no matter how small, it makes sense to call this the capacity of the bandlimited WGN channel. This is not an issue in the design of a communication system, since filters must be used, and it is widely recognized that they cannot be entirely bandlimited.

8.8 Convolutional codes

The theory of coding, and particularly of coding theorems, concentrates on block codes, but convolutional codes are also widely used and have essentially no block structure. These codes can be used whether bandwidth is highly constrained or not. We give an example below where there are two output bits for each input bit. Such a code is said to have rate 1/2 (in input bits per channel bit). More generally, such codes produce an m-tuple of output bits for each b-tuple of input bits for arbitrary integers $0 < b < m$. These codes are said to have rate b/m.

A convolutional code looks very much like a discrete filter. Instead of having a single input and output stream, however, we have b input streams and m output streams. For the example of a convolutional code in Figure 8.8, the number of input streams is $b = 1$ and the number of output streams is $m = 2$, thus producing two output bits per input bit. There is another difference between a convolutional code and a discrete filter; the inputs and outputs for a convolutional code are binary and the addition is modulo 2.

As indicated in Figure 8.8, the outputs for this convolutional code are given by

$$\begin{aligned} U_{k,1} &= D_k \oplus D_{k-1} \oplus D_{k-2}, \\ U_{k,2} &= D_k \qquad\quad\; \oplus D_{k-2}. \end{aligned}$$

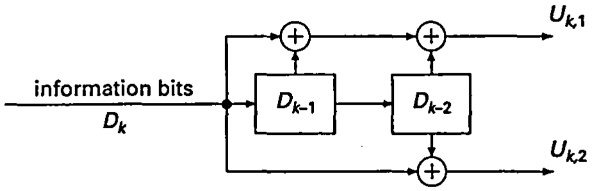


Figure 8.8. Example of a convolutional code.

Thus, each of the two output streams are linear modulo 2 convolutions of the input stream. This encoded pair of binary streams can now be mapped into a pair of signal streams such as antipodal signals $\pm a$. This pair of signal streams can then be interleaved and modulated by a single stream of Nyquist pulses at twice the rate. This baseband waveform can then be modulated to passband and transmitted.

The structure of this code can be most easily visualized by a "trellis" diagram as illustrated in Figure 8.9.

To understand this trellis diagram, note from Figure 8.8 that the encoder is characterized at any epoch k by the previous binary digits, D_{k-1} and D_{k-2}. Thus the encoder has four possible states, corresponding to the four possible values of the pair D_{k-1}, D_{k-2}. Given any of these four states, the encoder output and the next state depend only on the current binary input. Figure 8.9 shows these four states arranged vertically and shows time horizontally. We assume the encoder starts at epoch 0 with $D_{-1} = D_{-2} = 0$.

In the convolutional code of Figure 8.8 and 8.9, the output at epoch k depends on the current input and the previous two inputs. In this case, the *constraint length* of the code is 2. In general, the output could depend on the input and the previous n inputs, and the constraint length is then defined to be n. If the constraint length is n (and a single binary digit enters the encoder at each epoch k), then there are 2^n possible states, and the trellis diagram contains 2^n rather than 4 nodes at each time instant.

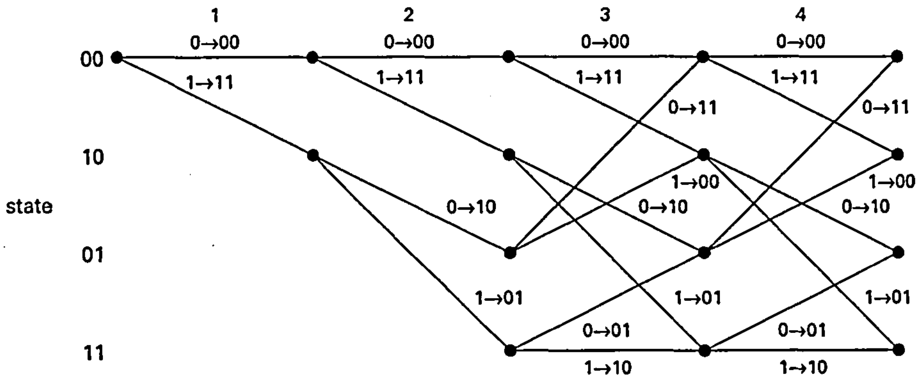


Figure 8.9. Trellis diagram; each transition is labeled with the input and corresponding output.

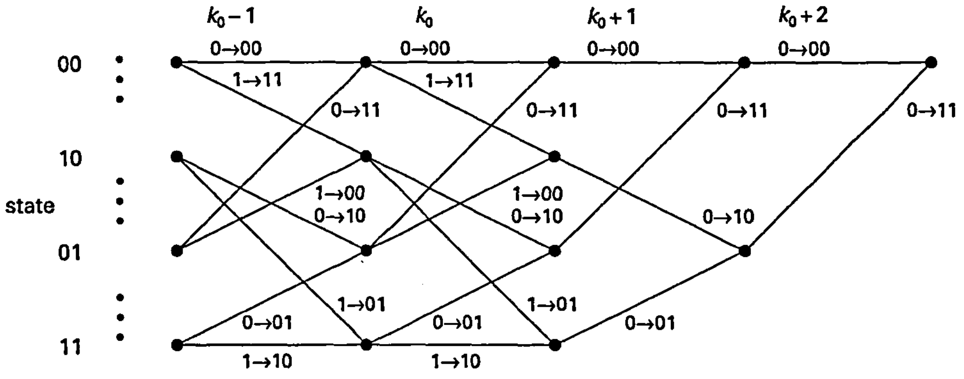


Figure 8.10. Trellis termination.

As we have described convolutional codes above, the encoding starts at time 1 and then continues forever. In practice, because of packetization of data and various other reasons, the encoding usually comes to an end after some large number, say k_0, of binary digits have been encoded. After D_{k_0} enters the encoder, two final 0s enter the encoder, at epochs k_0+1 and k_0+2, and four final encoded digits come out of the encoder. This restores the state of the encoder to state 0, which, as we see later, is very useful for decoding. For the more general case with a constraint length of n, we need n final 0s to restore the encoder to state 0. Altogether, k_0 inputs lead to $2(k_0+n)$ outputs, for a code rate of $k_0/[2(k_0+n)]$. This is referred to as a terminated rate 1/2 code. Figure 8.10 shows the part of the trellis diagram corresponding to this termination.

8.8.1 Decoding of convolutional codes

Decoding a convolutional code is essentially the same as using detection theory to choose between each pair of codewords, and then choosing the best overall (the same as done for the orthogonal code). There is one slight conceptual difference in that, in principle, the encoding continues forever. When the code is terminated, however, this problem does not exist, and in principle one takes the maximum likelihood (ML) choice of all the (finite length) possible codewords.

As usual, assume that the incoming binary digits are iid and equiprobable. This is reasonable if the incoming bit stream has been source encoded. This means that the codewords of any given length are equally likely, which then justifies ML decoding.

Maximum likelihood detection is also used so that codes for error correction can be designed independently of the source data to be transmitted.

Another issue, given iid inputs, is determining what is meant by probability of error. In all of the examples discussed so far, given a received sequence of symbols, we have attempted to choose the codeword that minimizes the probability of error for the entire codeword. An alternative would have been to minimize the probability of error individually for each binary information digit. It turns out to be easier to minimize the sequence error probability than the bit error probability. This, in fact, is what happens when we use ML detection between codewords, as suggested above.

In decoding for error correction, the objective is almost invariably to minimize the sequence probability of error. Along with the convenience suggested here, a major reason is that a binary input is usually a source-coded version of some other source sequence or waveform, and thus a single output error is often as serious as multiple errors within a codeword. Note that ML detection on sequences is assumed in what follows.

8.8.2 The Viterbi algorithm

The Viterbi algorithm is an algorithm for performing ML detection for convolutional codes. Assume for the time being that the code is terminated as in Figure 8.10. It will soon be seen that whether or not the code is terminated is irrelevant. The algorithm will now be explained for the convolutional code in Figure 8.8 and for the assumption of WGN; the extension to arbitrary convolutional codes will be obvious except for the notational complexity of the general case. For any given input $d_1, \dots, d_{k_0}$, let the encoded sequence be $u_{1,1}, u_{1,2}, u_{2,1}, u_{2,2}, \dots, u_{k_0+2,2}$, and let the channel output, after modulation, addition of WGN, and demodulation, be $v_{1,1}, v_{1,2}, v_{2,1}, v_{2,2}, \dots, v_{k_0+2,2}$.

There are 2^{k_0} possible codewords, corresponding to the 2^{k_0} possible binary k_0-tuples $d_1, \dots, d_{k_0}$, so a naive approach to decoding would be to compare the likelihood of each of these codewords. For large k_0, even with today's technology, such an approach would be prohibitive. It turns out, however, that by using the trellis structure of Figure 8.9, this decoding effort can be greatly simplified.

Each input $d_1, \dots, d_{k_0}$ (i.e. each codeword) corresponds to a particular path through the trellis from epoch 1 to k_0+2, and each path, at each epoch k, corresponds to a particular trellis state.

Consider two paths $d_1, \dots, d_{k_0}$ and $d'_1, \dots, d'_{k_0}$ through the trellis that pass through the same state at time k^+ (i.e. at the time immediately after the input and state change at epoch k) and remain together thereafter. Thus, $d_{k+1}, \dots, d_{k_0} = d'_{k+1}, \dots, d'_{k_0}$. For example, from Figure 8.8, we see that both $(0, \dots, 0)$ and $(1, 0, \dots, 0)$ are in state 00 at 3^+ and both remain in the same state thereafter. Since the two paths are in the same state at k^+ and have the same inputs after this time, they both have the same encoder outputs after this time. Thus $u_{k+1,i}, \dots, u_{k_0+2,i} = u'_{k+1,i}, \dots, u'_{k_0+2,i}$ for $i = 1, 2$.

Since each channel output rv $V_{k,i}$ is given by $V_{k,i} = U_{k,i} + Z_{k,i}$ and the Gaussian noise variables $Z_{k,i}$ are independent, this means that, for any channel output $v_{1,1}, \dots, v_{k_0+2,2}$,

$$\frac{f(v_{1,1}, \dots, v_{k_0+2,2} | d_1, \dots, d_{k_0})}{f(v_{1,1}, \dots, v_{k_0+2,2} | d'_1, \dots, d'_{k_0})} = \frac{f(v_{1,1}, \dots, v_{k,2} | d_1, \dots, d_{k_0})}{f(v_{1,1}, \dots, v_{k,2} | d'_1, \dots, d'_{k_0})}.$$

In plain English, this says that if two paths merge at time k^+ and then stay together, the likelihood ratio depends on only the first k output pairs. Thus if the right side exceeds 1, then the path $d_1, \ldots, d_{k_0}$ is more likely than the path $d'_1, \ldots, d'_{k_0}$. This conclusion holds no matter how the final inputs $d_{k+1}, \ldots, d_{k_0}$ are chosen.

We then see that when two paths merge at a node, no matter what the remainder of the path is, the most likely of the paths is the one that is most likely at the point of the merger. Thus, whenever two paths merge, the least likely of the paths can be eliminated at that point. Doing this elimination successively from the smallest k for which paths merge (3 in our example), there is only one survivor for each state at each epoch.

To be specific, let $h(d_1, \ldots, d_k)$ be the state at time k^+ with input $d_1, \ldots, d_k$. For our example, $h(d_1, \ldots, d_k) = (d_{k-1}, d_k)$. Let

$$f_{\max}(k, s) = \max_{h(d_1,\ldots,d_k)=s} f(v_{1,1}, \ldots, v_{k,2} | d_1, \ldots, d_k).$$

These quantities can then be calculated iteratively for each state and each time k by the following iteration:

$$f_{\max}(k+1, s) = \max_{r:r\rightarrow s} f_{\max}(k, r) \cdot f(v_{k,1}|u_1(r\rightarrow s)) f(v_{k,2}|u_2(r\rightarrow s)), \qquad (8.92)$$

where the maximization is over the set of states r that have a transition to state s in the trellis and $u_1(r\rightarrow s)$ and $u_2(r\rightarrow s)$ are the two outputs from the encoder corresponding to a transition from r to s.

This expression is simplified (for WGN) by taking the log, which is proportional to the negative squared distance between $\boldsymbol{v}$ and $\boldsymbol{u}$. For the antipodal signal case in the example, this may be further simplified by simply taking the dot product between $\boldsymbol{v}$ and $\boldsymbol{u}$. Letting $L(k, s)$ be this dot product,

$$L(k+1, s) = \max_{r.r\rightarrow s} L(k, r) + v_{k,1}u_1(r\rightarrow s) + v_{k,2}u_2(r\rightarrow s). \qquad (8.93)$$

What this means is that at each epoch $(k+1)$, it is necessary to calculate the inner product in (8.93) for each link in the trellis going from k to $k+1$. These must be maximized over r for each state s at epoch $(k+1)$. The maximum must then be saved as $L(k+1, s)$ for each s. One must, of course, also save the paths taken in arriving at each merging point.

Those familiar with dynamic programming will recognize this recursive algoriothm as an example of the dynamic programming principle.

The complexity of the entire computation for decoding a block of k_0 information bits is proportional to $4(k_0+2)$. In the more general case, where the constraint length of the convolutional coder is n rather than 2, there are 2^n states and the computational complexity is proportional to $2^n(k_0+n)$. The Viterbi algorithm is usually used in cases where the constraint length is moderate, say 6–12, and in these situations the computation is quite moderate, especially compared with 2^{k_0}.

Usually one does not wait until the end of the block to start decoding. When the above computation is performed at epoch k, all the paths up to k' have merged for k' a few constraint lengths less than k. In this case, one can decode without any bound on k_0, and the error probability is viewed in terms of "error events" rather than block error.

8.9 Summary of detection, coding, and decoding

This chapter analyzed the last major segment of a general point-to-point communication system in the presence of noise, namely how to detect the input signals from the noisy version presented at the output. Initially the emphasis was on detection alone; i.e., the assumption was that the rest of the system had been designed and the only question remaining was how to extract the signals.

At a very general level, the problem of detection in this context is trivial. That is, under the assumption that the statistics of the input and the noise are known, the sensible rule is maximum a-posteriori probability decoding: find the a-posteriori probability of all the hypotheses and choose the largest. This is somewhat complicated by questions of whether to carry out sequence detection or bit detection, but these questions are details in a sense.

At a more specific level, however, the detection problem led to many interesting insights and simplifications, particularly for WGN channels. A particularly important simplification is the principle of irrelevance, which says that components of the received waveform in degrees of freedom not occupied by the signal of interest (or statistically related signals) can be ignored in detection of those signals. Looked at in another way, this says that matched filters could be used to extract the degrees of freedom of interest.

The last part of the chapter discussed coding and decoding. The focus changed here to the question of how coding can change the input waveforms so as to make the decoding more effective. In other words, a MAP detector can be designed for any signal structure, but the real problem is to design both the signal structure and detection for effective performance.

At this point, the noisy-channel coding theorem comes into the picture. If $R < C$, then the probability of error can be reduced arbitrarily, meaning that finding the optimal code at a given constraint length is slightly artificial. What is needed is a good trade-off between error probability and the delay and complexity caused by longer constraint lengths.

Thus the problem is not only to overcome the noise, but also to do this with reasonable delay and complexity. Chapter 9 considers some of these problems in the context of wireless communication.

8.10 Appendix: Neyman–Pearson threshold tests

We have seen in the preceding sections that any binary MAP test can be formulated as a comparison of a likelihood ratio with a threshold. It turns out that many other detection rules can also be viewed as threshold tests on likelihood ratios. One of the most important binary detection problems for which a threshold test turns out to be essentially optimum is the *Neyman–Pearson test.* This is often used in those situations in which there is no sensible way to choose a-priori probabilities. In the Neyman–Pearson test, an acceptable value α is established for $\Pr\{e|U = 1\}$, and, subject to

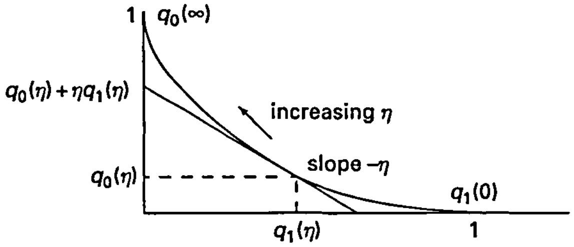


Figure 8.11. The error curve; $q_1(\eta)$ and $q_0(\eta)$ are plotted as parametric functions of η.

the constraint $\Pr\{e|U=1\} \le \alpha$, the Neyman–Pearson test minimizes $\Pr\{e|U=0\}$. We shall show in what follows that such a test is essentially a threshold test. Before demonstrating this, we need some terminology and definitions.

Define $q_0(\eta)$ to be $\Pr\{e|U=0\}$ for a threshold test with threshold η, $0<\eta<\infty$, and similarly define $q_1(\eta)$ as $\Pr\{e|U=1\}$. Thus, for $0<\eta<\infty$,

$$q_0(\eta)=\Pr\{\Lambda(V)<\eta|U=0\}; \qquad q_1(\eta)=\Pr\{\Lambda(V)\geq\eta|U=1\}. \tag{8.94}$$

Define $q_0(0)$ as $\lim_{\eta\to 0} q_0(\eta)$ and $q_1(0)$ as $\lim_{\eta\to 0} q_1(\eta)$. Clearly, $q_0(0)=0$, and in typical situations $q_1(0)=1$. More generally, $q_1(0)=\Pr\{\Lambda(V)>0|U=1\}$. In other words, $q_1(0)<1$ if there is some set of observations that are impossible under $U=0$ but have positive probability under $U=1$. Similarly, define $q_0(\infty)$ as $\lim_{\eta\to\infty} q_0(\eta)$ and $q_1(\infty)$ as $\lim_{\eta\to\infty} q_1(\eta)$. We have $q_0(\infty)=\Pr\{\Lambda(V)<\infty\}$ and $q_1(\infty)=0$.

Finally, for an arbitrary test A, threshold or not, denote $\Pr\{e\,|\,U=0\}$ as $q_0(A)$ and $\Pr\{e|U=1\}$ as $q_1(A)$.

Using (8.94), we can plot $q_0(\eta)$ and $q_1(\eta)$ as parametric functions of η; we call this the *error curve*.[14] Figure 8.11 illustrates this error curve for a typical detection problem such as (8.17) and (8.18) for antipodal binary signalling. We have already observed that, as the threshold η is increased, the set of v mapped into $\tilde{U}=0$ decreases. Thus $q_0(\eta)$ is an increasing function of η and $q_1(\eta)$ is decreasing. Thus, as η increases from 0 to ∞, the curve in Figure 8.11 moves from the lower right to the upper left.

Figure 8.11 also shows a straight line of slope $-\eta$ through the point $(q_1(\eta), q_0(\eta))$ on the error curve. The following lemma shows why this line is important.

Lemma 8.10.1 *For each η, $0<\eta<\infty$, the line of slope $-\eta$ through the point $(q_1(\eta), q_0(\eta))$ lies on or beneath all other points $(q_1(\eta'), q_0(\eta'))$ on the error curve, and also lies beneath $(q_1(A), q_0(A))$ for all tests A.*

Before proving this lemma, we give an example of the error curve for a discrete observation space.

Example 8.10.1 (Discrete observations) Figure 8.12 shows the error curve for an example in which the hypotheses 0 and 1 are again mapped $0\to +a$ and $1\to -a$.

[14] In the radar field, one often plots $1-q_0(\eta)$ as a function of $q_1(\eta)$. This is called the receiver operating characteristic (ROC). If one flips the error curve vertically around the point 1/2, the ROC results.

v	$P(v\|0)$	$P(v\|1)$	$\Lambda(v)$
3	0.4	0.1	4
1	0.3	0.2	3/2
−1	0.2	0.3	2/3
−3	0.1	0.4	1/4

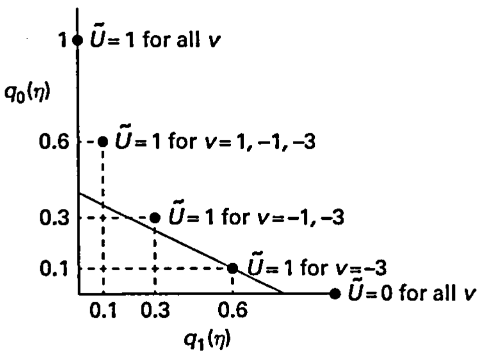


Figure 8.12. Error curve for a discrete observation space. There are only five points making up the "curve," one corresponding to each of the five distinct threshold rules. For example, the threshold rule $\tilde{U}=1$ only for $v=-3$ yields $(q_1(\eta), q_0(\eta)) = (0.6, 0.1)$ for all η in the range 1/4 to 2/3. A straight line of slope $-\eta$ through that point is also shown for $\eta = 1/2$. Lemma 8.10.1 asserts that this line lies on or beneath each point of the error curve and each point $(q_1(A), q_0(A)$ for any other test. Note that as η increases or decreases, this line will rotate around the point $(0.6, 0.1)$ until η becomes larger than 2/3 or smaller than 1/4; it then starts to rotate around the next point in the error curve.

Assume that the observation V can take on only four discrete values $+3, +1, -1, -3$. The probabilities of each of these values, conditional on $U=0$ and $U=1$, are given in Figure 8.12. As indicated, the likelihood ratio $\Lambda(v)$ then takes the values 4, 3/2, 2/3, and 1/4, corresponding, respectively, to $v = 3, 1, -1$, and -3.

A threshold test at η decides $\tilde{U}=0$ if and only if $\Lambda(V) \geq \eta$. Thus, for example, for any $\eta \leq 1/4$, all possible values of v are mapped into $\tilde{U}=0$. In this range, $q_1(\eta)=1$ since $U=1$ always causes an error. Also $q_0(\eta)=0$ since $U=0$ never causes an error. In the range $1/4 < \eta \leq 2/3$, since $\Lambda(-3)=1/4$, the value -3 is mapped into $\tilde{U}=1$ and all other values into $\tilde{U}=0$. In this range, $q_1(\eta)=0.6$, since, when $U=1$, an error occurs unless $V=-3$.

In the same way, all threshold tests with $2/3 < \eta \leq 3/2$ give rise to the decision rule that maps -1 and -3 into $\tilde{U}=1$ and 1 and 3 into $\tilde{U}=0$. In this range, $q_1(\eta) = q_0(\eta) = 0.3$. As shown, there is another decision rule for $3/2 < \eta \leq 4$ and a final decision rule for $\eta > 4$.

The point of this example is that a finite observation space leads to an error curve that is simply a finite set of points. It is also possible for a continuously varying set of outputs to give rise to such an error curve when there are only finitely many possible likelihood ratios. Figure 8.12 illustrates what Lemma 8.10.1 means for error curves consisting only of a finite set of points.

Proof of Lemma 8.10.1 Consider the line of slope $-\eta$ through the point $(q_1(\eta), q_0(\eta))$. From plane geometry, as illustrated in Figure 8.11, we see that the vertical axis intercept of this line is $q_0(\eta) + \eta q_1(\eta)$. To interpret this line, define p_0 and p_1 as a-priori probabilities such that $\eta = p_1/p_0$. The overall error probability for the corresponding MAP test is then given by

$$\begin{aligned} q(\eta) &= p_0 q_0(\eta) + p_1 q_1(\eta) \\ &= p_0[q_0(\eta) + \eta q_1(\eta)]; \quad \eta = p_1/p_0. \end{aligned} \tag{8.95}$$

Similarly, the overall error probability for an arbitrary test A with the same a-priori probabilities is given by

$$q(A) = p_0[q_0(A) + \eta q_1(A)]. \tag{8.96}$$

From Theorem 8.1.1, $q(\eta) \leq q(A)$, so, from (8.95) and (8.96), we have

$$q_0(\eta) + \eta\, q_1(\eta) \leq q_0(A) + \eta\, q_1(A). \tag{8.97}$$

We have seen that the left side of (8.97) is the vertical axis intercept of the line of slope $-\eta$ through $(q_1(\eta), q_0(\eta))$. Similarly, the right side is the vertical axis intercept of the line of slope $-\eta$ through $(q_1(A), q_0(A))$. This says that the point $(q_1(A), q_0(A))$ lies on or above the line of slope $-\eta$ through $(q_1(\eta), q_0(\eta))$. This applies to every test A, which includes every threshold test. □

Lemma 8.10.1 shows that if the error curve gives $q_0(\eta)$ as a differentiable function of $q_1(\eta)$ (as in the case of Figure 8.11), then the line of slope $-\eta$ through $(q_1(\eta), q_0(\eta))$ is a tangent, at point $(q_1(\eta), q_0(\eta))$, to the error curve. Thus in what follows we call this line the *η-tangent* to the error curve. Note that the error curve of Figure 8.12 is not really a curve, but rather a discrete set of points. Each η-tangent, as defined above and illustrated in the figure for $\eta = 2/3$, still lies on or beneath all of these discrete points. Each η-tangent has also been shown to lie below all achievable points $(q_1(A), q_0(A))$, for each arbitrary test A.

Since for each test A the point $(q_1(A), q_0(A))$ lies on or above each η-tangent, it also lies on or above the supremum of these η-tangents over $0 < \eta < \infty$. It also follows, then, that, for each η', $0 < \eta' < \infty$, $(q_1(\eta'), q_0(\eta'))$ lies on or above this supremum. Since $(q_1(\eta'), q_0(\eta'))$ also lies on the η'-tangent, it lies on or beneath the supremum, and thus must lie on the supremum. We conclude that each point of the error curve lies on the supremum of the η-tangents.

Although all points of the error curve lie on the supremum of the η-tangents, all points of the supremum are not necessarily points of the error curve, as seen from Figure 8.12. We shall see shortly, however, that all points on the supremum are achievable by a simple extension of threshold tests. Thus we call this supremum the *extended error curve*.

For the example in Figure 8.11, the extended error curve is the same as the error curve itself. For the discrete example in Figure 8.12, the extended error curve is shown in Figure 8.13.

To understand the discrete case better, assume that the extended error function has a straight line portion of slope $-\eta^*$ and horizontal extent γ. This implies that the distribution function of $\Lambda(V)$ given $U = 1$ has a discontinuity of magnitude γ at η^*. Thus there is a set $\mathcal{V}^*$ of one or more v with $\Lambda(v) = \eta^*$, $\Pr\{\mathcal{V}^*|U = 1\} = \gamma$, and $\Pr\{\mathcal{V}^*|U = 0\} = \eta^*\gamma$. For a MAP test with threshold η^*, the overall error probability

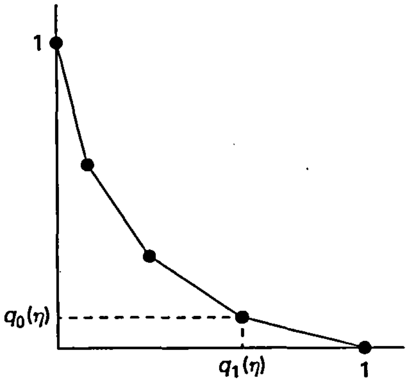


Figure 8.13. Extended error curve for the discrete observation example of Figure 8.12. From Lemma 8.10.1, for each slope $-\eta$, the η-tangent touches the error curve. Thus, the line joining two adjacent points on the error curve must be an η-tangent for its particular slope, and therefore must lie on the extended error curve.

is not affected by whether $v \in \mathcal{V}^*$ is detected as $\tilde{U} = 0$ or $\tilde{U} = 1$. Our convention is to detect $v \in \mathcal{V}^*$ as $\tilde{U} = 0$, which corresponds to the lower right point on the straight line portion of the extended error curve. The opposite convention, detecting $v \in \mathcal{V}^*$ as $\tilde{U} = 1$ reduces the error probability given $U = 1$ by γ and increases the error probability given $U = 0$ by $\eta^*\gamma$, i.e. it corresponds to the upper left point on the straight line portion of the extended error curve.

Note that when we were interested in MAP detection, it made no difference how $v \in \mathcal{V}^*$ was detected for the threshold η^*. For the Neyman–Pearson test, however, it makes a great deal of difference since $q_0(\eta^*)$ and $q_1(\eta^*)$ are changed. In fact, we can achieve any point on the straight line in question by detecting $v \in \mathcal{V}^*$ randomly, increasing the probability of choosing $\tilde{U} = 0$ to approach the lower right endpoint. In other words, the extended error curve is the curve relating q_1 to q_0 using a randomized threshold test. For a given η^*, of course, only those $v \in \mathcal{V}^*$ are detected randomly.

To summarize, the Neyman–Pearson test is a randomized threshold test. For a constraint α on $\Pr\{e|U = 1\}$, we choose the point α on the abscissa of the extended error curve and achieve the corresponding ordinate as the minimum $\Pr\{e|U = 1\}$. If that point on the extended error curve lies within a straight line segment of slope η^*, a randomized test is used for those observations with likelihood ratio η^*.

Since the extended error curve is a supremum of straight lines, it is a convex function. Since these straight lines all have negative slope, it is a monotonic decreasing[15] function. Thus, Figures 8.11 and 8.13 represent the general behavior of extended error curves, with the slight possible exception mentioned above that the endpoints need not have one of the error probabilities equal to 1.

The following theorem summarizes the results obtained for Neyman–Pearson tests.

[15] To be more precise, it is strictly decreasing between the endpoints $(q_1(\infty), q_0(\infty))$ and $(q_1(0), q_0(0))$.

Theorem 8.10.1 *The extended error curve is convex and strictly decreasing between* $(q_1(\infty), q_0(\infty))$ *and* $(q_1(0), q_0(0))$. *For a constraint* α *on* $\Pr\{e|U=1\}$, *the minimum value of*$\Pr\{e|U=0\}$ *is given by the ordinate of the extended error curve corresponding to the abscissa* α *and is achieved by a randomized threshold test.*

There is one more interesting variation on the theme of threshold tests. If the a-priori probabilities are unknown, we might want to minimize the maximum probability of error. That is, we visualize choosing a test followed by nature choosing a-priori probabilities to maximize the probability of error. Our objective is to minimize the probability of error under this worst case assumption. The resulting test is called a minmax test. It can be seen geometrically from Figures 8.11 or 8.13 that the minmax test is the randomized threshold test at the intersection of the extended error curve with a 45° line from the origin.

If there is symmetry between $U=0$ and $U=1$ (as in the Gaussian case), then the extended error curve will be symmetric around the 45° degree line, and the threshold will be at $\eta=1$ (i.e. the ML test is also the minmax test). This is an important result for Gaussian communication problems, since it says that ML detection, i.e. minimum distance detection, is robust in the sense of not depending on the input probabilities. If we know the a-priori probabilities, we can do better than the ML test, but we can do no worse.

8.11 Exercises

8.1 (Binary minimum cost detection)

(a) Consider a binary hypothesis testing problem with a-priori probabilities p_0, p_1 and likelihoods $f_{V|U}(v|i)$, $i=0,1$. Let C_{ij} be the cost of deciding on hypothesis j when i is correct. Conditional on an observation $V=v$, find the expected cost (over $U=0,1$) of making the decision $\tilde{U}=j$ for $j=0,1$. Show that the decision of minimum expected cost is given by

$$\tilde{U}_{\text{mincost}} = \arg\min_j \left[C_{0j}p_{U|V}(0|v) + C_{1j}p_{U|V}(1|v) \right].$$

(b) Show that the min cost decision above can be expressed as the following threshold test:

$$\Lambda(v) = \frac{f_{V|U}(v|0)}{f_{V|U}(v|1)} \begin{array}{c} \geq^{\tilde{U}=0} \\ <_{\tilde{U}=1} \end{array} \frac{p_1(C_{10}-C_{11})}{p_0(C_{01}-C_{00})} = \eta.$$

(c) Interpret the result above as saying that the only difference between a MAP test and a minimum cost test is an adjustment of the threshold to take account of the costs; i.e., a large cost of an error of one type is equivalent to having a large a-priori probability for that hypothesis.

8.2 Consider the following two equiprobable hypotheses:

$$U=0: V_1 = a\cos\Theta + Z_1, \quad V_2 = a\sin\Theta + Z_2;$$
$$U=1: V_1 = -a\cos\Theta + Z_1, \quad V_2 = -a\sin\Theta + Z_2.$$

Assume that Z_1 and Z_2 are iid $\mathcal{N}(0,\sigma^2)$ and that Θ takes on the values $\{-\pi/4, 0, \pi/4\}$, each with probability 1/3. Find the ML decision rule when V_1, V_2 are observed. [Hint. Sketch the possible values of V_1, V_2 for $Z=0$ given each hypothesis. Then, without doing any calculations, try to come up with a good intuitive decision rule. Then try to verify that it is optimal.]

8.3 Let

$$V_j = S_jX_j + Z_j, \qquad \text{for } 1 \le j \le 4,$$

where $\{X_j; 1 \le j \le 4\}$ are iid $\mathcal{N}(0,1)$ and $\{Z_j; 1 \le j \le 4\}$ are iid $\mathcal{N}(0,\sigma^2)$ and independent of $\{X_j; 1 \le j \le 4\}$. Assume that $\{V_j; 1 \le j \le 4\}$ are observed at the output of a communication system and the input is a single binary random variable U which is independent of $\{Z_j; 1 \le j \le 4\}$ and $\{X_j; 1 \le j \le 4\}$. Assume that $S_1, \dots, S_4$ are functions of U, with $S_1 = S_2 = U \oplus 1$ and $S_3 = S_4 = U$.

(a) Find the log likelihood ratio

$$\mathrm{LLR}(v) = \ln\left(\frac{f_{V|U}(v|0)}{f_{V|U}(x^n|1)}\right).$$

(b) Let $\mathcal{E}_a = |V_1|^2 + |V_2|^2$ and $\mathcal{E}_b = |V_3|^2 + |V_4|^2$. Explain why $\{\mathcal{E}_a, \mathcal{E}_b\}$ form a sufficient statistic for this problem and express the log likelihood ratio in terms of the sample values of $\{\mathcal{E}_a, \mathcal{E}_b\}$.

(c) Find the threshold for ML detection.

(d) Find the probability of error. [Hint. Review Exercise 6.1.] Note: we will later see that this corresponds to binary detection in Rayleigh fading.

8.4 Consider binary antipodal MAP detection for the real vector case. Modify the picture and argument in Figure 8.4 to verify the algebraic relation between the squared energy difference and the inner product in (8.22).

8.5 Derive (8.37), i.e. that $\sum_{k,j} y_{k,j}b_{k,j} = (1/2)\int y(t)b(t)\mathrm{d}t$. Explain the factor of 1/2.

8.6 In this problem, you will derive the inequalities

$$\left(1-\frac{1}{x^2}\right)\frac{1}{x\sqrt{2\pi}}\mathrm{e}^{-x^2/2} \le Q(x) \le \frac{1}{x\sqrt{2\pi}}\mathrm{e}^{-x^2/2}, \qquad \text{for } x>0, \qquad (8.98)$$

where $Q(x) = (2\pi)^{-1/2}\int_x^\infty \exp(-z^2/2)\mathrm{d}z$ is the "tail" of the Normal distribution. The purpose of this is to show that, when x is large, the right side of this inequality is a very tight upperbound on $Q(x)$.

(a) By using a simple change of variable, show that

$$Q(x) = \frac{1}{\sqrt{2\pi}}\mathrm{e}^{-x^2/2}\int_0^\infty \exp\left(-y^2/2 - xy\right)\mathrm{d}y.$$

(b) Show that

$$1 - y^2/2 \leq \exp(-y^2/2) \leq 1.$$

(c) Use parts (a) and (b) to establish (8.98).

8.7 (Other bounds on $Q(x)$)

(a) Show that the following bound holds for any γ and η such that $0 \leq \gamma$ and $0 \leq \eta$:

$$Q(\gamma + \eta) \leq Q(\gamma) \exp[-\eta\gamma - \eta^2/2].$$

[Hint. Start with $Q(\gamma+\eta) = (1/\sqrt{2\pi}) \int_{\gamma+\eta} \exp[-x^2/2] dx$ and use the change of variable $y = x - \eta$.]

(b) Use part (a) to show that, for all $\eta \geq 0$,

$$Q(\eta) \leq \frac{1}{2} \exp[-\eta^2/2].$$

(c) Use part (a) to show that, for all $0 \leq \gamma \leq w$,

$$\frac{Q(w)}{\exp[-w^2/2]} \leq \frac{Q(\gamma)}{\exp[-\gamma^2/2]}.$$

Note: equation (8.98) shows that $Q(w)$ goes to 0 with increasing w as a slowly varying coefficient times $\exp[-w^2/2]$. This demonstrates that the coefficient is decreasing for $w \geq 0$.

8.8 (Orthogonal signal sets) An *orthogonal signal set* is a set $\mathcal{A} = \{a_m, \ 0 \leq m \leq M-1\}$ of M orthogonal vectors in $\mathbb{R}^M$ with equal energy E; i.e., $\langle a_m, a_j \rangle = E\delta_{mj}$.

(a) Compute the spectral efficiency ρ of $\mathcal{A}$ in bits per two dimensions. Compute the average energy E_b per information bit.

(b) Compute the minimum squared distance $d^2_{\min}(\mathcal{A})$ between these signal points. Show that every signal has $M-1$ nearest neighbors.

(c) Let the noise variance be $N_0/2$ per dimension. Describe a ML detector on this set of M signals. [Hint. Represent the signal set in an orthonormal expansion where each vector is collinear with one coordinate. Then visualize making binary decisions between each pair of possible signals.]

8.9 (Orthogonal signal sets; continuation of Exercise 8.8) Consider a set $\mathcal{A} = \{a_m, 0 \leq m \leq M-1\}$ of M orthogonal vectors in $\mathbb{R}^M$ with equal energy E.

(a) Use the union bound to show that $\Pr(e)$, using ML detection, is bounded by

$$\Pr(e) \leq (M-1) Q(\sqrt{E/N_0}).$$

(b) Let $M \to \infty$ with $E_b = E/\log M$ held constant. Using the upperbound for $Q(x)$ in Exercise 8.7(b), show that if $E_b/N_0 > 2\ln 2$, then $\lim_{M\to\infty} \Pr(e) = 0$. How close is this to the ultimate Shannon limit on E_b/N_0? What is the limit of the spectral efficiency ρ?

8.10 (Lowerbound to $\Pr(e)$ for orthogonal signals)

(a) Recall the exact expression for error probability for orthogonal signals in WGN from (8.49):

$$\Pr(e) = \int_{-\infty}^{\infty} f_{W_0|A}(w_0|a_0) \Pr\left(\bigcup_{m=1}^{M-1}(W_m \geq w_0|A = a_0)\right) \mathrm{d}w_0.$$

Explain why the events $W_m \geq w_0$ for $1 \leq m \leq M-1$ are iid conditional on $A = a_0$ and $W_0 = w_0$.

(b) Demonstrate the following two relations for any w_0:

$$\Pr\left(\bigcup_{m=1}^{M-1}(W_m \geq w_0|A = a_0)\right) = 1 - [1 - Q(w_0)]^{M-1}$$
$$\geq (M-1)Q(w_0) - \frac{[(M-1)Q(w_0)]^2}{2}.$$

(c) Define γ_1 by $(M-1)Q(\gamma_1) = 1$. Demonstrate the following:

$$\Pr\left(\bigcup_{m=1}^{M-1}(W_m \geq w_0|A = a_0)\right) \geq \begin{cases} \dfrac{(M-1)Q(w_0)}{2} & \text{for} \quad w_0 > \gamma_1; \\ \dfrac{1}{2} & \text{for} \quad w_0 \leq \gamma_1. \end{cases}$$

(d) Show that

$$\Pr(e) \geq \frac{1}{2}Q(\alpha - \gamma_1).$$

(e) Show that $\lim_{M\to\infty} \gamma_1/\gamma = 1$, where $\gamma = \sqrt{2\ln M}$. Use this to compare the lowerbound in part (d) to the upperbounds for cases (1) and (2) in Section 8.5.3. In particular, show that $\Pr(e) \geq 1/4$ for $\gamma_1 > \alpha$ (the case where capacity is exceeded).

(f) Derive a tighter lowerbound on $\Pr(e)$ than part (d) for the case where $\gamma_1 \leq \alpha$. Show that the ratio of the log of your lowerbound and the log of the upperbound in Section 8.5.3 approaches 1 as $M \to \infty$. Note: this is much messier than the bounds above.

8.11 Section 8.3.4 discusses detection for binary complex vectors in WGN by viewing complex n-dimensional vectors as $2n$-dimensional real vectors. Here you will treat the vectors directly as n-dimensional complex vectors. Let $\boldsymbol{Z} = (Z_1, \ldots, Z_n)^\mathsf{T}$ be a vector of complex iid Gaussian rvs with iid real and imaginary parts, each $\mathcal{N}(0, N_0/2)$. The input $\boldsymbol{U}$ is binary antipodal, taking on values $\boldsymbol{a}$ or $-\boldsymbol{a}$. The observation $\boldsymbol{V}$ is $\boldsymbol{U}+\boldsymbol{Z}$.

(a) The probability density of $\boldsymbol{Z}$ is given by

$$f_{\boldsymbol{z}}(\boldsymbol{z}) = \frac{1}{(\pi N_0)^n} \exp \sum_{j=1}^{n} \frac{-|z_j|^2}{N_0} = \frac{1}{(\pi N_0)^n} \exp \frac{-\|\boldsymbol{z}\|^2}{N_0}.$$

Explain what this probability density represents (i.e. probability per unit what?)

(b) Give expressions for $f_{V|U}(v|a)$ and $f_{V|U}(v|-a)$.

(c) Show that the log likelihood ratio for the observation v is given by

$$\mathrm{LLR}(v) = \frac{-\|v-a\|^2 + \|v+a\|^2}{N_0}.$$

(d) Explain why this implies that ML detection is minimum distance detection (defining the distance between two complex vectors as the norm of their difference).

(e) Show that LLR(v) can also be written as $4\Re(\langle v, a\rangle)/N_0$.

(f) The appearance of the real part, $\Re(\langle v, a\rangle)$, in part (e) is surprising. Point out why log likelihood ratios must be real. Also explain why replacing $\Re(\langle v, a\rangle)$ by $|\langle v, a\rangle|$ in the above expression would give a nonsensical result in the ML test.

(g) Does the set of points $\{v : \mathrm{LLR}(v) = 0\}$ form a complex vector space?

8.12 Let D be the function that maps vectors in $\mathcal{C}^n$ into vectors in $\mathcal{R}^{2n}$ by the mapping

$$a = (a_1, a_2, \dots, a_n) \to (\Re a_1, \Re a_2, \dots, \Re a_n, \Im a_1, \Im a_2, \dots, \Im a_n) = D(a).$$

(a) Explain why $a \in \mathcal{C}^n$ and $\mathrm{i}a$ ($\mathrm{i} = \sqrt{-1}$) are contained in the 1D complex subspace of $\mathcal{C}^n$ spanned by a.

(b) Show that $D(a)$ and $D(\mathrm{i}a)$ are orthogonal vectors in $\mathcal{R}^{2n}$.

(c) For $v, a \in \mathcal{C}^n$, the projection of v on a is given by $v_{|a} = (\langle v, a\rangle/\|a\|)\,(a/\|a\|)$. Show that $D(v_{|a})$ is the projection of $D(v)$ onto the subspace of $\mathcal{R}^{2n}$ spanned by $D(a)$ and $D(\mathrm{i}a)$.

(d) Show that $D((\Re(\langle v, a\rangle)/\|a\|)\,(a/\|a\|))$ is the further projection of $D(v)$ onto $D(a)$.

8.13 Consider 4-QAM with the four signal points $u = \pm a \pm \mathrm{i}a$. Assume Gaussian noise with spectral density $N_0/2$ per dimension.

(a) Sketch the signal set and the ML decision regions for the received complex sample value y. Find the exact probability of error (in terms of the Q function) for this signal set using ML detection.

(b) Consider 4-QAM as two 2-PAM systems in parallel. That is, a ML decision is made on $\Re(u)$ from $\Re(v)$ and a decision is made on $\Im(u)$ from $\Im(v)$. Find the error probability (in terms of the Q function) for the ML decision on $\Re(u)$ and similarly for the decision on $\Im(u)$.

(c) Explain the difference between what has been called an error in part (a) and what has been called an error in part (b).

(d) Derive the QAM error probability directly from the PAM error probability.

8.14 Consider two 4-QAM systems with the same 4-QAM constellation:

$$s_0 = 1+\mathrm{i}, \quad s_1 = -1+\mathrm{i}, \quad s_2 = -1-\mathrm{i}, \quad s_3 = 1-\mathrm{i}.$$

For each system, a pair of bits is mapped into a signal, but the two mappings are different:

$$\text{mapping 1:} \quad 00 \to s_0, \quad 01 \to s_1, \quad 10 \to s_2, \quad 11 \to s_3;$$

$$\text{mapping 2:} \quad 00 \to s_0, \quad 01 \to s_1, \quad 11 \to s_2, \quad 10 \to s_3.$$

The bits are independent, and 0s and 1s are equiprobable, so the constellation points are equally likely in both systems. Suppose the signals are decoded by the minimum distance decoding rule and the signal is then mapped back into the two binary digits. Find the error probability (in terms of the Q function) for each bit in each of the two systems.

8.15 Re-state Theorem 8.4.1 for the case of MAP detection. Assume that the inputs $U_1, \dots, U_n$ are independent and each have the a-priori distribution $p_0, \dots, p_{M-1}$. [Hint. Start with (8.43) and (8.44), which are still valid here.]

8.16 The following problem relates to a digital modulation scheme called minimum shift keying (MSK). Let

$$s_0(t) = \begin{cases} \sqrt{\frac{2E}{T}}\cos(2\pi f_0 t) & \text{if } 0 \le t \le T; \\ 0 & \text{otherwise,} \end{cases}$$

and

$$s_1(t) = \begin{cases} \sqrt{\frac{2E}{T}}\cos(2\pi f_1 t) & \text{if } 0 \le t \le T; \\ 0 & \text{otherwise.} \end{cases}$$

(a) Compute the energy of the signals $s_0(t), s_1(t)$. You may assume that $f_0 T \gg 1$ and $f_1 T \gg 1$.

(b) Find conditions on the frequencies f_0, f_1 and the duration T to ensure both that the signals $s_0(t)$ and $s_1(t)$ are orthogonal and that $s_0(0) = s_0(T) = s_1(0) = s_1(T)$. Why do you think a system with these parameters is called minimum shift keying?

(c) Assume that the parameters are chosen as in part (b). Suppose that, under $U = 0$, the signal $s_0(t)$ is transmitted and, under $U = 1$, the signal $s_1(t)$ is transmitted. Assume that the hypotheses are equally likely. Let the observed signal be equal to the sum of the transmitted signal and a white Gaussian process with spectral density $N_0/2$. Find the optimal detector to minimize the probability of error. Draw a block diagram of a possible implementation.

(d) Compute the probability of error of the detector you have found in part (c).

8.17 Consider binary communication to a receiver containing k_0 antennas. The transmitted signal is $\pm a$. Each antenna has its own demodulator, and the received signal after demodulation at antenna k, $1 \le k \le k_0$, is given by

$$V_k = U g_k + Z_k,$$

where U is $+a$ for $U = 0$ and $-a$ for $U = 1$. Also, g_k is the gain of antenna k and $Z_k \sim \mathcal{N}(0, \sigma^2)$ is the noise at antenna k; everything is real and $U, Z_1, Z_2, \ldots, Z_{k_0}$ are independent. In vector notation, $V = Ug + Z$, where $V = (v_1, \ldots, v_{k_0})^{\mathsf{T}}$, etc.

(a) Suppose that the signal at each receiving antenna k is weighted by an arbitrary real number q_k and the signals are combined as $Y = \sum_k V_k q_k = \langle V, q \rangle$. What is the ML detector for U given the observation Y?
(b) What is the probability of error, $\Pr(e)$, for this detector?
(c) Let $\beta = \langle g, q \rangle / \|g\| \|q\|$. Express $\Pr(e)$ in a form where q does not appear except for its effect on β.
(d) Give an intuitive explanation why changing q to cq for some nonzero scalar c does not change $\Pr(e)$.
(e) Minimize $\Pr(e)$ over all choices of q. [Hint. Use part (c).]
(f) Is it possible to reduce $\Pr(e)$ further by doing ML detection on $V_1, \ldots, V_{k_0}$ rather than restricting ourselves to a linear combination of those variables?
(g) Redo part (b) under the assumption that the noise variables have different variances, i.e. $Z_k \sim \mathcal{N}(0, \sigma_k^2)$. As before, $U, Z_1, \ldots, Z_{k_0}$ are independent.
(h) Minimize $\Pr(e)$ in part (g) over all choices of q.

8.18 (a) The Hadamard matrix H_1 has the rows 00 and 01. Viewed as binary codewords, this is rather foolish since the first binary digit is always 0 and thus carries no information at all. Map the symbols 0 and 1 into the signals a and $-a$, respectively, $a > 0$, and plot these two signals on a 2D plane. Explain the purpose of the first bit in terms of generating orthogonal signals.
(b) Assume that the mod-2 sum of each pair of rows of H_b is another row of H_b for any given integer $b \geq 1$. Use this to prove the same result for H_{b+1}. [Hint. Look separately at the mod-2 sum of two rows in the first half of the rows, two rows in the second half, and two rows in different halves.]

8.19 (RM codes)

(a) Verify the following combinatorial identity for $0 < r < m$:

$$\sum_{j=0}^{r} \binom{m}{j} = \sum_{j=0}^{r-1} \binom{m-1}{j} + \sum_{j=0}^{r} \binom{m-1}{j}.$$

[Hint. Note that the first term above is the number of binary m-tuples with r or fewer 1s. Consider separately the number of these that end in 1 and end in 0.]

(b) Use induction on m to show that $k(r, m) = \sum_{j=0}^{r} \binom{m}{j}$. Be careful how you handle $r = 0$ and $r = m$.

8.20 (RM codes) This exercise first shows that $\mathrm{RM}(r, m) \subset \mathrm{RM}(r+1, m)$ for $0 \leq r < m$. It then shows that $d_{\min}(r, m) = 2^{m-r}$.

(a) Show that if $\mathrm{RM}(r-1, m-1) \subset \mathrm{RM}(r, m-1)$ for all r, $0 < r < m$, then

$$\mathrm{RM}(r-1, m) \subset \mathrm{RM}(r, m) \qquad \text{for all } r,\ 0 < r \leq m.$$

Note: be careful about $r = 1$ and $r = m$.

(b) Let $x = (u, u \oplus v)$, where $u \in \mathrm{RM}(r, m-1)$ and $v \in \mathrm{RM}(r-1, m-1)$. Assume that $d_{\mathrm{min}}(r, m-1) \leq 2^{m-1-r}$ and $d_{\mathrm{min}}(r-1, m-1) \leq 2^{m-r}$. Show that if x is nonzero, it has at least 2^{m-r} 1s. [Hint (1). For a linear code, d_{min} is equal to the weight (number of 1s) in the minimum-weight nonzero codeword.] [Hint (2). First consider the case $v = 0$, then the case $u = 0$. Finally use part (a) in considering the case $u \neq 0, v \neq 0$, under the subcases $u = v$ and $u \neq v$.]

(c) Use induction on m to show that $d_{\mathrm{min}} = 2^{m-r}$ for $0 \leq r \leq m$.