ECE 2414 Digital Communications

Lecture Notes

Multimedia University of Kenya

Electrical and Telecommunication Engineering
Semester 1, 2024

Contents

Prerequisites

. MTE 2212 Probability and Statistics
ECE 2326 Analogue Communication Systems

Purpose

To enable the student to learn the different types of digital modulation and demodulation techniques and the sampling process used in digital communication.

Expected Learning Outcomes

At the end of this course, the student should be able to:

1. Describe the different types of digital modulation and demodulation techniques.
2. Analyze the sampling process used in digital communication.
3. Measure noise levels in a digital communication system.

Course Content

Elements of Digital Communication

Model of digital communication systems, Noisy communications channels, the channel capacity of a discrete memoryless channel, Hartley Shanon Law, Bandwidth–S/N tradeoff, pulse Amplitude modulation, PCM, Delta Modulation, adaptive delta modulation.

Sampling Theory

Sampling Theorem, Natural sampling, Flat top sampling, signal recovery & holding, Quantization of signal, Quantization error.

Digital Carrier Modulation Transmission and Reception

Introduction, ASK Spectrum, ASK Modulator, Frequency Shift Keying (FSK), Bandwidth and FSK Spectrum, FSK Detection Using PLL, Binary Phase Shift Keying, Binary PSK Spectrum, Quadrature Phase Shift Keying (QPSK), Differential PSK.

Data Transmission

A baseband signal receiver, probability of error, the optimum filter, white noise-the matched filter, probability of error of the matched filter, coherent reception: correlation, application of coherent reception in PSK and FSK. Correlation receiver for QPSK. Noise in Pulse Code & Delta Modulation Systems.

Introduction to Machine Learning and its Application to Communications Systems.

Laboratory Work

Digital Signaling, Study of Pulse Amplitude Modulation (PAM) and Demodulation, Study of Pulse Width Modulation (PWM) and Demodulation, Study of Pulse Position Modulation (PPM) and Demodulation, To study Pulse Code Modulation (PCM) and demodulation and observe the waveforms, Study of Amplitude Shift Keying (ASK) Modulator and Demodulator, Study of Phase Shift

Keying (PSK) Modulator and Demodulator, Study of Frequency Shift Keying (FSK) Modulator and Demodulator, Noise in Digital Systems and Error Detection.

Mode of Delivery

Lectures, tutorials, group discussions, presentations and practical.

Instructional Materials and/or Equipment

Whiteboards, whiteboard markers, LCD projector, scientific calculator, relevant laboratory equipment and components, and computers and internet.

Course Assessment

- Continuous Course Assessment Tests: 30%
- End of Semester Examination: 70%

Course Textbook

1. John Proakis and Masoud Salehi (2007), *Digital Communications*, McGraw-Hill Science/Engineering/Math, 5th edition, ISBN-13: 978-0072957167.
2. Leon W. Couch (2012), *Digital & Analog Communication Systems*, Prentice-Hall, 8th edition. ISBN-13: 978-0132915380.
3. Lathi B. P. and Zhi Ding (2009), *Modern Digital and Analog Communication Systems*, Oxford University Press, USA; 4th edition, ISBN-13: 978-0195331455.

Reference Textbook

1. Jim Stinson (2007), *Digital Communication & Production*, Goodheart-Willcox Co, 2nd edition. ISBN-13: 978-1590707678.
2. Robert G. Gallager (2008), *Principles of Digital Communication*, Cambridge University Press, illustrated edition. ISBN-13: 9780521879071.
3. Carlson A. Bruce (2009), *Communication Systems*, McGraw-Hill Science/Engineering/Math, 5th edition, ISBN-13: 978-0073380407.

Preface

This set of notes was written to accompany the ETE 2414 Digital Communications Lectures which have been mostly written by Oxford University's Prof Justin Coon for his B13 lectures on Communications. These lectures are primarily based on material covered in the following four textbooks:

1. Principles of Digital Communication, 2nd Edition by R. G. Gallager (2008)
2. Digital Communications, 3rd Edition by I. Glover and P. Grant (2009)
3. Digital Communications: Fundamentals and Applications, 3rd Edition by B. Sklar and F. J. Harris (2021)

4. John Proakis and Masoud Salehi (2007), Digital Communications, McGraw-Hill Science/Engineering/Math, 5th edition, ISBN-13: 978-0072957167.

Most students will find Glover and Grant (2009) and Sklar and Harris (2021) accessible, but perhaps a little disjointed in places. Gallager (2008) is beautifully written, but is intended as a first-year graduate text. Nevertheless, it presents a simple, yet thorough and systematic approach that some students will find useful. As always, the best advice is to consult widely when learning. You may find the description of a concept to be opaque in one text only to discover that it is perfectly clear in another. These notes largely follow the Gallager (2008) structure, but more advanced detail is omitted in favour of clarity, and additional effort is given to explain core concepts and to provide sufficient mathematical details. Some of this detail is included for completeness and to help students to fully understand the theory, but will not be examined. Such material has been annotated by a dagger (†) symbol.

Chapter 1

Introduction to Digital Communications

1.1 Introduction to Digital Communications

In this chapter, we will introduce the fundamental components of digital communication systems along with some basic concepts. Many of these concepts will be explored in depth later in the course, but it is important and instructive to obtain a good overarching picture of modern digital communication theory before diving into the details. First, we address the question of why digital communication is preferred over analogue communication today.

1.1.1 The Argument for Communicating Digitally

Historically, communication between a transmitting device and a receiving device was achieved by modulating a carrier waveform with an analogue [1] message signal. In such a system, the message signal could be conveyed in a discrete time manner, whereby it would remain constant for a period of time before changing to a (possibly) new value in a continuous range. Or, it could vary continuously with time. Such systems still exist, particularly in the form of amplitude modulation (AM) and frequency modulation (FM) transmissions, but these are being increasingly phased out of use for mainstream applications in favour of so-called digital technology.

So, what does digital mean in this context? To understand fully, we must wind back the clock to 1928, when Harry Nyquist published an important paper that provided details about how to achieve distortionless transmission over telegraph lines [2]. One of Nyquist's fundamental conclusions in that work can be summarised as follows. Suppose that a sequence of rectangular pulses, each with its own amplitude and a fixed duration of T seconds, is sent over an ideal telegraph line [3] that admits frequencies ranging from zero to W hertz (Hz). Then, at the output of the telegraph line, each pulse can be resolved unambiguously [4] as long as the rate of pulse transmission, taken to be $1/T$ pulses per second, does not exceed $2W$.

Incidentally, it is Nyquist's theorem on distortionless signalling that yields the well known Nyquist rate of $2W$ samples per second, where a sample in this context corresponds to the amplitude of the pulse. The sampling theorem and the associated rate that is usually taught in engineering classrooms across the world is often mistakenly attributed to Nyquist. In fact, Claude Shannon stated and proved the theorem in a seminal paper published in 1948, but referred to the theorem as being "common knowledge" to those versed in the theory of communication. It is actually thought that V . A. Kotelnikov was the first to prove the sampling theorem in a paper published in 1933[5].

Nyquist showed that rectangular pulse amplitudes that are sent through an ideal channel with bandwidth W can be resolved perfectly at the receiver as long as the pulse rate does not exceed $2W$. Ralph Hartley extended this result in 1928, almost trivially, to state his famous theorem

[1] By "analogue", we mean that a continuum of values can be taken.

[2] Nyquist, H. (1928). Certain topics in telegraph transmission theory. Transactions of the American Institute of Electrical Engineers, 47(2), 617-644.

[3] The telegraph line here can be equally viewed as any ideal communication channel.

on the capacity of an ideal band-limited channel. The logic of Hartley's argument was based on the notion that one could quantify information as the logarithm of the number of possible pulse amplitudes in a given period of time. Hence, if one of M discrete pulse amplitudes could be chosen for each pulse of duration T, then one could convey information at a rate of

$$R = \frac{\log M}{T} \leq 2W \log M \quad \text{bits per second} \tag{1.1}$$

where the inequality follows from Nyquist's theorem on distortionless signalling [6]. The upper bound on R is typically referred to as the capacity of an ideal band-limited channel.

Of course, the problem with Hartley's result was that it was overly simplistic. Real communication systems always suffer from the effects of noise, which results from the random behaviour of electrons in active electronic components. This noise plays havoc with analogue systems, in which the receiver attempts to filter out as much noise as possible in order to obtain a reasonably accurate representation of the transmitted message.

In the decade leading up to 1948, Claude Shannon, inspired by the work of Nyquist and Hartley, completely rephrased the question of how to communicate in noisy systems. Shannon's approach was not to simply remove noise from the received signal, but rather to reconstruct the signal from scratch at the receiver based on noisy observations of the transmitted message. A quick mental test of Shannon's proposed approach of reconstruction reveals its limitations in the sense of analogue modulation. Indeed, to reconstruct the message, one would have to somehow consider the (uncountably) infinite set of possible analogue signals in order to be sure the most likely transmitted message is selected. But, Shannon was interested in discrete signals, in which case it is easy to imagine enumerating all possible messages from a predefined finite set in order to choose the most likely transmitted message. In this context, the message observed at the output of the detector [7] at the receiver is either a perfect replica of the transmitted message, or it is completely different. So, where analogue modulation will always result in a signal that is not quite perfect, but could resemble the transmitted message in some sense, so-called digital modulation is an all-or-nothing game.

But, surely, this digital game would lead to more problems than the analogue approach? Herein lies the incontrovertible argument for communicating digitally. In his seminal paper published in 1948[8], Claude Shannon proved that it is possible to communicate reliably in a noisy system as long as the rate of communication does not exceed a threshold known as the channel capacity. Shannon's theory was mathematical. His proof relied on the notion of choosing a discrete message from a finite set of possible messages at the transmitter. He then proposed the idea of an encoder that mapped the chosen message to a

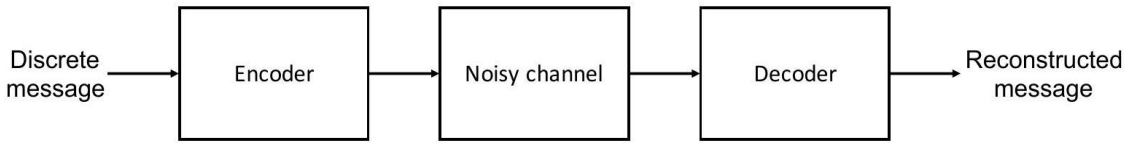


Figure 1: Basic model of communication system envisaged by Shannon.

string of binary digits [9] (bits) called a codeword before conveying the message through a noisy channel to an intended receiver. There, the noisy message is decoded to obtain a replica of the transmitted message (see Figure 1).

Shannon's genius was in the formulation of these system operations. The decoder that Shannon used, for example, is not even practical. The seemingly abstract decoding rule it implements is

[4] If the pulse rate exceeds the $2W$ limit, it turns out that distortion from neighbouring pulses affects each individual pulse, and resolution cannot be guaranteed. We will discuss this in more detail later.

[5] Luke, H. D. (1999). The origins of the sampling theorem. IEEE Communications Magazine, 37(4), 106-108.

[6] The notation log is taken to be base- 2 in this course. We reserve ln for the natural logarithm. All other bases are explicitly noted in the subscript of the log operator, e.g., $\log_{10}$.

[7] The detector is a specific function that a digital communication receiver carries out. We will discuss detection in detail later in the course.

[8] Shannon, C. E. (1948). A mathematical theory of communication. The Bell System Technical Journal, 27(3), 379−423

based on properties of the law of large numbers. No such decoder exists in practice today. However, Shannon's intuition told him that by defining the digital communication system in the way he did, he would be able to obtain the desired result.

Shannon's theorem was not perfect. Indeed, the encoding rule he used was not constructive. In fact, he assumed the encoder randomly mapped each message to a codeword; he did not provide a structured way of performing this mapping. In his theory, this oversight did not matter. He was able to overcome the problem by considering infinitely long codewords, i.e., infinitely long bit sequences were used to represent each message. In this sense, the notion of "reliable communication" referred to above is more formally defined as the event where the probability that a transmitted codeword, corrupted by noise, is decoded in error at the receiver [10] tends to zero as the length of the codeword goes to infinity.

If Shannon's theory was so abstract, then why is it so important? The channel coding theorem described briefly above is, indeed, abstract. However, it can be used to compute the capacity of practical systems. This capacity gives an upper bound on the rate at which communication can take place reliably in practice. For example, one of the most famous formulas in communication theory is

$$C = W \log(1 + \text{SNR}) \tag{1.2}$$

which is the capacity of a band-limited system that experiences additive Gaussian noise at the receiver where the power of the transmission is constrained. The parameter SNR denotes the signal-to-noise ratio [11] at the receiver, and W is the baseband bandwidth of the channel over which information is transmitted. Referring to (1.2), it is interesting, and perhaps a little surprising, to note that capacity increases monotonically (indefinitely) with both bandwidth and SNR, i.e., given enough spectrum and/or transmit power, it is theoretically possible to communicate at arbitrarily high rates with zero probability of decoding error. Shannon derived this equation in his 1948 paper, and engineers continue to use this formula for practical system analysis and design. In fact, it was only recently that code designers began to develop practical encoding and decoding schemes that come close to facilitating reliable communication at rates that approach the capacity.

To summarise, Shannon was able to show, at least theoretically, that it is possible to design digital communication systems that outperform analogue counterparts in terms of received message fidelity. During the past 70 years, engineers and researchers have worked hard to develop practical digital communication systems that come extremely close to performing in the manner that Shannon's theory promises. Analogue systems will always be limited by noise, whereas we now know a great deal about how to design digital systems to cope with such challenges.

1.1.2 Information Sources

All digital communication systems are designed to encode and convey information generated at a source and then reconstruct that information at the receiver. Examples of information sources include collections of text, sound waves generated from speech, still images taken by a camera (film or digital), and moving images captured by a video recording device. These examples illustrate the different forms that source data can take. For example, collections of text are formed from characters drawn from a discrete set, whereas sound waves generated from speech are continuous in nature (although effectively bounded in magnitude and frequency).

In any systematic theory of digital communication, it is important to classify information sources into categories that we can model and work with. The three main types of sources that cover pretty much all practical applications are summarised as follows, and examples of messages generated by these sources are illustrated in Figure 2.

- Discrete sources generate sequences of elements belonging to a finite set. Typically, the set consists of contiguous nonnegative integers, from one to M, for example. Of course, sources such as text also fall into this category. This should not be too confusing, since we can easily map each character of text to a unique integer if we so wish. Discrete sources are, in some

[9] This is where the moniker digital communication comes from.

[10] Here, a decoding error refers to the case where the decoder reconstructs a message from the original set of possible messages that differs from the actual message that was encoded and sent from the transmitter.

[11] We will discuss this parameter in detail later.

ways, the simplest of source types. They are important for two reasons: (1) they model many practical systems of interest, such as systems that convey text or RGB values for each pixel of a discrete image to an intended receiver; and (2), as we will see, analogue sources can be converted to discrete sources through sampling and/or quantisation.

- Analogue sequence sources are similar to discrete sources, except that the elements in the generated sequence can take on a continuum of values. A sampled speech signal is an example of such a source. In this case, the speech samples are confined to the dynamic range of the human voice, but they can take on any value in that continuous interval.
- Analogue waveform sources are generalisations of analogue sequence sources. An analogue waveform source generates a waveform, which could be time-limited or frequency-limited (but not both), from a set of possible waveforms (which could be infinite). Information is encoded in the particular waveform that is output from the source. Note that it is possible for an analogue waveform source to generate a sequence of waveforms, with information encoded separately in each. We will revisit this example later.

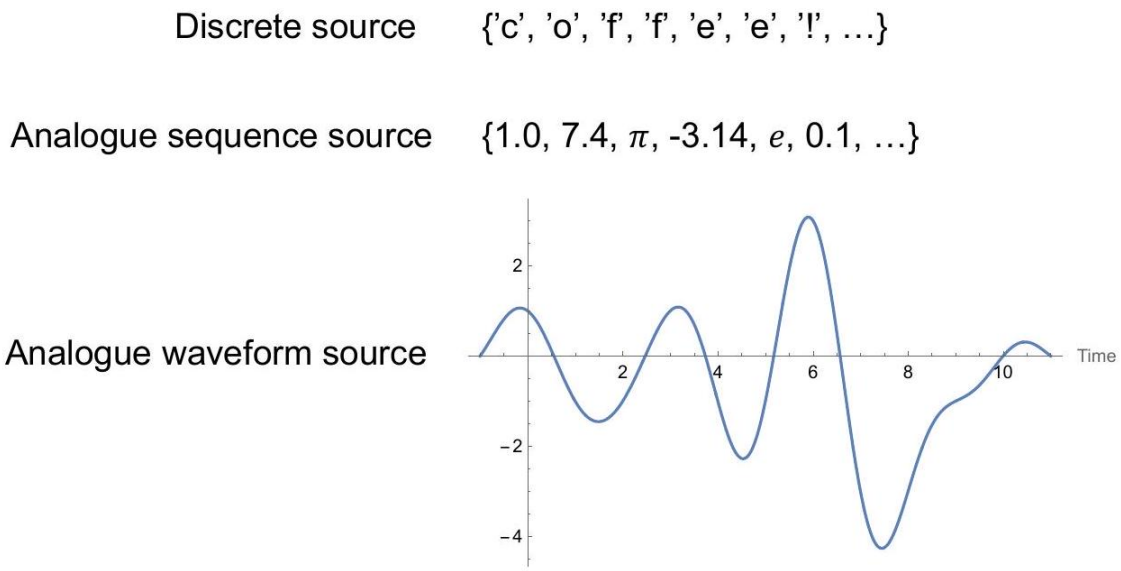


Figure 2: Examples of messages generated by different sources.

1.1.3 Digital Representation of Information

As discussed in Section 1.1, Shannon's theory relies on the idea that information source messages are encoded to yield binary strings [12]. For now, we refrain from discussing the exact method of encoding, but instead focus on the general idea of representing a source message digitally.

In the case of discrete sources, the mapping from source symbols [13] to bit strings is relatively straightforward. Suppose a discrete source generates an M ary symbol every T seconds, i.e., each symbol is drawn from a set consisting of M possible symbols, labelled 1 to M. We will refer to the time devoted to each such symbol is a symbol period [14]. If M is a power of two, then we can employ a generalisation of Hartley's logic to represent each source symbol as a unique string of $\log M$ bits.

For analogue sequence sources, the mapping to bit strings requires an additional step. Each of the source symbols in this case is drawn from an uncountably infinite set. Hence, we would only be able to describe such symbols as binary strings if we use an infinite number of bits per symbol. Instead, we must first quantise each source symbol, which effectively involves mapping all values

[12] The term "string" is used to denote a finite collection of entities. In contrast, the term "sequence" typically refers to a countably infinite collection.
[13] The term "symbol" refers to a discrete number, character, or similar entity that represents information to be conveyed to an intended receiver.
[14] Gallager (2008) uses the term epoch to denote a symbol period.

in the uncountably infinite set covered by the source to a finite set of discrete values. These values will typically be real numbers rather than natural numbers, as is the case for discrete sources. Hence, it is often helpful to apply an additional step whereby these real-valued quantised symbols are mapped to a set of natural numbers, say 1 to M. In doing so, the process of quantisation (and subsequent mapping) converts an analogue sequence source to a discrete source, and mapping each discrete quantised symbol to a string of bits follows easily.

Analogue waveform sources are a little more complicated to deal with than the other two types of sources. It is, perhaps, easiest to think about this case through the example of speech waveforms. Clearly, an infinite number of speech waveforms of a given length of time exists. To encode such a waveform as a binary string, we must first sample the waveform, which yields a discrete set of analogue values. In other words, sampling converts the analogue waveform source to an analogue sequence source. We can then quantise the samples and map the resulting symbols to bit strings.

It is worth noting that analogue waveform sources can be thought of in a different manner. Specifically, if we consider a band-limited waveform (instead of time-limited as noted above), then the sampling theorem [15] guarantees that we can represent the waveform perfectly (in theory) by using a discrete set of samples equal in number to twice the bandwidth. Consequently, the sampling operation is information lossless, and thus the analogue waveform source is not only converted to an analogue sequence source, but the two are, in fact, equivalent.

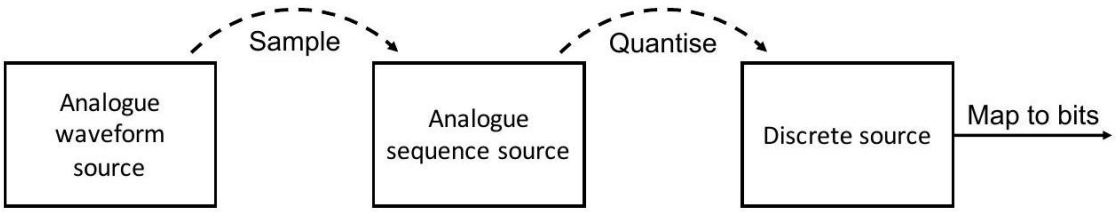


Figure 3: Depiction of the relationship between different source types in the context of the source-to-bit mapping operation.

1.1.4 Channels and Modulation

As discussed in the previous section, source encoders map source messages to bit strings. In any digital communication system, the aim is to convey these bit strings (or, technically, the information they represent) to an intended receiving device. Transmission is inherently a physical process, and thus it is desirable to map the bits generated by the source encoder to a physical waveform (or waveforms) that can be physically measured at the receiver. Such waveforms may describe fluctuations in voltage observed on a transmission line connecting the transmitter and the receiver, or they may represent electromagnetic waves that travel from a transmitting antenna to a receiving antenna in a wireless system. Other examples exist. The common feature in all examples is the idea that one must encode the bits generated through source encoding onto analogue waveforms.

In the case where the source, itself, is an analogue waveform, it may seem odd to first sample, quantise, and encode each source waveform only to go back to the analogue domain in order to transmit the message. However, the importance of following this process can be observed when we consider the properties of the channel over which the message will be conveyed. When we speak of a communication channel, we are typically referring to the physical medium between the transmitting and receiving devices as well as (in some cases) the front-end circuitry at each device, such as transmit/receive filters and amplifiers. Even though this definition of a channel is somewhat ambiguous, several very important, general models of a communication channel exist. These include the additive Gaussian noise (AGN) channel, the linear time-invariant (LTI) channel, and the linear time-varying channel.

The AGN channel is the simplest of the basic models. As noted earlier, received waveforms are corrupted by noise in all communication systems. This noise limits the capacity of the system, and engineers have developed a number of ways to reproduce messages reasonably reliably at a

[15] We will discuss the sampling theorem in more detail later in the course.

receiver, even when the noise level is significant. The random quantum effects that give rise to noise at the receiver are statistically independent and abundant. Adding these effects together yields a macroscopic change in, for example, the observed voltage level of a received waveform. These changes can often be modelled statistically as a zero-mean Gaussian process [16]. If a transmitted waveform is represented by the function $x(t)$, then the received waveform $V(t)$ can be modelled as

$$V(t) = x(t) + Z(t)$$

where $Z(t)$ denotes the AGN process [17].

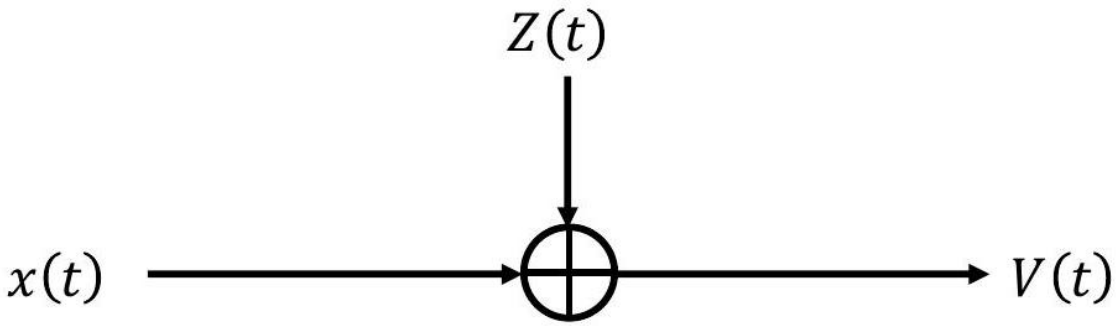


Figure 4: The additive Gaussian noise channel model.

Most communication channels exhibit filtering properties, whereby the spectrum of the input to the channel (i.e., the transmitted waveform) is shaped in some way. The LTI system model accurately represents these scenarios. This model is characterised by the channel output (i.e., the received waveform) being represented by the convolution of the channel input $x(t)$ and the channel impulse response (CIR) $h(t)$. In addition, these convolved waveforms are affected by AGN at the receiver. The resulting received waveform is modelled

by the equation [18]

$$V(t) = (h * x)(t) + Z(t) = \int_{-\infty}^{\infty} h(\tau)x(t-\tau)\mathrm{d}\tau + Z(t)$$

Note that the LTI model captures temporal dispersion that some channels may introduce. Wireless channels are typical examples where this phenomenon is often observed, owing to the idea that a waveform broadcast into the wireless medium will be reflected and refracted many times, with different components of the wave travelling along different paths from the transmitting antenna to the receiving antenna. Thus, from the receiver's perspective, the transmitted waveform will be spread out in time.

The third important model is the linear time-varying model. This model is more advanced and is often used to represent quickly changing mobile communication systems. The mathematics resemble the LTI case, but where the CIR changes with the time variable t, giving

$$V(t) = \int_{-\infty}^{\infty} h(t,\tau)x(t-\tau)\mathrm{d}\tau + Z(t)$$

Now, let us return to the idea of representing bit strings as physical waveforms in preparation for transmission over a noisy channel. This encoding operation is often referred to as modulation. The terminology derives from early analogue communication methods, whereby a message waveform was combined with a carrier waveform prior to transmission. Technically, the term "modulation" can be used to refer to the variation in pitch, volume, or tone of one's voice. Or, more generally, it relates to the act of controlling or influencing something. In conventional analogue modulation schemes, the carrier waveform is controlled, or modified, by the message waveform. In this sense,

[16] The addition of many independent microscopic changes ensures that the central limit theorem can be invoked, which yields the Gaussian result.

[17] Here, we use what will become standard notation: lower-case letters are reserved for deterministic quantities; upper-case letters represent random quantities. Note that even though the transmitted waveform is deterministic in this case, the received waveform is random owing to $Z(t)$.

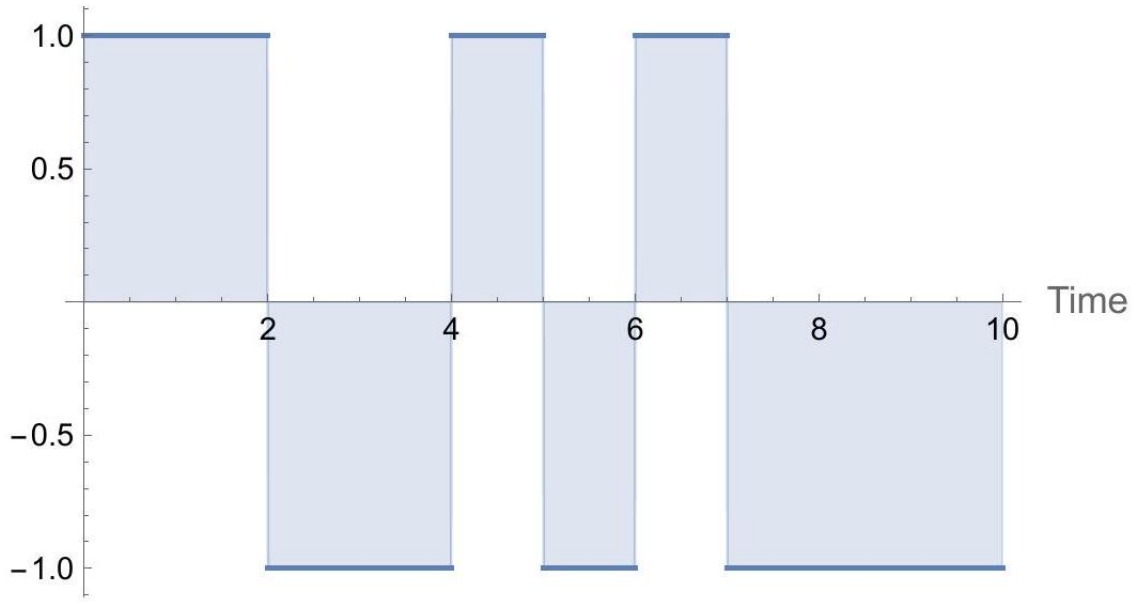
the message waveform modulates the carrier. In the digital case, the same idea applies for passband transmissions; but, in general, the principle of modulation refers to any waveform being controlled, or defined, by a message signal.

In fact, it is probably more accurate to refer to digital modulation as digital encoding and do away with the term "modulation" altogether. Gallager (2008) presents more detailed arguments for this change in terminology. But, old habits die hard, and we will often use the conventional phrase "digital modulation" throughout this course.

The primary concern when choosing a digital modulation strategy is to ensure the effects of channel noise are ameliorated as much as possible. Without having to dig too deeply, it is fairly easy to deduce that this criterion implies the modulated waveforms should be as different as possible. Indeed, if two waveforms representing different bit strings only differ slightly, then noise could easily corrupt one waveform to the point of rendering it closer in structure to the other, which would lead to a decoding error at the receiver.

A simple example of the idea of separating waveforms can be seen in binary pulse amplitude modulation (PAM). In this scheme, a single bit is encoded in a (rectangular) pulse of duration T seconds with unit magnitude. Note that the rate of this transmission would be $1/T$ bits per second. A typical mapping for this binary scheme is as follows: a one is encoded in a positive pulse, and a zero is encoded in a negative pulse (Figure 5). We could, instead, choose the zero to be encoded in the absence of a pulse, for example - a technique known as on-off keying (OOK). However, it seems reasonable that the antipodal scheme would perform better in a noisy channel, since the distance between the positive and negative pulses is greater than the distance between the positive and zero pulses [19]. We will more carefully define the term "distance" as it relates to waveforms and signals a little later.

The modulation process described here is known as baseband modulation. In many systems, such as wireless networks, the baseband waveform is modulated to passband. This process is known as frequency conversion. As we are altering a signal at a lower set of frequencies to occupy a higher set of frequencies, we give this translation the specific title of up-conversion. In any case, the modulated signal is transmitted through the channel, and



Encoded string: 121

Figure 5: Example of a binary string encoded using binary antipodal PAM.

the inverse processes corresponding to those operations described so far are performed in reverse

[18] Here, we use the notation $(x * h)(t)$ to denote the convolution of x and h as a function of the independent variable t. Some texts use the looser notation $x(t) * h(t)$, or similar.

[19] Of course, we could choose an OOK scheme where the positive pulse has amplitude equal to two units, in which case there would be no difference in performance between the antipodal and OOK schemes.

order to recover the transmitted message. For example, a received passband waveform is down-converted and demodulated at baseband, and an estimate of the transmitted message is reconstructed. A simple diagram illustrating the sequential processes of source encoding, modulation, frequency conversion, and their inverses is shown in Figure 6. More detailed system diagrams such as this can be found in Sklar and Harris (2021) and Glover and Grant (2009).

1.1.5 Forward Error Correction

Shannon's famous noisy channel coding theorem promises that reliable communication can be achieved in theory if the rate of communication does not exceed the capacity of the channel. For the AGN channel, this capacity is given by the celebrated equation

$$C = W \log(1 + \text{SNR}) \quad \text{bits per second}$$

as discussed earlier. We also stated above that Shannon's result relies on the idea of encoding source messages such that they are represented by very long codewords. We have briefly covered the idea of source encoding, but we did

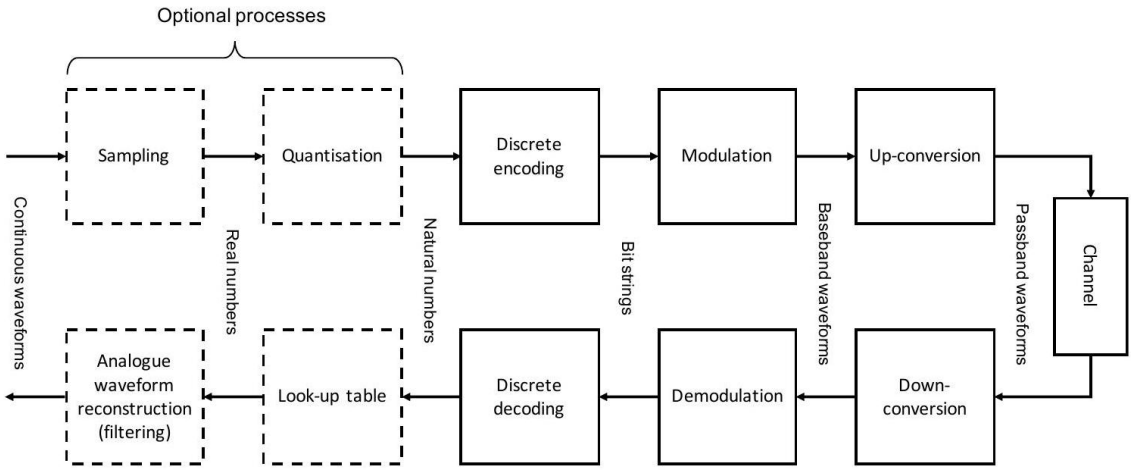


Figure 6: Illustration of basic processes related to source encoding/decoding, modulation, and frequency conversion. The look-up table process is designed to map the discrete sequence of natural numbers output by the discrete decoder back to real values. In the case where an analogue waveform needs to be constructed from these values, sampling theory is invoked, which is technically carried out with a linear filtering operation.

not specifically make reference to long codeword lengths. In fact, we have not linked source encoding to Shannon's theorem at all.

In practice, source encoding is typically thought of as a separate process to the encoding operation that is required to invoke Shannon's promised result, although this is not strictly necessary. It is simply the process of getting information into a binary form. Of course, there are good ways and bad ways to do this, but we will not concern ourselves too much with such matters in this course [20]. The optional additional step that follows source encoding, which is needed to achieve the capacity bound, is the process of forward error correction (FEC) encoding.

FEC encoding - also referred to as error correction coding or error control coding - effectively involves taking strings of bits at the output of the source encoder and mapping these to bit strings of longer length. Many such encoding processes operate on strings of a given length, say k, and output strings of another constant length, say $n > k$. These block codes are said to

have a code rate of $r = k/n$.

Note that there are 2^k possible input strings in a block coding scheme, which is typically a much smaller number than the number of possible length- n codewords. Hence, there are many different ways to map the set of input strings to a set of codewords [21]. Some mappings will be better than

[20] Source encoding is a fascinating and active topic of study and research. To do it justice, one would need to develop a number of tools that underpin the field of information theory, which goes beyond the scope of this course.

others. The idea is that the FEC encoder should be designed to maximise the distance between all pairs of codewords [22]. By doing so, the message is made to be somewhat robust to errors introduced in the channel.

To understand why this is the case, consider the scenario where all codewords differ by at least two bits. If the noise in the channel causes one bit in a codeword to be flipped (e.g., from zero to one), then a properly designed decoder should be able to recognise that the resulting received bit string is not a valid codeword. In this instance, the decoder declares that an error has been introduced, and it either attempts to correct the error, or a request for a retransmission of that codeword is conveyed back to the transmitter.

Figure 7 illustrates an example of this scenario. Referring to the figure, the blue codeword is transmitted over the noisy channel in this example. The channel corrupts the third bit. The decoder is able to recognise that an error has occurred, but it cannot correct the error in this case. It does not know which bit has been flipped. All it can do is to consider all possible codewords and decide on the most likely transmitted codeword given its observation. However, without further information, it determines that both the second and fourth codewords were transmitted with equal probability, since the received bit string differs from each by only one bit; hence, both of these codewords satisfy the "most likely transmitted codeword" condition.

The field of coding theory is rich and active. Applications extend well beyond conventional communication systems. For example, new FEC methods are constantly being developed for solid-state storage [23]. The subject of

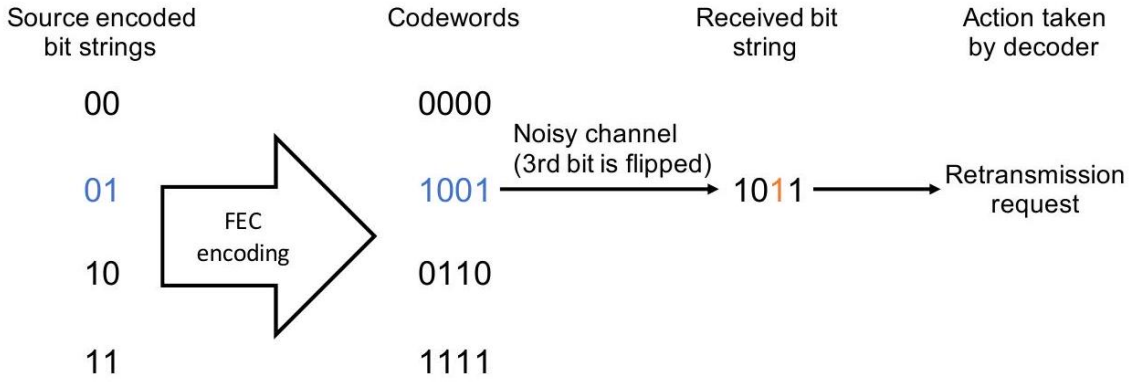


Figure 7: Illustration of FEC encoding and decoding operations.

FEC encoding and decoding is vast and would require a full course devoted to its study to do it justice. Hence, we will not cover further details here. Chapter 8 of Gallager (2008) gives a nice general perspective, although it is a little advanced in its presentation. Chapter 10 in Glover and Grant (2009) and Chapters 6, 7, and 8 in Sklar and Harris (2021) also provide some further information. The former is a shorter introduction (but with far more information than I have provided here), whereas the latter is a bit more complete and gives more details of techniques used in practice today, such as turbo codes and low-density parity check (LDPC) codes.

1.1.6 The Layering Principle and Standardised Interfaces

In Figure 6, we showed the main digital transmission and reception processes we will discuss in this course. We also highlighted the symmetry in these processes and showed how the signal format changes as it progresses inwardly toward the channel and then back outwardly at the receiver. This basic notion of progressing through different stages (and their inverses) in a digital system is known as layering. The layering principle is useful in creating standardised interfaces, whereby

[21] In fact, there are $\binom{2^n}{2^k}$ such ways, where $\binom{a}{b} = \frac{a!}{(a-b)!b!}$ is the binomial coefficient.

[22] More specifically, most encoders are designed to maximise the minimum distance, where the distance between two codewords is defined as the number of places that the bit strings differ.

[23] Incidentally, although storage does not appear to be related to communication theory, it happens to involve encoding information digitally, conveying this to a storage medium where random physical processes

blocks of a system can be designed (and replaced) in a modular format without affecting other blocks. You may recognise this concept in other areas of engineering as well.

So far, we have discussed layering at a low level. For example, for an ana-

DLC header	IP header	TCP header	Data packet	DLC trailer

Figure 8: Simplified example of a packet generated by a layered protocol for transmission of source encoded data over the Internet.

logue waveform source, we have described the sampling layer, the quantisation layer, the discrete encoding layer, etc. In fact, most of these processes are referred to collectively as the physical layer, or PHY layer. The exception is the source encoder and decoder, which often sit outside of the PHY layer.

The PHY layer's job is to process any source encoded message in preparation for transmission over a physical channel. Of course, we must think about other aspects of the communication system, such as how and where to route the message in a multi-user network. Hence, several other (higher) layers exist in modern communication systems.

Let us take the example of the Internet. We will adopt the simplified example described in Section 1.4 of Gallager (2008). In Internet transmissions, the source encoder sits at the highest layer (furthest away from the PHY layer). The bit strings output by the source encoder are arranged into packets, which just contain the data bits that describe the message that we wish to transmit. A transmission control protocol (TCP) header, which is another bit string that contains information about the source and destination ports, is appended to the beginning of each packet. Following this, an internet protocol (IP) header is added to facilitate the routing of the message to the correct host device. Finally, a data link control (DLC) header is added, and a DLC trailer is appended to the end of the packet. The DLC bit strings facilitate the communication between two end-user devices in a network and often contain additional FEC bits. An illustration of this packet structure is given in Figure 8.

Moving from the TCP/IP to DLC processes, we travel down the layers towards the PHY layer. The DLC header/trailer processing is done in the socalled Data Link Layer (DLL), which is typically considered to be the second layer in the Open Systems Interconnection (OSI) model of computing and telecommunications. The PHY layer is considered to be the first layer.

deteriorate the stored data over time, and ultimately read (received) at a future date. In other words, storage = communication!

Chapter 2

Mathematical Preliminaries

2.1 Fourier Series and Transforms

2.2 Probability and Random Processes

We begin with some of the basic concepts that underpin the subject of probability theory. A systematic study would start with a discussion of measure theory, but we will mostly forego such a discussion in favour of building quickly to operational aspects of probability that are typically more useful for engineers. Nevertheless, we must at least introduce the setting in which our study of probability takes place: the probability space.

2.2.1 Probability Spaces

The setting of probability is the triple $(\Omega, \mathcal{A}, \mathrm{P})$. This is sometimes called the probability space. Here, Ω represents a sample space that contains the possible outcomes that may arise from a simple experiment. The second quantity $\mathcal{A}$ is a collection of subsets of Ω. Each subset is called an event. Formally, $\mathcal{A}$ must be a σ-algebra. The properties of σ-algebras are beyond the scope of these notes, and we will generally not concern ourselves with such technical details here. The third quantity is a probability measure. As we will see, P serves to measure the probability that a given event in $\mathcal{A}$ occurs.

Consider the following illuminating example. Suppose we have a weighted coin, which, upon flipping, comes up heads $100 \times p$ percent of the time, i.e., with probability p. In this case, the sample space is $\Omega = \{$ heads, tails $\}$. The associated σ-algebra is

$$\mathcal{A} = \{\emptyset, \{ \text{ heads } \}, \{ \text{ tails } \}, \{ \text{ heads , tails } \}\}$$

We see that $\mathcal{A}$ is the collection of all possible subsets of Ω in this case, which includes the empty set $\emptyset$ and Ω, itself. This is called the power set and is sometimes represented as 2^{Ω}. Now, if we construct a simple experiment whereby we flip the coin once, we can mathematically represent the probabilities of the events as follows:

$$\begin{array}{ll} \mathrm{P}[\emptyset] = 0, & \mathrm{P}[\{ \text{ heads } \}] = p \\ \mathrm{P}[\{ \text{ tails } \}] = 1 - p, & \mathrm{P}[\{ \text{ heads , tails } \}] = 1 \end{array}$$

The first probability $P[\emptyset]$ is read as "the probability that neither heads nor tails occurs". The final probability $\mathrm{P}[\{$ heads, tails $\}]$ is read as "the probability that either a heads or a tails occurs". For this particular experiment, we flip the coin once; hence, we must have either a heads or a tails. This is why $\mathrm{P}[\emptyset] = 0$ and $\mathrm{P}[\{$ heads, tails $\}] = 1$. We said that, by definition, the probability of a heads appearing is p. Hence, we must have that the probability that a tails appears is $1 - p$.

Note that, in general, $\mathrm{P}[\Omega] = 1$. Furthermore, for any $A \in \mathcal{A}$, we have that $\mathrm{P}[A] = 1 - \mathrm{P}\left[A^c\right]$, where $A^c = \Omega \backslash A$ is the complement of A. These statements are true for any probability space. Also note that P somehow measures the likelihood that an event (the argument of P) occurs in a given experiment. We can use this formalism to construct more complicated experiments. For

example, we can flip the coin two times and predict the likelihood that a heads will turn up on both occasions. If both coin flips are independent of each other (more on independence later), we can write such a probability as

$$\mathrm{P}[\{\text{ heads }\},\{\text{ heads }\}] = p^2$$

Note that the set braces separate the two events in our more complicated experiment. The probabilities multiply in this case, since for the two events to occur in a collective manner, we must have that each occurs separately. Probability expressions such as these are read "the probability that {event one } and {event two } occur". We will not consider such extensions further at this point, but will press ahead with some of the basic mechanics of probability in a more general setting.

2.3 Problem 1.1: Coin toss

Consider the probability space defined above for the coin toss experiment. For each probability expression below, write the number of coin tosses that are implied. Also, write expressions for these probabilities in terms of the probability p that a heads occurs.

$\mathrm{P}[\{\text{ tails }\}]$,	$\mathrm{P}[\{\text{ heads }\},\{\text{ tails }\}]$,	$\mathrm{P}[\{\text{ tails }\},\{\text{ tails }\},\{\text{ heads }\}]$,
$\mathrm{P}[\emptyset,\{\text{ tails }\}]$,	$\mathrm{P}[\Omega,\emptyset]$,	$\mathrm{P}[\{\text{ heads }\},\Omega]$.

2.3.1 Conditional Probability

Let $\{A_1,\ldots,A_n\}$ be a set of mutually exclusive events, and let B be an arbitrary event with nonzero probability. (See Figure 1.1.) The conditional probability of A_i given the occurrence of B can be written as

$$\mathrm{P}\left[A_i \mid B\right] = \frac{\mathrm{P}\left[A_i, B\right]}{\mathrm{P}[B]} \tag{1.1}$$

When we measure the conditional probability, we are actually measuring the likelihood that a particular event occurs preconditioned on the knowledge that another event is true. Similarly, we have

$$\mathrm{P}\left[B \mid A_i\right] = \frac{\mathrm{P}\left[A_i, B\right]}{\mathrm{P}\left[A_i\right]} \tag{1.2}$$

Now, if the union of all A_i events yields the set Ω, i.e., $\bigcup_{i=1}^{n} A_i = \Omega$, then we can write

$$\begin{aligned}\mathrm{P}\left[A_i \mid B\right] &= \frac{\mathrm{P}\left[A_i, B\right]}{\mathrm{P}[B]} \\ &= \frac{\mathrm{P}\left[A_i, B\right]}{\sum_{j=1}^{n} \mathrm{P}\left[B, A_j\right]} \\ &= \frac{\mathrm{P}\left[A_i, B\right]}{\sum_{j=1}^{n} \mathrm{P}\left[B \mid A_j\right] \mathrm{P}\left[A_j\right]}\end{aligned} \tag{1.3}$$

To see that the second line is true, recall that $\mathrm{P}[\Omega] = 1$. This is known as the total probability theorem. It frequently arises in engineering applications.

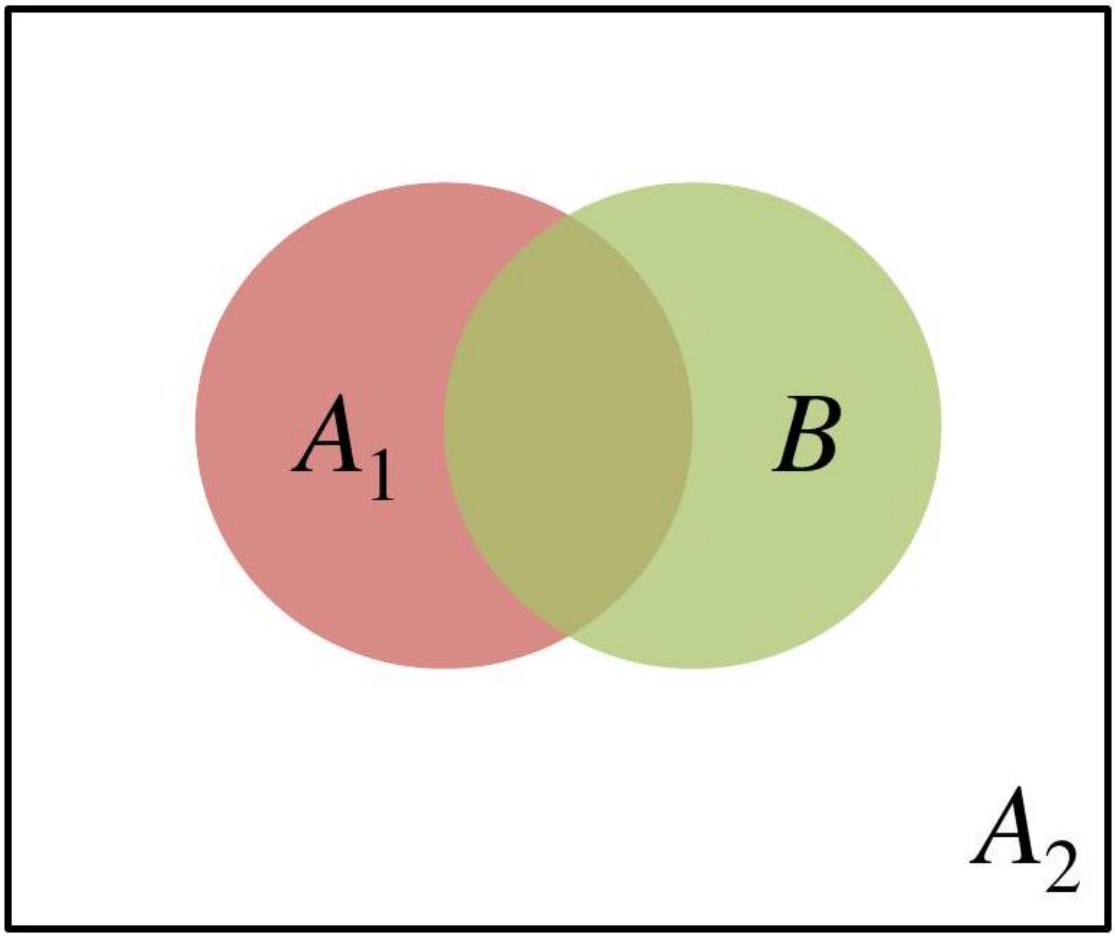


Figure 1.1: Sample space $\Omega = A_1 \cup A_2$. Events A_1 and A_2 are mutually exclusive.

2.4 Problem 1.2: Conditional Die Roll

Consider a six-sided, uniformly weighted die, where the faces hold the natural numbers 1 to 6 . Suppose we conduct the simple experiment that involves rolling the die once and recording the result.
(i) What is the sample space Ω ?
(ii) What is the σ-algebra $\mathcal{A}$?
(iii) What are the probabilities corresponding to the occurence of each number (face)?
(iv) What is the probability that a 2 is rolled given that the roll is even?

2.4.1 Statistical Independence

Two events A and B are statistically independent if and only if

$$\mathrm{P}[A \mid B] = \mathrm{P}[A] \tag{1.4}$$

That is, knowledge of whether or not B has occurred has no bearing on the probability that A will occur. Applying the rules of conditional probability, we see that, under statistical independence, we can write

$$\mathrm{P}[A, B] = \mathrm{P}[A]\mathrm{P}[B] \tag{1.5}$$

2.5 Problem 1.3: Independent bits?

Suppose we have a machine that generates three bits sequentially. Call these X_1, X_2, and X_3. The probability that $X_i = 1$ is $1/2$ for $i = 1, 2.X_3$, which is the third bit generated, is 1 if the sum of the first two bits is odd and $X_3 = 0$ if the sum of the first two bits is even. Note that the first two bits are generated independently.
(i) What is the probability that $X_2 = 0$?
(ii) What is the probability that the first two bits are both 1 ?
(iii) What is the probability that the three-bit string generated by the machine contains exactly two ones?
(iv) What is the probability that the three-bit string generated by the machine is 000 ?
(v) Are the first and third bits in the sequence statistically independent? Explain.

2.5.1 Random Variables

It is rather cumbersome to work directly with outcomes and events. To develop a calculus that can be used to study more complicated experiments, we often require a mapping from the sample space Ω to a set of real or complex numbers. Such a map is called a random variable. The difference between a random variable and a deterministic variable is that we have no way of knowing what the former is a priori [1], yet the latter can be assigned whatever value we wish.

Consider the coin toss experiment. Let the map $X : \Omega \to \{0, 1\}$ be defined as follows:

$$X = 1, \quad \text{if the coin toss results in a heads,}$$
$$X = 0, \text{ if the coin toss results in a tails.}$$

Then for any $\omega \in \Omega, X(\omega)$ is a mapping of the observed outcome ω to either 1 or 0 . So X is a random variable that describes the coin toss experiment. Instead of concerning ourselves with the probability of an event (a heads or tails in this case), we can equivalently work with the probability that $X = x$, where $x \in \{0, 1\}$. We usually write such statements as, for example, $\mathrm{P}[X = 1]$. This may seem like a slight abuse of notation, since, technically, the probability measure P operates on Ω, not the space to which X maps values in Ω, i.e., the image of Ω. Suffice to say that a systematic development of probability theory exists whereby the operation $\mathrm{P}[X = 1]$ is well defined and equivalent to $\mathrm{P}[\{$ heads $\}]$ in the case of our coin toss experiment. We will say no more about the underlying theory here; the interested student is referred to the text by A. Klenke for complete details.

At this point, it is important to make a comment on notation. In most contexts, uppercase letters (e.g., X) are used to represent random variables, while the corresponding lowercase letters (e.g., x) denote outcomes. This convention usually makes it clear when one is talking about a random quantity and when a deterministic quantity is the subject of investigation.

For a real random variable X, the cumulative distribution function (CDF) is defined as

$$F_X(x) = \mathrm{P}[X \leq x], \quad x \in X(\Omega) \tag{1.6}$$

where $X(\Omega)$ is the support [2] of the random variable X. Note that the subscript " X " is used in the functional notation above to discern a CDF related to the random variable X from one related to another random variable. So the CDF of a random variable Y would be written as $F_Y(x) = \mathrm{P}[Y \leq x]$. It is important to recognise that for a random variable X, we have that

$$\mathrm{P}[X \leq x] = 1 - \mathrm{P}[X > x]$$

This is analogous to the earlier statement that we made concerning events:

$$\mathrm{P}[A] = 1 - \mathrm{P}\left[A^c\right]$$

[1] A priori - Literally, "from the earlier"; here, used to signify knowledge that is present before conducting and experiment or making an observation.

If X is a continuous, real random variable, it often has associated with it a probability density function (PDF). When $X(\Omega) = (-\infty, \infty)$, the relationship between the CDF and PDF $p_X(x)$ of X is given by

$$p_X(x) = \frac{\partial F_X(x)}{\partial x} \quad \Leftrightarrow \quad F_X(x) = \int_{-\infty}^{x} p_X(u)\mathrm{d}u \tag{1.7}$$

The PDF and CDF definitions given above work for discrete random variables as well. For example, we can write

$$p_X(x) = \sum_{i=1}^{n} \mathrm{P}\left[X = x_i\right] \delta\left(x - x_i\right) \tag{1.8}$$

in which case the CDF is given by

$$F_X(x) = \sum_{i=1}^{x} \mathrm{P}\left[X = x_i\right] \tag{1.9}$$

where the sum runs to the number x_i closest to, but not greater than x. In the discrete case, all of the relevant information about X is contained in the probabilities, and the

PDF is often not used directly. Often, we will write that X has distribution $\{p_1, \ldots, p_n\}$, where $p_i := \mathrm{P}\left[X = x_i\right]$, when we want to define the random variable.

Note that for the examples given above

$$\lim_{x \to -\infty} F_X(x) = 0 \tag{1.10}$$

and

$$\lim_{x \to \infty} F_X(x) = 1 \tag{1.11}$$

Also note that for a continuous random variable, $\mathrm{P}[X = x] = 0$ since a single point has zero probability (i.e., zero measure). Consequently, $\mathrm{P}[X < x] = \mathrm{P}[X \leq x] = F_X(x)$.

2.6 Problem 1.4: Weighted die

Suppose we have an eight-sided, weighted die, the sides of which show the numbers $\{4, 4, 8, 8, 16, 16, 16, 16\}$. In a single die roll experiment, we can define the random variable X to be the number that comes up when we roll the die once. Hence, X is real and discrete. The die is weighted such that the probability that a roll shows a particular side is equal to the reciprocal of the number shown on the side rolled.

(i) What is the support of X ?

(ii) What is the probability that a 16 is rolled, i.e., $\mathrm{P}[X = 16]$?

(iii) What is the probability that the roll is less than or equal to 8 ?

(iv) What is the distribution of the side that comes up when we roll the die once? (NB: We are not talking about the number that appears on the side in this case, but the side, itself.)

2.6.1 Multiple Random Variables

Multiple random variables in a system should not, initially, be treated separately. They have a joint distribution, more specifically a joint CDF and a joint PDF. We have already touched on this notion in the context of the multiple coin toss experiment discussed in section 1.1. For discrete random variables, the idea is relatively straightforward, as evidenced by the following example problem, which you should work through.

[2] Informally, the support of a random variable is the set of possible values it can take.

2.7 Problem 1.5: Multiple coin tosses

Suppose we conduct an experiment whereby we toss a coin n times. Each toss is independent of others. The probability that a heads appears is p. We let the random variable $X_i \in \{0, 1\}$ signify the result of the i th toss, where $X_i = 1$ if a heads comes up. The probability of a given sequence of tosses $(x_1, \ldots, x_n) \in \{0, 1\}^n$ is written as $\mathrm{P}[x_1, \ldots, x_n]$.
(i) What is the probability that the first $n - 1$ tosses come up heads and the final toss comes up tails?
(ii) Show that

$$\mathrm{P}[x_1, \ldots, x_n] = p^{\sum_i x_i}(1-p)^{n-\sum_i x_i}$$

(iii) What is the probability that an even number of heads appears?

For two real random variables X_1 and X_2, these are defined by the relations

$$F_{X_1,X_2}(x_1, x_2) = \int_{-\infty}^{x_1}\int_{-\infty}^{x_2} p_{X_1,X_2}(u, v)\mathrm{d}v\ \mathrm{d}u \tag{1.12}$$

and

$$p_{X_1,X_2}(x_1, x_2) = \frac{\partial^2 F_{X_1,X_2}(x_1, x_2)}{\partial x_1 \partial x_2} \tag{1.13}$$

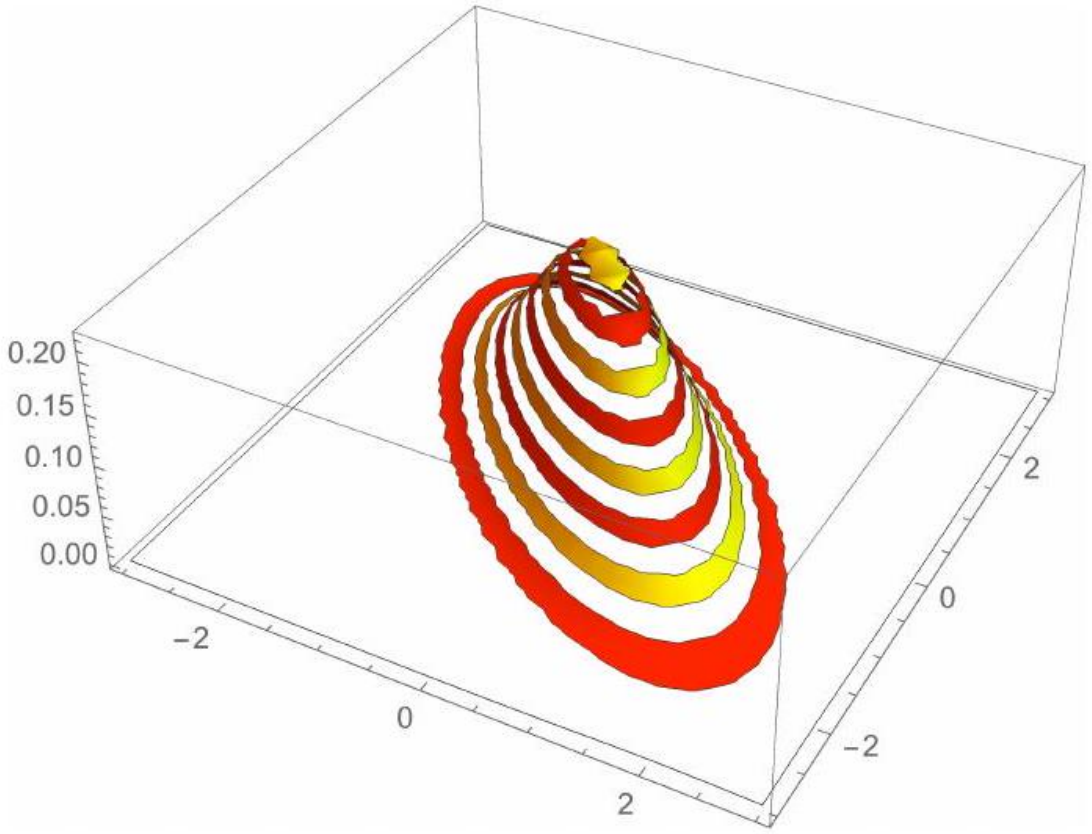
respectively. Figure 1.2 provides an illustration of these functions for the specific case of jointly normal random variables. We will write more about the normal distribution later. The definitions extend naturally to more random variables. When writing joint PDFs and CDFs, the subscript notation can become cumbersome, so it is sometimes omitted in favour of simplicity. Note that we can marginalise a distribution by integrating the PDF over one (or more) of the variables. For example,

$$p(x_1) = \int_{-\infty}^{\infty} p(x_1, x_2)\,\mathrm{d}x_2 \tag{1.14}$$

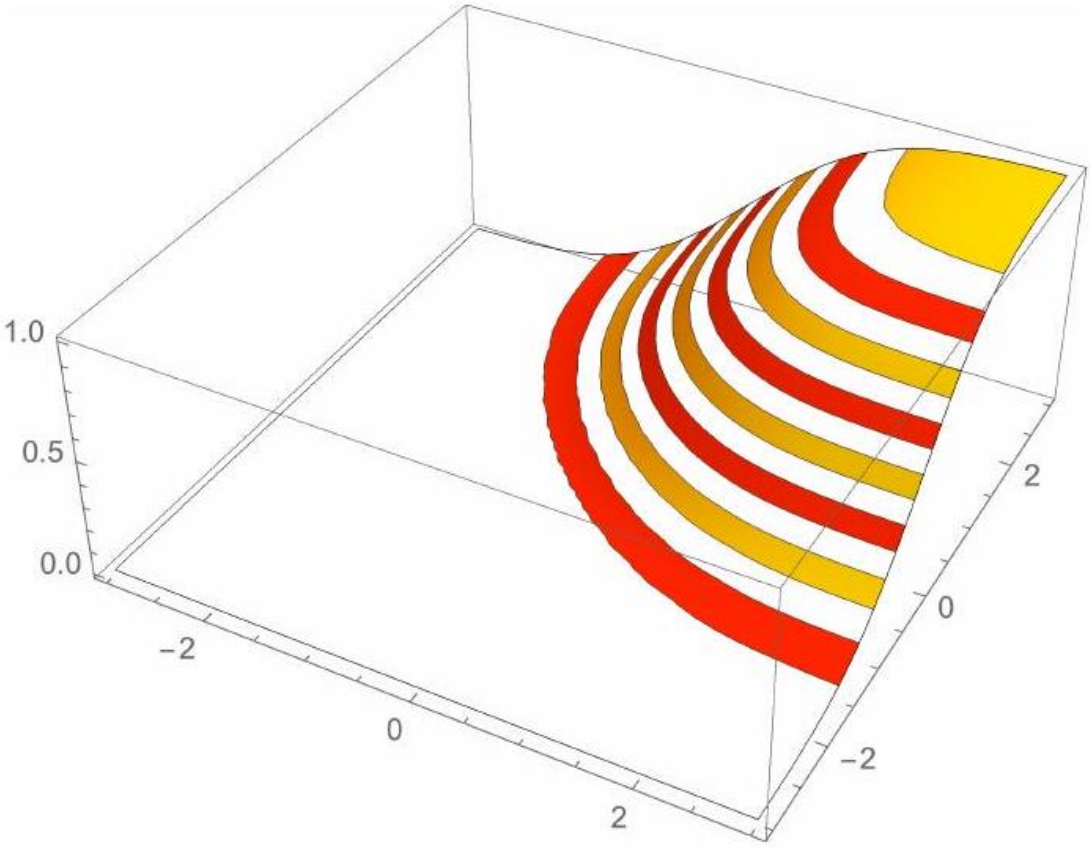
2.7.1 Conditional Probability and Random Variables

The concept of conditional probability works exactly the same with random variables as was discussed above for explicitly defined events. Here, an event is simply a possible set of values the random variable (or variables) can take. For example, with a slight abuse of notation, we might have $A = \{X \leq x\}$ or, more generally, $B = \{X \in \mathcal{B}\}$ for some set $\mathcal{B}$ defined to have a nonempty intersection with the support of X. Now we can construct the conditional probability

$$\mathrm{P}[X \leq x \mid X \in \mathcal{B}] = \frac{\mathrm{P}[X \leq x, X \in \mathcal{B}]}{\mathrm{P}[X \in \mathcal{B}]} = \frac{\mathrm{P}[X \in \mathcal{B} \mid X \leq x]\mathrm{P}[X \leq x]}{\mathrm{P}[X \in \mathcal{B}]} \tag{1.15}$$



(a) Joint PDF.



(b) Joint CDF.

Figure 1.2: Joint probability density and distribution functions (correlated bivariate normal).

One must be careful when conditioning on a continuous random variable taking a point value, e.g., $X = x$. If X is continuously defined, then the "event" $\{X = x\}$ has zero measure, and therefore $\mathrm{P}[X = x] = 0$. So it does not make sense to condition on such events in this way, since the resulting probability will always be zero. However, one can invoke the idea of probability density in such cases. Instead of considering the event $\{X = x\}$, we consider the event $\{x \leq X \leq x + \delta x\}$. Then, for two random variables X and Y, we have

$$\begin{aligned}
\mathrm{P}[Y \leq y \mid x \leq X \leq x + \delta x] &= \frac{\mathrm{P}[Y \leq y, x \leq X \leq x + \delta x]}{\mathrm{P}[x \leq X \leq x + \delta x]} \\
&= \frac{\int_{-\infty}^{y} \int_{x}^{x+\delta x} p_{X,Y}(u, v) \mathrm{d}u \ \mathrm{d}v}{\int_{x}^{x+\delta x} p_X(u) \mathrm{d}u} \\
&= \frac{F_{X,Y}(x + \delta x, y) - F_{X,Y}(x, y)}{F_X(x + \delta x) - F_X(x)}
\end{aligned} \tag{1.16}$$

Dividing the numerator and denominator of this fraction by δx and letting $\delta x \to 0$ yields

$$\begin{aligned}
\mathrm{P}[Y \leq y \mid X = x] &= F_{Y|X}(y \mid X = x) \\
&= \frac{\partial F_{X,Y}(x, y)/\partial x}{\partial F_X(x)/\partial x} \\
&= \frac{\int_{-\infty}^{y} p_{X,Y}(x, v) \mathrm{d}v}{p_X(x)}
\end{aligned} \tag{1.17}$$

Differentiating with respect to y gives

$$p_{Y|X}(y \mid x) = \frac{p_{X,Y}(x,y)}{p_X(x)} \tag{1.18}$$

Hence, we see that the law of conditional probability applies to density functions in the case of continuous random variables.

A standard example that illustrates the use of conditional densities arises when we wish to calculate the probability $\mathrm{P}[XY \leq z]$ for some continuous random variables X and Y and a real scalar z. We can make progress by conditioning on Y in the following way:

$$\begin{aligned}
\mathrm{P}[XY \leq z] &= \iint_{\{(x,y):xy \leq z\}} p_{X,Y}(x,y)\mathrm{d}x\ \mathrm{d}y \\
&= \iint_{\{(x,y):xy \leq z\}} p_{X|Y}(x \mid y)p_Y(y)\mathrm{d}x\ \mathrm{d}y \\
&= \int_{-\infty}^{\infty} p_Y(y) \int_{-\infty}^{z/y} p_{X|Y}(x \mid y)\mathrm{d}x\ \mathrm{d}y \\
&= \int_{-\infty}^{\infty} p_Y(y) F_{X|Y}(z/y \mid Y = y)\mathrm{d}y
\end{aligned} \tag{1.19}$$

2.8 Problem 1.6: Convolution

Recall that the convolution of two functions f and g is defined as

$$(f * g)(t) := \int f(x)g(t-x)\mathrm{d}x$$

Consider two independent random variables X and Y, and let a third random variable Z be equal to their sum, i.e., $Z = X + Y$. Show that the PDF of Z is the convolution of the PDFs of X and Y. (Hint: Begin by writing the CDF of Z as the double integral of $p_{X,Y}(x,y)$ over the region $\{(x,y) : x + y \leq z\}$.)

2.8.1 Functions of Random Variables

You will often encounter situations where a transformation of a random variable (or variables) will occur. Indeed, in the problem above, you encountered $Z = X + Y$. For now, let us consider the simpler example

$$Y = aX + b \tag{1.20}$$

where X is a random variable, and $a > 0$ and b are deterministic scalars. Thus, Y is a random variable through the affine mapping. If we want the probability of Y, we simply write

$$\mathrm{P}[Y \leq y] = \mathrm{P}[aX + b \leq y] = \mathrm{P}[X \leq (y-b)/a] \tag{1.21}$$

We can differentiate with respect to y to find the PDF, which in this case is

$$p_Y(y) = \frac{1}{a} p_X((y-b)/a) \tag{1.22}$$

written as a function of y. If $a < 0$, we would simply negate this term (due to the reversal of the inequality symbol in the probability expression above arising from the division of both sides by a negative number). So in general, for the affine mapping given previously, we can write the PDF of the transformed random variable Y as

$$p_Y(y) = \frac{1}{|a|} p_X((y-b)/a) \tag{1.23}$$

In general, for a mapping $Y = g(X)$ with inverse mapping $X = g^{-1}(Y)$, the CDF of Y can be manipulated to give

$$\mathrm{P}[Y \leq y] = \mathrm{P}[g(X) \leq y] = \mathrm{P}\left[X \leq g^{-1}(y)\right] \tag{1.24}$$

where the $\leq$ sign in the last expression can only be written if g is monotonically increasing in the argument. If it is monotonically decreasing, the inequality would be reversed. And if g is not monotonic, then a more subtle analysis would be needed (indeed, the inverse may not exist). Using this reasoning along with the chain rule for derivatives, it should be easy to see that differenting the right-hand side with respect to y yields the PDF

$$p_Y(y) = \left|\frac{\partial g^{-1}(y)}{\partial y}\right| p_X\left(g^{-1}(y)\right) \tag{1.25}$$

This can be extended to mappings of sequences of random variables to other sequences of random variables, where the absolute value of the prefactor becomes the Jacobian determinant, as is observed with coordinate transformations in multivariate calculus.

2.9 Problem 1.7: The reciprocal transformation

Let X be a strictly positive random variable with PDF $p_X(x) = e^{-x}$ for $x > 0$.
(i) Show that p_X is a valid density function by integrating over the support of X.
(ii) Write expressions for $\mathrm{P}[X \leq x]$ and $\mathrm{P}[X > x]$, and verify that $\mathrm{P}[X \leq x] = 1 - \mathrm{P}[X > x]$ for this example.
(iii) Find an expression for the PDF of the random variable $Y = X^{-1}$ by employing the reasoning given in eq. (1.24) and the surrounding text, then verify that the formula given in eq. (1.25) gives the same result.

2.9.1 Statistical Averages

Let X be a discrete random variable that can take on n values, the i th such value being denoted as x_i. Let $p_i = \mathrm{P}\left[X = x_i\right]$. Then the average, or expected value, or expectation of X is given by

$$\mathrm{E}[X] = \sum_{i=1}^{n} x_i p_i$$

Clearly, if $p_i = 1/n$ for all i, then X is uniformly distributed and the expected value is just the arithmetic mean.

The notion of the expected value can be adapted to compute statistical averages in many scenarios. Consider, for example, a real-valued, continuous random variable X. The expected value of X in this case is [3]

$$\mathrm{E}[X] = \int x p_X(x) \mathrm{d}x \tag{1.26}$$

In a manner that echos calculus and physics, we say the nth moment of X is

$$\mathrm{E}\left[X^n\right] = \int x^n p_X(x) \mathrm{d}x \tag{1.27}$$

The quantity

$$\mathrm{E}\left[(X - \mathrm{E}[X])^n\right] = \int (x - \mathrm{E}[X])^n p_X(x) \mathrm{d}x \tag{1.28}$$

is known as the nth central moment of X. In the case of a mapping from one random variable to another, such as $Y = g(X)$, the expected value of Y can either be written as

$$\mathrm{E}[Y] = \int y p_Y(y) \mathrm{d}y \tag{1.29}$$

or as

$$\mathrm{E}[g(X)] = \int g(x)p_X(x)\mathrm{d}x \tag{1.30}$$

This subtlety can be very useful in practice when the integral of $yp_Y(y)$ appears difficult or intractable. A particularly useful mapping is $g(X) = (X - \mathrm{E}[X])^2$, which gives the variance of X :

$$\mathrm{V}[X] = \int (x - \mathrm{E}[X])^2 p_X(x)\mathrm{d}x \tag{1.31}$$

The standard deviation of X is defined as the square root of its variance, i.e.,

$$\sigma_X = \sqrt{\mathrm{V}[X]} \tag{1.32}$$

The (m, n) th joint moment of continuous variables X and Y is given by

$$\mathrm{E}\left[X^m Y^n\right] = \iint x^m y^n p_{X,Y}(x, y)\mathrm{d}x\ \mathrm{d}y \tag{1.33}$$

The joint moment $\mathrm{E}[XY]$ is related to the correlation of X and Y. Specifically, the correlation coefficient for X and Y is defined as

$$\rho_{XY} = \frac{\mathrm{E}[XY] - \mathrm{E}[X]\mathrm{E}[Y]}{\sigma_X \sigma_Y} \tag{1.34}$$

where σ_X and σ_Y are the standard deviations of X and Y, respectively. These random variables are said to be uncorrelated if $\mathrm{E}[XY] = \mathrm{E}[X]\mathrm{E}[Y]$. Note that if X and Y are statistically independent, we have

$$\begin{aligned}
\mathrm{E}[XY] &= \iint xyp_{X,Y}(x, y)\mathrm{d}x\ \mathrm{d}y \\
&= \iint xyp_X(x)p_Y(y)\mathrm{d}x\ \mathrm{d}y \\
&= \left(\int xp_X(x)\mathrm{d}x\right)\left(\int yp_Y(y)\mathrm{d}y\right) \\
&= \mathrm{E}[X]\mathrm{E}[Y]
\end{aligned} \tag{1.35}$$

So statistical independence implies uncorrelatedness. However, this rule does not work the other way around unless X and Y are marginally and jointly normal random variables.

The joint central moment is defined as

$$\mathrm{E}\left[(X - \mathrm{E}[X])^m (Y - \mathrm{E}[Y])^n\right] = \iint (x - \mathrm{E}[X])^m (y - \mathrm{E}[Y])^n p_{X,Y}(x, y)\mathrm{d}x\ \mathrm{d}y \tag{1.36}$$

The joint central moment with $m = n = 1$ is known as the covariance of X and Y.
When computing the expectation of a function of two (or more) random variables, conditioning can be used to simplify matters. We can, in effect, fix one variable and average over the other. Then we can average the result of that calculation over the variable we originally fixed. More rigorously, let g be a function of two random variables, X and Y. Then we can write

$$\begin{aligned}
\mathrm{E}[g(X, Y)] &= \iint g(x, y)p_{X,Y}(x, y)\mathrm{d}x\ \mathrm{d}y \\
&= \iint g(x, y)p_{Y|X}(y \mid x)p_X(x)\mathrm{d}x\ \mathrm{d}y \\
&= \int p_X(x) \int g(x, y)p_{Y|X}(y \mid x)\mathrm{d}y\ \mathrm{d}x \\
&= \int p_X(x)\mathrm{E}[g(X, Y) \mid X]\mathrm{d}x \\
&= \mathrm{E}[\mathrm{E}[g(X, Y) \mid X]]
\end{aligned} \tag{1.37}$$

[3] All integrals in this short section are over the support of the respective random variables.

A word on notation is in order. In the second to last line above, $\mathrm{E}[\cdot \mid \cdot]$ represents the expectation of the first argument conditioning on the second argument. In other words,
we assume the second argument is not actually a random variable when we compute the expectation. In the last line, we have two expectations. The inner expectation is taken with respect to the random variable Y while conditioning on X as noted above. The outer expectation is the remaining average taken over the distribution of X.

Finally, it is important to note that, for discrete random variables, the integrals in the expressions given above become summations. The rest of the results are identical, i.e., one can, in an analogous way, compute the variance, moments, correlation coefficients, etc.

2.10 Problem 1.8: Expected value and variance

Let X be a nonnegative, continuous random variable with PDF $p_X(x) = e^{-x}$ for $x \geq 0$. Compute the expected value and variance of X.

2.10.1 Some Useful Probability Distributions

So far, we have discussed probability and random variables using general formalisms. The exception has been in some of the problems, where particular distributions have surreptitiously been introduced. For example, the uniform distribution was defined such that the probability of each outcome has equal measure. Here, we review some further basic, but very useful, probability distributions.

Binomial Distribution: Consider a coin toss, where a heads occurs with probability p and a tails is observed with probability $1 - p$. We can define a random variable for each toss, say X, which can take a value of " 1 " for a heads and " 0 " for a tails. We might ask the question: what is the probability that we observe k heads in n tosses of the coin (where, clearly, $k \leq n$)? Since each coin toss is independent of the others, we can mathematically express the number of heads observed by the random variable

$$Y = \sum_{i=1}^{n} X_i$$

where X_i is the variable related to the i th toss. The probability that $Y = k$ is just the probability that heads comes up k times multiplied by the probability that tails is observed the rest of the time, and all of this is multiplied by the number of combinations one might observe k heads in n trials of the experiment, which is just the binomial coefficient $\binom{n}{k}$. More succinctly, we have that Y follows a binomial distribution such that

$$\mathrm{P}[Y = k] = \binom{n}{k} p^k (1-p)^{n-k} \tag{1.38}$$

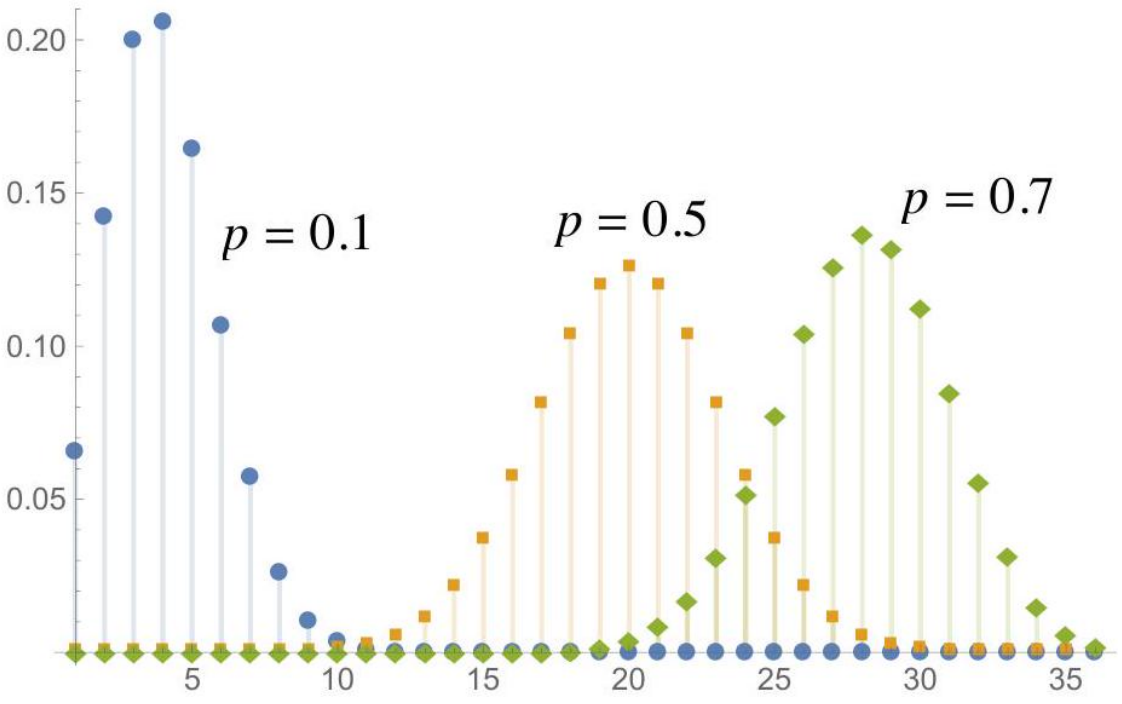


Figure 1.3: Binomial probability mass function ($n = 40$).

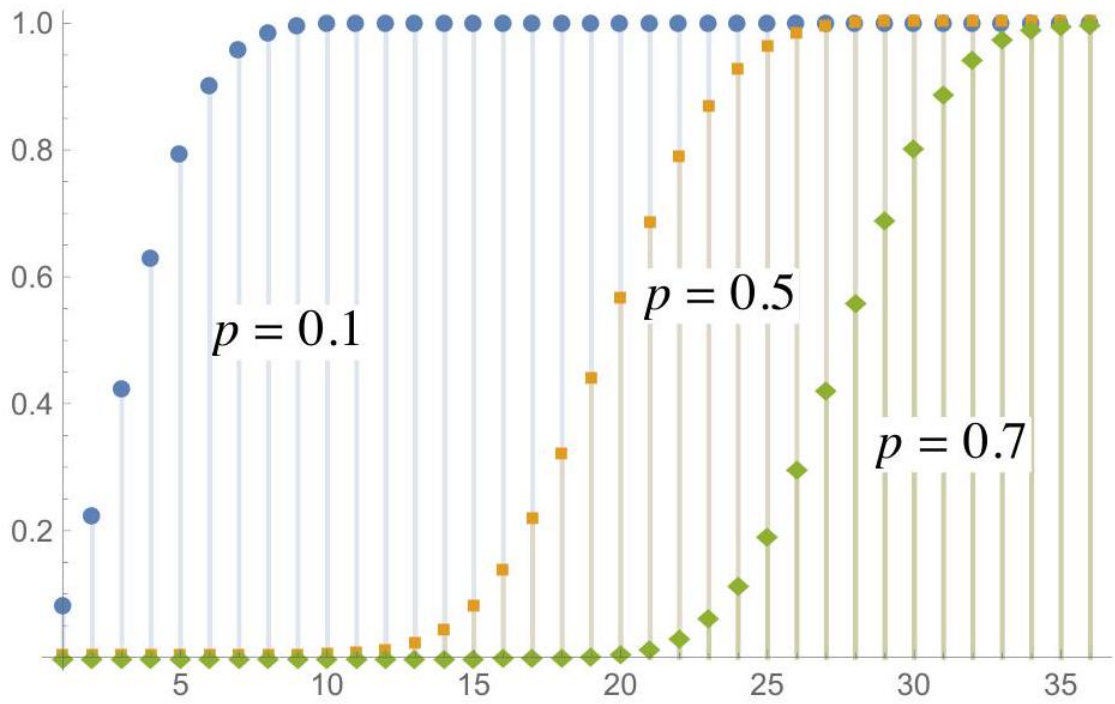


Figure 1.4: Binomial cumulative distribution ($n = 40$).

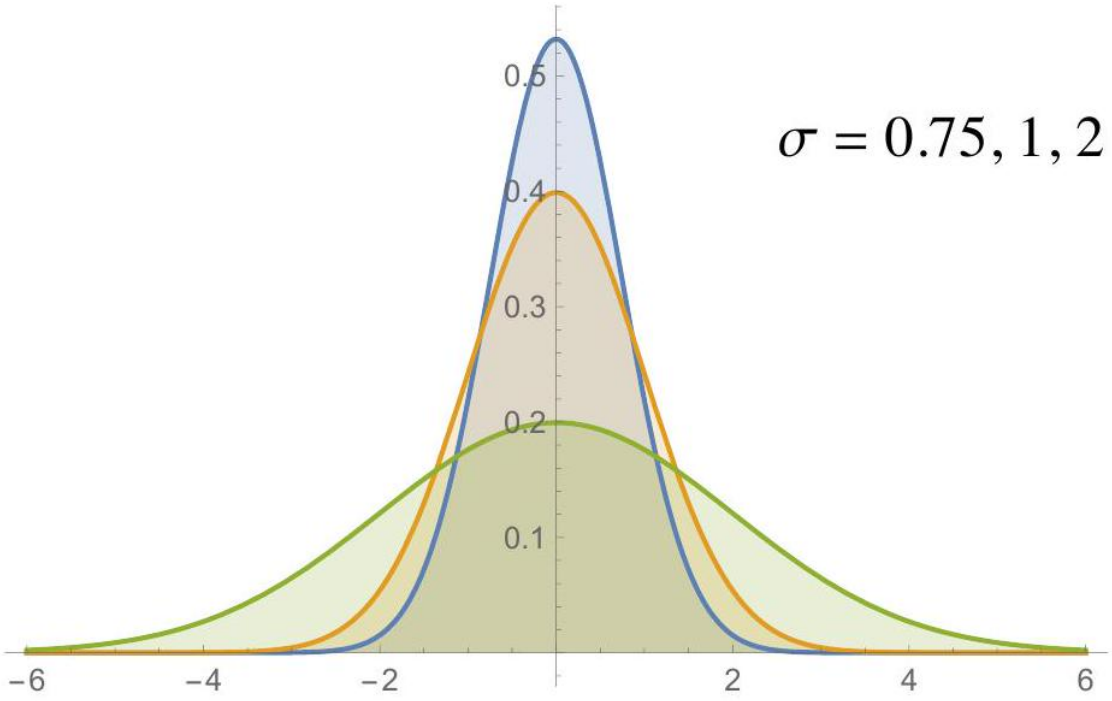


Figure 1.5: Gaussian density function ($\mu = 0$).

Normal (Gaussian) Distribution: The PDF of a normally distributed random variable X is

$$p_X(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{1.39}$$

The parameters μ and σ^2 are the mean (expected value) and variance of X, respectively. Hence, σ is the standard deviation. The CDF of the normal random variable is obtained by integrating:

$$\begin{aligned} F_X(x) &= \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{x} e^{-\frac{(u-\mu)^2}{2\sigma^2}} \, \mathrm{d}u \\ &\overset{(a)}{=} \frac{1}{2} \frac{2}{\sqrt{\pi}} \int_{-\infty}^{\frac{x-\mu}{\sqrt{2}\sigma}} e^{-t^2} \, \mathrm{d}t \\ &\overset{(b)}{=} \frac{1}{2} \left(1 + \mathrm{erf}\left(\frac{x-\mu}{\sqrt{2}\sigma} \right) \right) \end{aligned} \tag{1.40}$$

where (a) follows from a substitution of variables, and (b) results from the definition of the error function $\mathrm{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2} \, \mathrm{d}t$ along with the fact that integrating a normal PDF over half of the real line gives $1/2$. You will also often see the Gaussian tail function used to express the CDF. This function is defined as

$$Q(x) = \frac{1}{\sqrt{2\pi}} \int_x^{\infty} e^{-\frac{t^2}{2}} \, \mathrm{d}t \tag{1.41}$$

Note that Q is monotonically decreasing in the argument. Using this function, the CDF can be expressed as

$$F_X(x) = 1 - Q\left(\frac{x-\mu}{\sigma} \right) \tag{1.42}$$

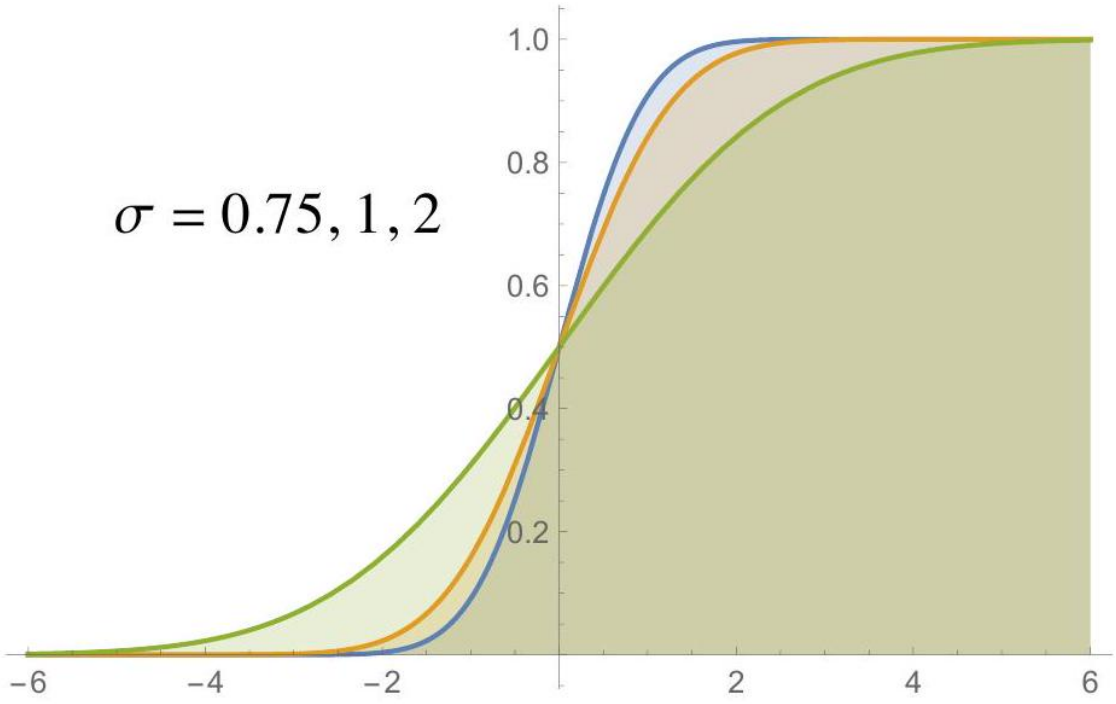


Figure 1.6: Gaussian distribution function ($\mu = 0$).

Central χ^2 Distribution: Suppose we sum the squares of n independent, identical normally distributed random variables to obtain a new variable

$$Y = \sum_{i=1}^{n} X_i^2 \tag{1.43}$$

If each of the Gaussian X_i variables is zero mean with variance σ^2, then Y has a central χ^2 distribution with n degrees of freedom. The PDF of Y is

$$p_Y(y) = \frac{y^{\frac{n}{2}-1} e^{-\frac{y}{2\sigma^2}}}{\sigma^n 2^{\frac{n}{2}} \Gamma\left(\frac{n}{2}\right)}, \quad y \geq 0 \tag{1.44}$$

The function $\Gamma(x)$ is known as the gamma function. It is an analytical continuation of the factorial function to all arguments in the complex plane. It is typically expressed as the integral

$$\Gamma(x) = \int_0^\infty t^{x-1} e^{-t} \, \mathrm{d}t \tag{1.45}$$

The CDF of Y can be obtained by integrating in the usual way. Unless n is even, however, it does not admit a simple closed form.

Exponential Distribution: If the random variable Y in the central χ^2 example above has two degrees of freedom, it is said to be exponentially distributed with mean $2\sigma^2$. The PDF of Y in this case is

$$p_Y(y) = \frac{1}{2\sigma^2} e^{-\frac{y}{2\sigma^2}}, \quad y \geq 0 \tag{1.46}$$

The CDF is

$$F_Y(y) = 1 - e^{-\frac{y}{2\sigma^2}} \tag{1.47}$$

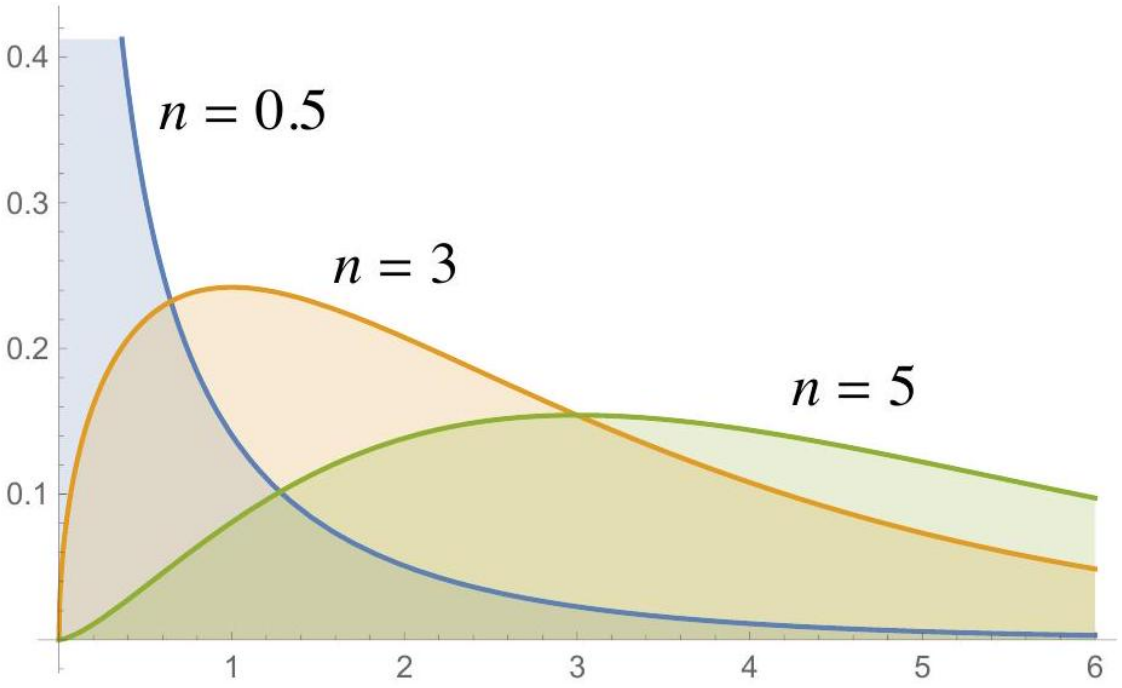


Figure 1.7: χ^2 density function ($\sigma = 1$).

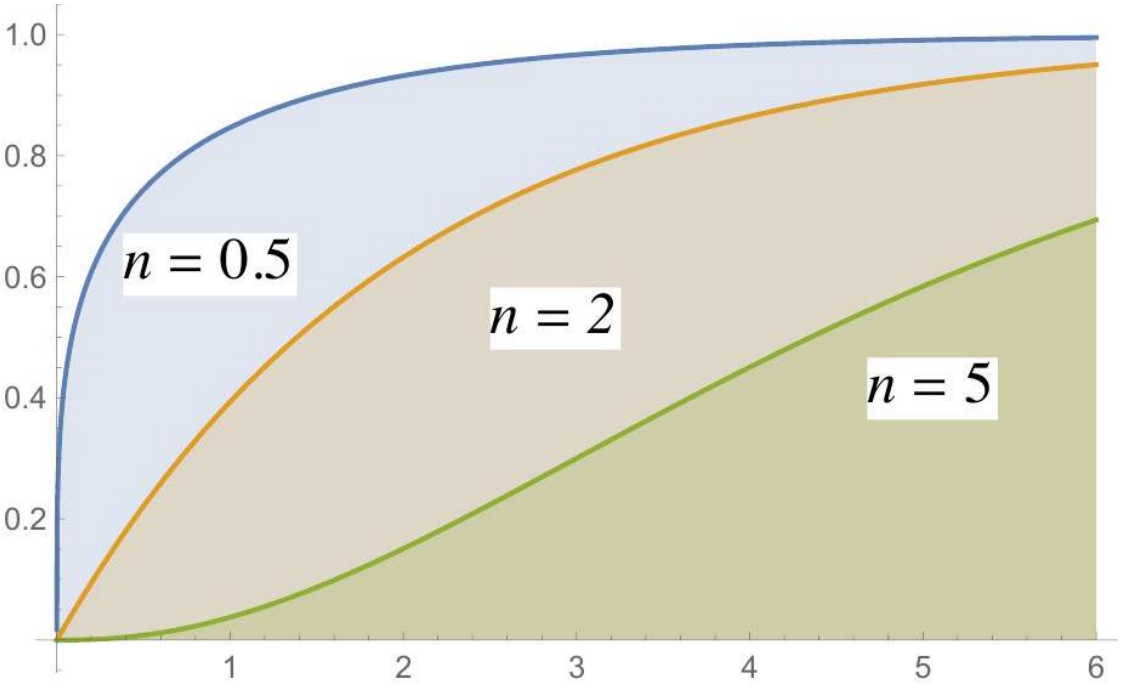


Figure 1.8: χ^2 distribution ($\sigma = 1$).

You met this distribution in some of the problems given in previous sections. In those cases, observe that $\sigma = 1/\sqrt{2}$.

2.11 Problem 1.9: Properties of the χ^2 distribution

Let Y be χ^2 distributed with $n = 5$ degrees of freedom and parameter σ.
(i) Show that the PDF of Y is a valid PDF. (Recall that $\Gamma(x) = \int_0^\infty t^{x-1}e^{-t}\ \mathrm{d}t$.)
(ii) Determine an expression for $\mathrm{E}[Y^m]$ in terms of σ and m, where $m > 0$.

2.11.1 Characteristic Function

An important function encountered in probability theory is the characteristic function, which is defined as

$$\psi_X(t) = \mathrm{E}\left[e^{itX}\right] \tag{1.48}$$

where $i = \sqrt{-1}$. Note that if the random variable X is continuous, we have

$$\psi_X(t) = \int_{-\infty}^{\infty} p_X(x) e^{itx} \, \mathrm{d}x \tag{1.49}$$

and thus, from Fourier analysis, it follows that

$$p_X(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \psi_X(t) e^{-itx} \, \mathrm{d}t \tag{1.50}$$

This should give some idea of how important characteristic functions are. All of the machinery we have at our disposal regarding Fourier transforms can be applied to analyse probability distributions.

In particular, consider the random variable Y formed by summing n independent, identically distributed (i.i.d.) random variables $\{X_k\}$:

$$Y = \sum_{k=1}^{n} X_k \tag{1.51}$$

The characteristic function of Y is

$$\begin{aligned}
\psi_Y(t) &= \mathrm{E}\left[e^{itY}\right] \\
&= \mathrm{E}\left[e^{it\sum_{k=1}^{n} X_k}\right] \\
&\overset{(a)}{=} \mathrm{E}\left[e^{itX_1}\right]^n \\
&\overset{(b)}{=} \psi_X(t)^n
\end{aligned}$$

where (a) follows from the i.i.d. property and (b) follows from the definition of the characteristic function. Hence, we can find the density function of Y by calculating the characteristic function of X_1, raising this to the nth power, and inverting the Fourier transform.

2.12 Problem 1.10: Convolution revisited

Let X and Y be independent, continuous, real-valued random variables that have identical distributions (i.e., they are i.i.d.). Let $Z = X + Y$. By considering the characteristic function of Z and properties of Fourier transforms, show that the PDF of Z is the convolution of the PDFs of X and Y.

2.13 Chapter 2

2.14 Stochastic Processes

Randomness in a given system is typically a function of time. Random variables that evolve with time are referred to as random processes or stochastic processes. To study the basic properties of stochastic processes, we require knowledge of a few more mathematical concepts and tools that can be built from those discussed in the preceding chapter.

2.14.1 Stationarity

A stochastic process $X(t)$ can be thought of as a random variable that is a function of time t. Thus, at a given time, say t_1, the process becomes a random variable $X_{t_1} = X(t_1)$. Of course, the process could be correlated at different times. Consequently, we can form a joint PDF to describe the statistics of the process at these times [1] :

$$p(x_{t_1}, x_{t_2}, \ldots, x_{t_n}) \tag{2.1}$$

But how do the statistics at these times relate to statistics at another set of times? Consider the case where we shift the original set of times by some fixed time τ and observe the corresponding PDF

$$p(x_{t_1+\tau}, x_{t_2+\tau}, \ldots, x_{t_n+\tau}) \tag{2.2}$$

The two PDFs written above may or may not be identical. In a system where the statistical properties do not fluctuate from the first set of times to the second set of times, we will have

$$p(x_{t_1}, x_{t_2}, \ldots, x_{t_n}) = p(x_{t_1+\tau}, x_{t_2+\tau}, \ldots, x_{t_n+\tau}) \tag{2.3}$$

If this equality is true for all n and τ, then the random process is said to be strict-sense stationary (SSS). We will observe another type of stationarity below.

2.15 Problem 2.1: The weather

Let $T(t)$ be a stochastic process that represents the temperature measured at a certain point on Earth at time t. Is $T(t)$ strict-sense stationary? Provide a physical argument that supports your answer.

2.15.1 Autocorrelation and Autocovariance

Statistical averages work for random processes in the same way as they do for random variables. A particularly important observable is the autocorrelation function, which, for a real-valued, continuous stochastic process $X(t)$, is defined as

$$\phi(t_1, t_2) = \mathrm{E}[X_{t_1} X_{t_2}] = \iint x_{t_1} x_{t_2} p(x_{t_1}, x_{t_2}) \, \mathrm{d}x_{t_1} \, \mathrm{d}x_{t_2} \tag{2.4}$$

If $X(t)$ is continuous and complex-valued, the autocorrelation function is defined as

$$\phi(t_1, t_2) = \frac{1}{2}\mathrm{E}\left[X_{t_1} X_{t_2}^*\right] = \frac{1}{2}\iint x_{t_1} x_{t_2}^* p(x_{t_1}, x_{t_2}) \, \mathrm{d}x_{t_1} \, \mathrm{d}x_{t_2} \tag{2.5}$$

Analogous definitions hold for discrete processes, where the integrals are replaced with sums and densities are replaced with probabilities.

When the process $X(t)$ is SSS, the autocorrelation function is the same for any constant shift of the arguments. This implies the function is not dependent upon t_1 and t_2 specifically, but on their difference $\tau = t_1 - t_2$. Thus, we can write

$$\phi(\tau) = \mathrm{E}[X_{t_1} X_{t_1-\tau}] \tag{2.6}$$

for some arbitrary t_1 when the process is real valued. Note that for a complex stochastic process, we have

$$\phi^*(\tau) = \frac{1}{2}\mathrm{E}\left[X_{t_1}^* X_{t_1-\tau}\right] = \frac{1}{2}\mathrm{E}\left[X_{t_1} X_{t_1+\tau}^*\right] = \phi(-\tau) \tag{2.7}$$

[1] Note that we dropped the subscript notation for the PDF here. We will adopt this simplified notation where the distribution or random variable in question is clear.

If the mean of the process $X(t)$ is independent of time (i.e., $\mathrm{E}[X(t)]$ = constant) and the autocorrelation function depends only on the time difference τ, then the process is said to be wide-sense stationary (WSS). Wide-sense stationarity is less strict than strict-sense stationarity, and often arises in the description of engineered systems.

Another important observable is the autocovariance function

$$\rho(t_1, t_2) = \mathrm{E}\left[\left(X_{t_1} - \mathrm{E}\left[X_{t_1}\right]\right)\left(X_{t_2} - \mathrm{E}\left[X_{t_2}\right]\right)\right] \tag{2.8}$$

$$= \iint \left(x_{t_1} - \mathrm{E}\left[X_{t_1}\right]\right)\left(x_{t_2} - \mathrm{E}\left[X_{t_2}\right]\right) p\left(x_{t_1}, x_{t_2}\right) \mathrm{d}x_{t_1}\ \mathrm{d}x_{t_2} \tag{2.9}$$

$$= \phi(t_1, t_2) - \mathrm{E}\left[X_{t_1}\right]\mathrm{E}\left[X_{t_2}\right] \tag{2.10}$$

The autocovariance function is effectively the autocorrelation function centralised about the mean of the process. This is analogous to the notion of moments and central moments in relation to random variables. When the process is WSS, the autocovariance function becomes

$$\rho(\tau) = \phi(\tau) - \mu^2 \tag{2.11}$$

where $\mu = \mathrm{E}\left[X_{t_1}\right] = \mathrm{E}\left[X_{t_2}\right]$ and $\tau = t_1 - t_2$.

2.16 Problem 2.2: Wide-sense stationarity

Let $X(n) \in \{0, 1\}$ denote a discrete-time random process where each random variable in the process represents an independent coin flip, where heads ($X(n) = 1$) appears with probability $1/2$. Compute the autocorrelation function of $X(n)$ and determine whether or not this process WSS.

Now suppose we change the coin with each flip, where the probability of a heads coming up on the nth flip is $1/2^n$ for $n = 1, 2, \ldots$ Let us denote this new process as $Y(n)$. Compute the autocorrelation function of $Y(n)$ and determine whether or not this process WSS.

2.16.1 Cross-Correlation and Cross-Covariance

We can compare two different stochastic processes by studying their cross-correlation. The cross-correlation function of two real-valued processes $X(t)$ and $Y(t)$ is defined as

$$\phi_{xy}(t_1, t_2) = \mathrm{E}\left[X_{t_1} Y_{t_2}\right] \tag{2.12}$$

If $X(t)$ and $Y(t)$ are complex, the definition becomes

$$\phi_{xy}(t_1, t_2) = \frac{1}{2}\mathrm{E}\left[X_{t_1} Y_{t_2}^*\right] \tag{2.13}$$

If $X(t)$ and $Y(t)$ are jointly and independently stationary, the cross-correlation function only depends on the time difference $\tau = t_1 - t_2$. Moreover, it can be shown that

$$\phi_{xy}(\tau) = \phi_{yx}(-\tau) \tag{2.14}$$

The cross-covariance function is given by

$$\rho_{xy}(t_1, t_2) = \phi_{xy}(t_1, t_2) - \mathrm{E}\left[X_{t_1}\right]\mathrm{E}\left[Y_{t_2}\right] \tag{2.15}$$

The cross-covariance is a centralised version of the cross-correlation. This is analogous to the relation between the autocorrelation and autocovariance functions.

2.17 Problem 2.3: Coin tosses

Let $X(n)$ and $Y(n)$ denote independent, discrete-time random processes, where each random variable in each process represents a coin flip that is independent of other flips. For the process $X(n)$, a heads appears with probability 1/4. For the process $Y(n)$, a heads appears with probability 3/4. Compute the crosscorrelation and cross-covariance functions of $X(n)$ and $Y(n)$.

Now let $Z(n)$ be a stochastic process that adheres to the following rule:

$$Z_n = \begin{cases} X_n, & \text{with probability } 1/2 \\ Y_n, & \text{with probability } 1/2 \end{cases}$$

That is, the nth sample is drawn uniformly from the set $\{X_n, Y_n\}$, which are, in turn, binary random variables with probabilities defined above. Compute the cross-correlation function of $X(n)$ and $Z(n)$. (Hint: Condition on the process $Z(n)$ takes values from, then compute the resulting expectations.)

2.17.1 Power Density Spectrum

It is often important to understand the spectral properties of signals that evolve with time. The Fourier transform gives us the machinery that we need to carry out this task. But the idea of taking the Fourier transform of a random signal is not well defined.

To make progress, consider a stationary stochastic process, which may represent a random signal in an engineering context. The mean of a stationary stochastic process is independent of time; it is a constant. However, we cannot immediately observe the signal's spectrum, since we do not actually know what the signal is for all time [2]. However, we can use the autocorrelation function of this process to obtain information about the signal's spectrum. The distribution of power as a function of frequency f is given by

$$\Phi(f) = \int_{-\infty}^{\infty} \phi(\tau) e^{-i2\pi f\tau} \mathrm{d}\tau \tag{2.16}$$

The inverse exists as a result of Fourier properties:

$$\phi(\tau) = \int_{-\infty}^{\infty} \Phi(f) e^{i2\pi f\tau} \mathrm{d}f \tag{2.17}$$

Note that

$$\phi(0) = \int_{-\infty}^{\infty} \Phi(f) \mathrm{d}f = \mathrm{E}\left[|X_{t_1}|^2\right] \geq 0 \tag{2.18}$$

represents the average power of the signal. $\Phi(f)$ is thus known as the power density spectrum [3].

If $X(t)$ is real, $\phi(\tau)$ is real and even, so $\Phi(f)$ is real and even. If $X(t)$ is complex, $\phi(\tau) = \phi^*(-\tau)$, so $\Phi(f)$ is real, but not even.

In the case of discrete-time processes, we can think of the same formalism but where the autocorrelation function is sampled uniformly every T seconds. Mathematically, the sampled autocorrelation function is

$$\phi_s(\tau) = \sum_{m=-\infty}^{\infty} \phi(\tau)\delta(\tau - mT) \tag{2.19}$$

where $\delta(\cdot)$ is the Dirac delta function. Taking the Fourier transform gives the power density spectrum

$$\Phi(f) = \sum_{m=-\infty}^{\infty} \phi(mT) e^{-i2\pi fmT} \tag{2.20}$$

[2] Recall that computing the Fourier transform of a signal involves integrating over the real line, which represents time.

Typically, the autocorrelation function of a WSS discrete-time process is written as a function of the sample index m, not the sample time mT. Hence, one often sees $\phi(m)$ instead of $\phi(mT)$ when studying discrete-time processes. Adopting this slightly modified notation, we can write the autocorrelation function of a discrete-time stochastic process by taking the inverse Fourier transform of the power density spectrum:

$$\phi(m) = T \int_{-\frac{1}{2T}}^{\frac{1}{2T}} \Phi(f) e^{i2\pi f mT} \, \mathrm{d}f \tag{2.21}$$

2.18 Problem 2.4: Proving the inverse transform

Prove that eq. (2.21) is true by substituting eq. (2.20) for $\Phi(f)$ and integrating term by term. (Hint: You may find Euler's formula, $e^{ix} = \cos x + i \sin x$, useful.)

2.18.1 Linear Time-Invariant System

Suppose we have a deterministic signal $x(t)$ that passes through a linear filter with impulse response $h(t)$. The output is given by the convolution of these two functions:

$$y(t) = (h * x)(t) = \int_{-\infty}^{\infty} h(\tau) x(t - \tau) \mathrm{d}\tau \tag{2.22}$$

Such a model is known as a linear time-invariant system. Now assume the initial signal is a WSS random process $X(t)$ with expectation $\mathrm{E}[X(t)] = \mu_x$. We wish to

analyse the output $Y(t)$ given the deterministic filter response $h(t)$. Such problems often arise in engineering applications.

To begin, we can write the mean (expectation) of $y(t)$ as

$$\mathrm{E}[y(t)] = \int_{-\infty}^{\infty} h(\tau) \mathrm{E}[x(t - \tau)] \mathrm{d}\tau = \mu_x \int_{-\infty}^{\infty} h(\tau) \mathrm{d}\tau = \mu_x H(0) \tag{2.23}$$

where $H(0)$ is the frequency response (i.e., the Fourier transform) of $h(t)$ at $f = 0$. By using the definition of the autocorrelation function, and assuming $Y(t)$ is complex, we have that

$$\phi_{yy}(t_1, t_2) = \frac{1}{2} \mathrm{E}\left[Y_{t_1} Y_{t_2}^*\right] \tag{2.24}$$

it can be shown that the stationarity of $X(t)$ implies $Y(t)$ is also stationary; indeed, we know the mean of $Y(t)$ does not change over time. Moreover, by taking the Fourier transform of ϕ_{yy}, it can be shown that the power density spectrum of the output $Y(t)$ can be written as

$$\Phi_{yy}(f) = \Phi_{xx}(f) |H(f)|^2 \tag{2.25}$$

where Φ_{xx} is the power density spectrum of $X(t)$. Thus, the power density spectra of $X(t)$ and $Y(t)$ are related by a simple multiplication by the power spectrum of the channel response.

Discrete-time systems can be treated in the same way as continuous-time systems. The difference arises in the definition of the convolution operator, which is no longer an integral, but is defined as

$$(x * y)(n) = \sum_{k=-\infty}^{\infty} x(k) y(n - k) \tag{2.26}$$

Going through the same process as was done for continuous-time systems, one can show that

$$\Phi_{yy}(f) = \Phi_{xx}(f) |H(f)|^2 \tag{2.27}$$

but in this case the power density spectrum is periodic.

[3] Sometimes, we use the term power spectral density instead.

2.19 Problem 2.5: Power density spectrum of an LTI system

Prove eq. (2.25) by direct calculation.

Chapter 3

Introduction to Information Theory

3.1 Entropy, Mutual Information and Relative Entropy

3.2 Sources and Coding

In this section, we will cover the basic ideas behind coding for different types of sources. Recall that the three sources we will be interested in are discrete sources, analogue sequence sources, and analogue waveform sources.

3.2.1 Coding for Discrete Sources

A discrete source can be viewed as a machine that generates numbers from the set $\mathcal{S} = \{1, \ldots, M\}$, where M is a positive integer [24]. If we begin with a source that generates characters that are different from this set (e.g., text), we can easily create a one-to-one mapping from that set to $\mathcal{S}$. Each source symbol is generated with its own probability; it is not necessarily the case that all source symbols are generated with equal probability.

The goal of a discrete source encoder is to efficiently represent the members of $\mathcal{S}$ as binary strings. There are two general approaches to doing so: fixedlength encoding and variable-length encoding. In both cases, it is possible to create an encoder that is surjective, but not one-to-one. In other words, we may create a rule whereby two distinct symbols in $\mathcal{S}$ are mapped to the same bit string. However, we will only concern ourselves with one-to-one mappings here, since this enables us to perfectly recover the source symbols in $\mathcal{S}$ at the decoder.

Fixed-Length Encoding

A fixed-length encoder simply maps each symbol in $\mathcal{S}$ to a unique bit string of length $\lceil \log M \rceil$, where the notation $\lceil a \rceil$ denotes the smallest integer greater than or equal to a. If M is a power of two, this mapping is complete in the sense that all bit strings of length $\lceil \log M \rceil = \log M$ represent source symbols. If M is not a power of two, then there exist more bit strings that are required to

[24] We use curly braces to denote sets. Here, the elements of the set $\mathcal{S}$ are listed explicitly. Frequently, we will write expressions such as $\{a_i\}$ to refer to the set $\{a_1, a_2, \ldots\}$ when the range of indices is understood from the context.

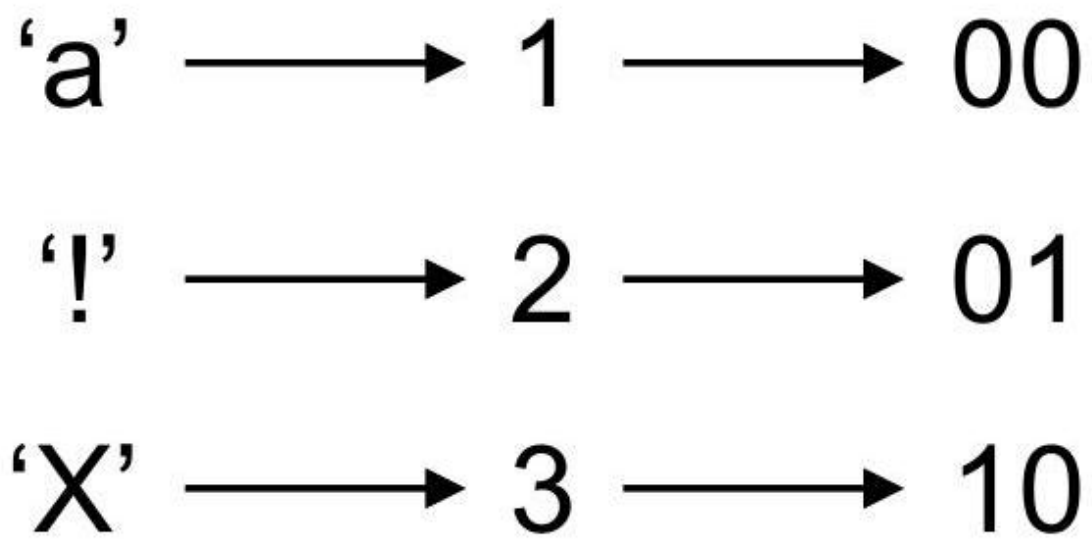


Figure 9: Example of fixed-length source encoding. The source generates three ASCII characters, which are mapped to the set $\mathcal{S} = \{1, 2, 3\}$. These are then encoded as two-bit strings. The rate is two bits per source symbol, which is somewhat inefficient, since we do not use the string 11. represent the source message, which is inefficient. The rate R of a fixed-length source code is

$$R = \lceil \log M \rceil \quad \text{bits per source symbol.}$$

A simple example of fixed-length encoding is shown in Figure 9.
One way to get around the inefficiency described above and in the Figure is to collect source symbols into groups of a given size and encode each group as a bit string rather than encoding each symbol and concatenating the resulting strings. To see why this is more efficient, consider the average rate

$$\bar{R} = \frac{1}{n} \lceil \log M^n \rceil \quad \text{bits per source symbol}$$

in the case of grouping n symbols together. We can bound the rate as follows:

$$\log M = \frac{1}{n} \log M^n \leq \bar{R} \leq \frac{1}{n} (\log M^n + 1) = \log M + \frac{1}{n}$$

Hence, as n grows large, the average rate tends to $\log M$, which is at least as good as $\lceil \log M \rceil$. The disadvantage of taking this approach lies with the increased complexity of encoding large source symbol strings.

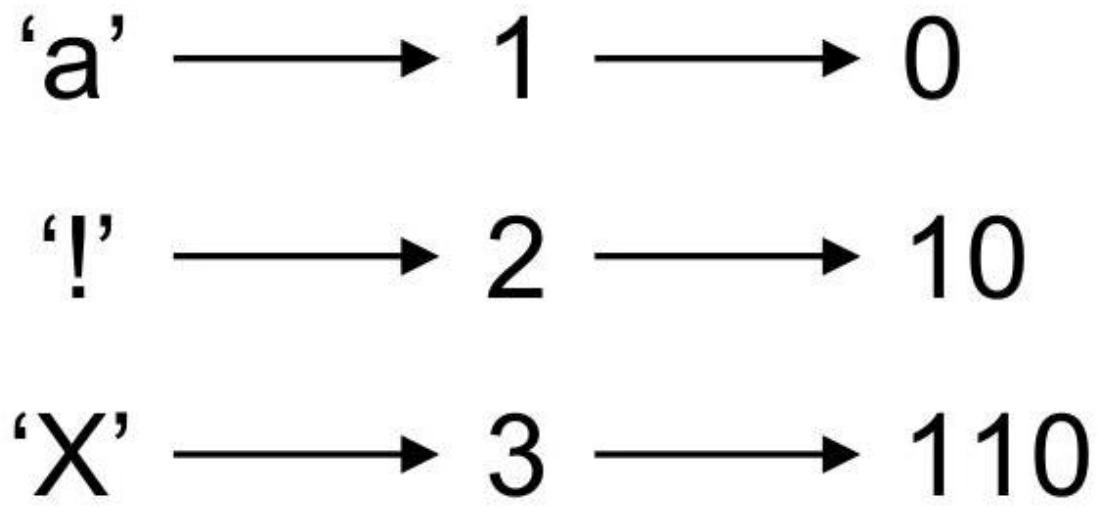


Figure 10: Example of a prefix code.

Variable-Length Encoding

Another way to overcome the inefficiencies inherent in fixed-length encoding is to map source symbols to bit strings with varying lengths. It makes sense to create a mapping whereby frequently appearing source symbols are mapped to short bit strings, and infrequently appearing symbols are encoded as longer bit strings. An immediate problem with this approach presents itself. If sequences of source symbols are represented by concatenating bit strings of different lengths, how will the decoder know when one bit string ends and another begins?

The answer to this question lies in the idea of the prefix code. A prefix code is a source code that is designed such that no codeword (i.e., bit string) is a prefix of another codeword. An example of such a code is shown in Figure 10. Note that the fixed-length codes described above are, by definition, prefix codes.

Prefix codes are instantaneously and uniquely decodable. In other words, as soon as a codeword arrives at the input of the decoder, it is decoded uniquely to obtain the original source symbol. It is important to note that there exist codes that are not instantaneously decodable, but are uniquely decodable. In this case, the decoder would need to consider strings of input bits that follow the string it wishes to decode in order to uniquely decode that codeword.

A quick glance at the example given in Figure 10 reveals that a more efficient code exists. Indeed, we could use the string 11 to represent ’ X ’ instead of 110 . The resulting code would still be prefix free, and we would have reduced the encoding overhead by one bit. Two questions arise from this
observation:

1. What is the optimal length of each codeword?

2. How can we construct optimal codes?

As noted above, a key consideration in designing variable-length codes is the relative frequency of occurrence of each source symbol. Claude Shannon and Robert Fano worked on this idea in the late 1940s and developed a couple of ways of encoding source symbols based on their probability distribution. In Shannon’s approach, source symbol i, which is generated with probability p_i, is encoded using a string of $\lceil \log(1/p_i) \rceil$ bits. A code generated by this method is known as a Shannon code. Fano’s approach yields codes with similar lengths, and thus the two approaches are often referred to together as Shannon-Fano coding. We will not go into details of how Fano encoding works. Suffice to say that Shannon coding, or more generally Shannon-Fano coding, comes within one bit of the optimal average source coding rate, which can be shown to be the entropy [25] of the source:

$$H\left(\{p_1,\ldots,p_M\}\right)=\sum_{i=1}^{M}p_i\log\frac{1}{p_i}\qquad\text{bits per source symbol.}$$

It is easy to see that Shannon coding yields an average rate that is close to the entropy. To do so, let l_i denote the number of bits in the string that represents the i th source symbol and write the average rate of the code as the expected value

$$\bar{R}=\sum_{i=1}^{M}p_i l_i=\sum_{i=1}^{M}p_i\left\lceil\log\frac{1}{p_i}\right\rceil$$

This gives the upper bound

$$\bar{R}\leq\sum_{i=1}^{M}p_i\left(\log\frac{1}{p_i}+1\right)=H\left(\{p_1,\ldots,p_M\}\right)+1$$

Shannon codes are not optimal, and the one-bit overhead can be costly when the number of symbols in the source symbol set $\mathcal{S}$ is small. Fortunately, a simple algorithm devised by David Huffman in 1952 can be used to generate uniquely and instantaneously decodable source codes that are optimal in the sense that the average rate is minimised.

To illustrate the operation of the Huffman coding algorithm, consider the source symbol set $\mathcal{S}=\{1,\ldots,6\}$ with corresponding probabilities

$$\begin{array}{ll}p_1=0.3 & p_4=0.1\\ p_2=0.3 & p_5=0.05\\ p_3=0.2 & p_6=0.05\end{array}$$

The algorithm works by constructing a tree with one leaf represented for each source symbol (see Figure 11). With reference to the figure, the algorithm begins with the two least probable symbols, 5 and 6 in this case, and a connection is formed between these. The sum of these probabilities is retained as the probability of either symbol appearing at the source device. The branches corresponding to 5 and 6 are labelled with a 0 and a 1 , respectively. The algorithm proceeds with the next two least probable events, which in this case are the selection of 4 and the selection of the "supersymbol" consisting of 5 and 6. The procedure outlined above is repeated. Once all branches have been labelled with a 0 or a 1 , the code is obtained by reading these values from right to left, following the branches toward each leaf. In the example shown in Figure 11, the mapping becomes

$$\begin{aligned}1&\mapsto 00\\ 2&\mapsto 01\\ 3&\mapsto 10\\ 4&\mapsto 110\\ 5&\mapsto 1110\\ 6&\mapsto 1111.\end{aligned}$$

[25] Note that the entropy is defined as the negative expected value of the logarithm of the source symbol probabilities.

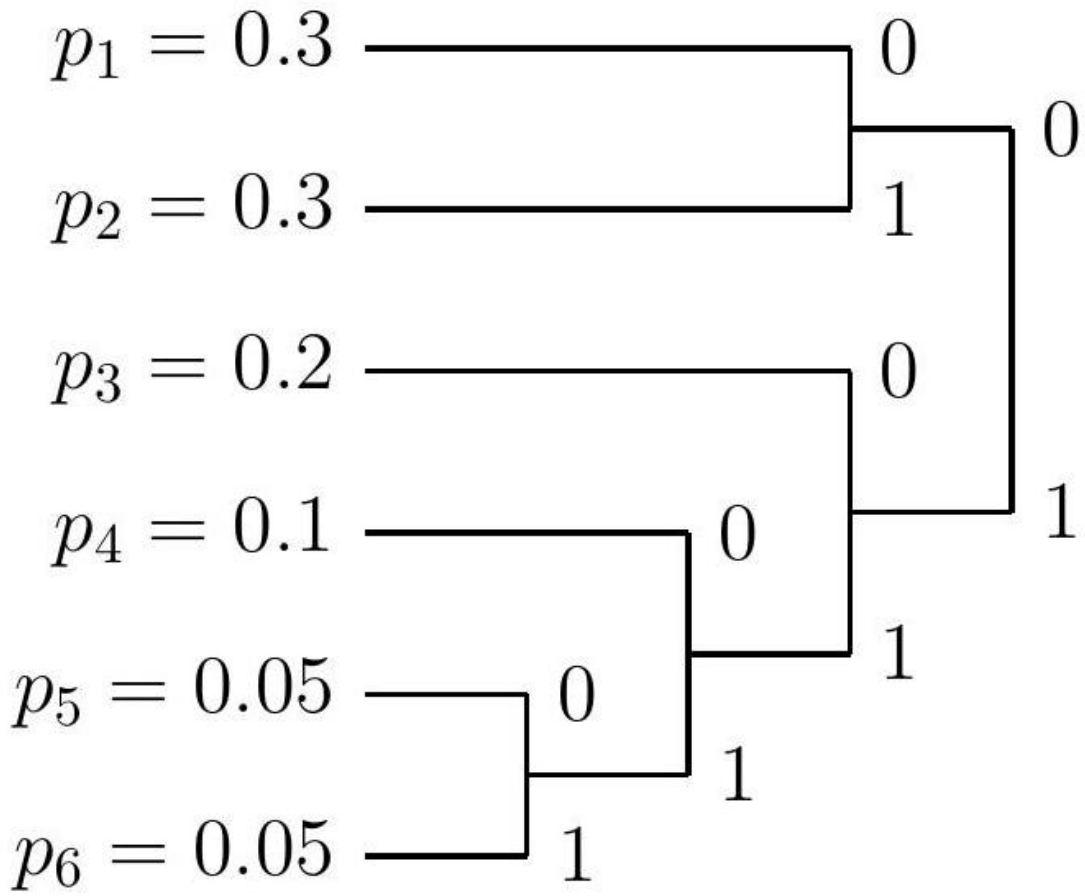


Figure 11: Example of the Huffman algorithm.

One can verify that the entropy for this source is $H(X) = 2.271$ and the average rate is $\bar{R} = 2.3$. It is worth noting that the code generated by the algorithm is not unique; indeed, if we swap around the 0 s and 1 s in Figure 11, we would still obtain an optimal code.

Further information about discrete source coding can be found in Chapter 13 of Sklar and Harris (2021), Chapter 2 of Gallager (2008), and Chapter 9 of Glover and Grant (2009).

3.2.2 Quantisation for Analogue Sequence Sources

Consider an analogue sequence source that generates a sequence of real numbers $s_1, s_2, \ldots$. The aim of a quantiser is to map each of these real numbers to a representation point in the finite set $\{q_1, \ldots, q_M\}$, where each representation point is, itself, a real number. Each representation point corresponds to a quantisation region. Let $\mathcal{R}_i$ denote the i th such region, and let q_i denote the representation point corresponding to $\mathcal{R}_i$. Figure 12 provides an illustration of quantisation regions and representation points. Notice that the regions are intervals in this example. It turns out that this will always be the case, as we will show below.

It is useful to define a measure of accuracy when quantising an analogue sequence. The usual approach is to use the squared error, which is defined

Figure 12: Quantisation regions and representation points.

as

$$e_i(s) = (s - q_i)^2$$

for a given source value s. With this measure of accuracy in mind, two questions arise:

1. Given a set of representation points, how should we define the quantisation regions?

2. Given a set of quantisation regions, how should we choose the representation points?

We answer these two questions in turn below. The approach taken here is general and follows that of Gallager (2008). For further information about quantisation, including practical considerations, you may wish to consult Chapter 13 of Sklar and Harris (2021).

Choosing Quantisation Regions for Given Representation Points

Suppose the values $q_1, \ldots, q_M$ are given to us. Now, suppose we wish to quantise an analogue value s. Whatever the quantisation regions are, it makes sense that they should partition the set of real numbers. In other words, their union should be equal to the reals, and no two regions should overlap. With this in mind, since s is a real number, it must lie in one of the regions.

Our goal is to choose the regions such that the squared error between s and the representation point that we map it to is minimised. Clearly, we can achieve this goal by mapping s to the closest point in the set $\{q_1, \ldots, q_M\}$. This suggests that the quantisation regions should be defined as the intervals with boundaries given by the midpoints between the different representation points. For example, for $1 < i < M$, we should choose the region $\mathcal{R}_i$ to be

$$\mathcal{R}_i = \left[\frac{q_{i-1} + q_i}{2}, \frac{q_i + q_{i+1}}{2}\right).$$

For the left-most and right-most representation points, we have

$$\mathcal{R}_1 = \left(-\infty, \frac{q_1 + q_2}{2}\right)$$

and

$$\mathcal{R}_M = \left[\frac{q_{M-1} + q_M}{2}, \infty\right)$$

Referring back to Figure 12, we see that the example shown adheres to this optimal choice of quantisations regions.

Choosing Representation Points for Given Quantisation Regions

Now, suppose we are given a partition of the real numbers that defines the quantisation regions $\mathcal{R}_1, \ldots, \mathcal{R}_M$, and we wish to select the representation points. In this case, it is helpful to regard the quantisation operation as a function of the source symbol. Indeed, for a given source value s, the quantised version can be written as the function $q(s)$, which just maps s to the correct representation point in the set $\{q_1, \ldots, q_M\}$.

As in the previous section, the goal here should be related to minimising the squared error between the source symbols and their quantised versions. However, in this case, we must consider all possible analogue source values that the sequence $s_1, s_2, \ldots$ can take. A logical approach to this problem is to treat each source value as a random variable (rv) S with density function $f_S(s)$. Now, we can think of the quantisation operation as a function of an rv. Hence, each quantised value Q is an rv, itself. If we wish, we can write the functional dependence between S and Q explicitly, e.g., $Q(S)$.

Based on this probabilistic formulation, we can analyse the average of the squared error over all possible source symbols. This is known as the mean-squared error (MSE) in the quantised source sequence and is written as

$$\text{MSE} = \text{E}\left[(S - Q(S))^2\right] = \int_{-\infty}^{\infty} f_S(s)(s - q(s))^2 \, \text{d}s$$

Partitioning the MSE according to the quantisation regions allows us to write

$$\text{MSE} = \sum_{i=1}^{M} \int_{\mathcal{R}_i} f_S(s) \, (s - q_i)^2 \, \text{d}s$$

Note that this step has permitted us to replace $q(s)$ with the representation point q_i in region $\mathcal{R}_i$ for all i.

Now, for each region, we can apply the law of total probability and the rule of condition probability to write [26]

$$\int_{\mathcal{R}_i} f_S(s)\,(s-q_i)^2\ \mathrm{d}s = \mathrm{P}\,(S \in \mathcal{R}_i) \int_{\mathcal{R}_i} f_i(s)\,(s-q_i)^2\ \mathrm{d}s$$

where $f_i(s)$ is the conditional density of the source symbol given that it is in the region $\mathcal{R}_i$. The probability $\mathrm{P}\,(S \in \mathcal{R}_i)$ does not depend on the representation points, which are the optimisation variables in our problem. Furthermore, the integral on the right-hand side is the expected value of $(S-q_i)^2$ given that S is in $\mathcal{R}_i$, where q_i is a deterministic parameter. Hence, to minimise the MSE , we must choose the representation points to minimise each of these integrals for $i = 1, \ldots, M$. Since the regions of integration are disjoint, we can carry out this minimisation separately for each integral.

Consider the conditional expected value of $(S-q_i)^2$ given $S \in \mathcal{R}_i$. We can write this as

$$\mathrm{E}\left[(S-q_i)^2 \mid S \in \mathcal{R}_i\right] = \mathrm{E}\left[S^2 \mid S \in \mathcal{R}_i\right] - 2q_i\mathrm{E}\left[S \mid S \in \mathcal{R}_i\right] + q_i^2$$

Differentiating with respect to q_i and setting the result equal to zero reveals

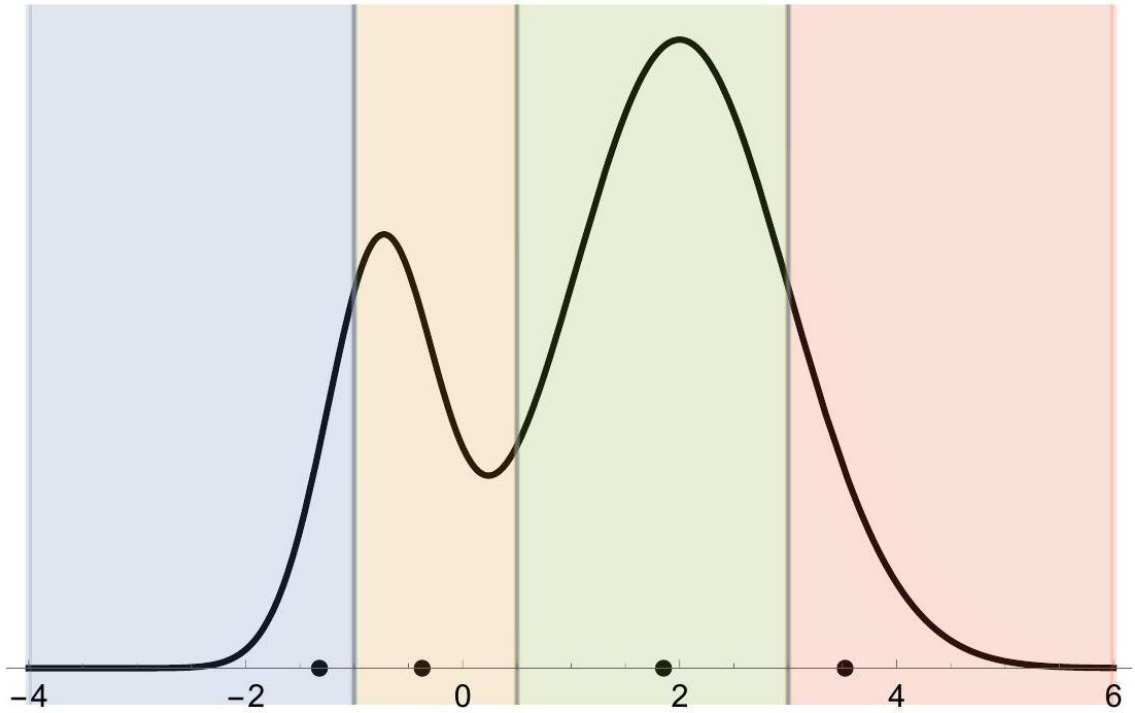


Figure 13: Illustration of the optimal representation points for fixed regions when the source is drawn from a mixture of two Gaussian distributions with means $-3/4$ and 2 and standard deviations $1/2$ and 1 . The mixture ratio is $1:3$, with the Gaussian centred to the left of the origin contributing less.

that the stationary point is

$$q_i = \mathrm{E}\left[S \mid S \in \mathcal{R}_i\right] = \int_{\mathcal{R}_i} s f_i(s)\mathrm{d}s$$

[26] Recall that $\mathrm{P}(A) = \sum_i \mathrm{P}\,(A, B_i)$ when the events $\{B_i\}$ are disjoint. Also, $\mathrm{P}\,(A, B_i) = \mathrm{P}\,(A \mid B_i)\,\mathrm{P}\,(B_i)$. Hence, $\mathrm{P}(A) = \sum_i \mathrm{P}\,(A \mid B_i)\,\mathrm{P}\,(B_i)$. Similar results hold for density functions. Applying these results here yields the stated equation.

which is clearly a minimum, since the quadratic is positive. So the optimal choice of representation points corresponds to the conditional means of the source symbols given that they reside in the disjoint regions.

Figure 13 illustrates this principle for an example where the source symbols are drawn from a Gaussian mixture. The specified regions are indicated in the figure, and the dots that appear on the horizontal axis are the optimal representation points. Observe that points are located in accordance with the bulk of the source distribution in each region.

3.2.3 Sampling for Analogue Waveform Sources

The aim of this section is twofold. First, we will review much of Fourier theory, which will be useful in developing models for digital modulation and demodulation. Second, we will describe how analogue waveform sources can
be sampled to convert them to analogue sequence sources. In doing so, we can invoke the machinery discussed above to encode analogue waveforms as binary strings.

Fourier Series

Recall that a periodic function $x(t)$ with period T can be represented as the (complex) Fourier series

$$x(t) = \sum_{k=-\infty}^{\infty} \tilde{x}_k e^{\mathrm{i}2\pi kt/T}$$

where $\mathrm{i} = \sqrt{-1}$ and $\{\tilde{x}_k\}$ is the set of Fourier coefficients. Also, recall that the l th coefficient is given by

$$\tilde{x}_l = \frac{1}{T} \int_{-T/2}^{T/2} x(t) e^{-\mathrm{i}2\pi lt/T} \, \mathrm{d}t$$

The coefficient formula follows from the properties of the complex exponential function. Specifically, we have that

$$\begin{aligned} \int_{-T/2}^{T/2} e^{\mathrm{i}2\pi kt/T} e^{-\mathrm{i}2\pi lt/T} \, \mathrm{d}t &= \frac{T}{\mathrm{i}2\pi(k-l)} \left(e^{\mathrm{i}\pi(k-l)} - e^{-\mathrm{i}\pi(k-l)} \right) \\ &= T \frac{\sin(\pi(k-l))}{\pi(k-l)} \\ &= T \, \mathrm{sinc}(k-l) \end{aligned}$$

The function $\mathrm{sinc}(k-l)$ is one when the argument is zero, and it is zero for all other integer arguments (Figure 14). To prove eq. (3), we simply substitute eq. (2) into eq. (3) and integrate term by term, using eq. (4) in the process.

Let us pause for a moment to discuss some matters of notation. First, note the definition of the sinc function:

$$\mathrm{sinc}(t) = \frac{\sin(\pi t)}{\pi t}$$

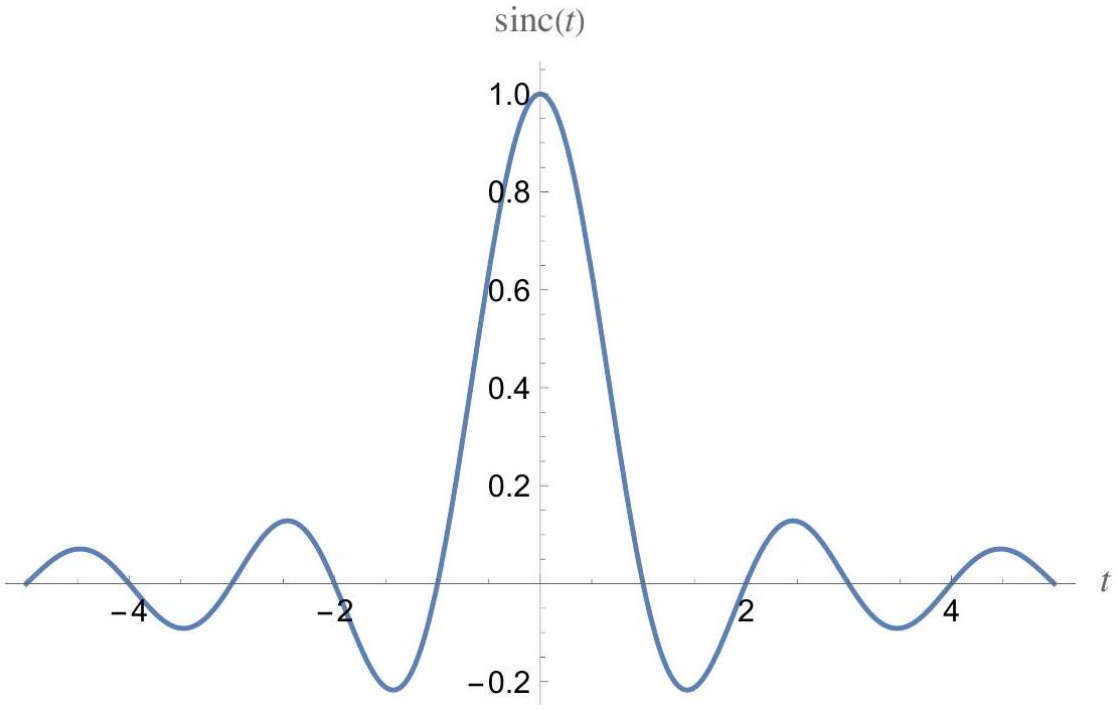


Figure 14: Illustration of sinc(t).

In particular, π is accounted for implicitly. Second, we have embellished the Fourier coefficients with a tilde, e.g., $\tilde{x}_k$. This notation is used to denote quantities that pertain to the frequency domain. Remember, the Fourier series given above is a decomposition of the function $x(t)$ as a sum of weighted sinusoids [27] with different frequencies, where the weight of each sinusoid is a Fourier coefficient.

It is important to note that the Fourier series given in eq. (2) converges pointwise, but not necessarily uniformly, which sometimes gives rise to the so-called Gibbs phenomenon. We will not concern ourselves with this issue in this course. For the interested reader, Gallager (2008) treats this problem rigorously starting in Section 4.3.

Orthogonal Expansions and Time-Limited Waveforms

The property derived in eq. (4) is effectively the orthogonality property of complex exponential functions. In this course, we will typically define two functions, say $\theta_k(t)$ and $\theta_l(t)$, to be orthogonal if

$$\int_{-\infty}^{\infty} \theta_k(t)\theta_l^*(t)\mathrm{d}t = 0, \quad \text{if } k \neq l$$

where the asterisk denotes the complex conjugate. In eq. (4), we observe that the integral is over a finite region. Nevertheless, we can still use the definition stated above to describe the orthogonality of complex exponentials if we first define the rectangular pulse function

$$
\sqcap(t) = \begin{cases} 1, & -1/2 \leq t \leq 1/2 \\ 0, & \text{otherwise} \end{cases}
$$

Now, if we let

$$\theta_k(t) = e^{\mathrm{i}2\pi kt/T} \sqcap (t/T)$$

[27] Of course, the complex exponential can be written as a linear combination of cosine and sine functions.

we see that the functions in the set $\{\theta_k(t)\}$ satisfy eq. (5).
Usually, Fourier series are used in the context of periodic functions. In fact, we can use Fourier series to represent time-limited functions in the following manner. Suppose we are concerned with only one period (centred at the origin) of the function $x(t)$ discussed above. To capture this period, we can write

$$x(t) \sqcap (t/T) = \sum_{k=-\infty}^{\infty} \tilde{x}_k e^{\mathrm{i}2\pi kt/T} \sqcap (t/T) = \sum_{k=-\infty}^{\infty} \tilde{x}_k \theta_k(t)$$

Again, the functions $\{\theta_k(t)\}$ are orthogonal. Hence, we have written the fundamental period of $x(t)$ as a so-called orthogonal expansion. Here, the expansion is in terms of the basis functions $\{\theta_k(t)\}$. The coefficients $\{\tilde{x}_k\}$ are still given by eq. (3), since the integral is limited to the same nonzero interval as the rectangular pulse function.

Notice that the Fourier series provides a one-to-one correspondence between a periodic or time-limited waveform and a sequence of numbers. Hence, if we have such a waveform, we could, in theory, represent it as a sequence of Fourier series coefficients. In this way, we can convert the analogue waveform source to an analogue sequence source.

Because the waveform is time-limited, an infinite number of Fourier coefficients would be needed to perfectly represent the waveform, which is not completely practical. Nevertheless, we could truncate the Fourier series and
quantise the set of coefficients corresponding to the resulting approximate waveform.

Let

$$y(t) = \sum_{k=-\infty}^{\infty} \tilde{y}_k \mathrm{e}^{\mathrm{i}2\pi kt/T} \sqcap (t/T)$$

denote the function that is described by the truncated and quantised series. Typically, most of the $\tilde{y}_k$ coefficients would be zero due to the truncation operation. We are naturally interested in the squared error of this waveform relative to $x(t)$ over the interval $[-T/2, T/2]$. The error is quantified by the difference-energy equation

$$\int_{-T/2}^{T/2} |x(t) - y(t)|^2 \, \mathrm{d}t = T \sum_{k=-\infty}^{\infty} |\tilde{x}_k - \tilde{y}_k|^2$$

where the right-hand side follows from the orthogonality of the complex exponential functions over the interval in question. Since this equation is the sum of squared errors over all coefficients, we can invoke the same quantisation analysis that we did earlier to design the quantiser and to study the accuracy of the method.

Fourier Transforms

Another way to study the frequency properties of waveforms that are not necessarily periodic is to utilise the Fourier transform. This transform maps a function of time to a function of frequency. We will use natural frequency [28] in this course, and thus the Fourier transform and its inverse are defined as follows:

$$\tilde{x}(f) = \int_{-\infty}^{\infty} x(t) e^{-\mathrm{i}2\pi ft} \, \mathrm{d}t$$
$$x(t) = \int_{-\infty}^{\infty} \tilde{x}(f) e^{\mathrm{i}2\pi ft} \, \mathrm{d}f$$

Recall the following properties of Fourier transforms:

[28] The unit of natural frequency is a cycle per second, or hertz (Hz). It is related to angular frequency ω through the relation $\omega = 2\pi f$ where f is the frequency in Hz .

$$
\begin{aligned}
ax(t) + by(t) &\leftrightarrow a\tilde{x}(f) + b\tilde{y}(f) && \text{linearity,}\\
x^*(-t) &\leftrightarrow \tilde{x}^*(f) && \text{conjugation,}\\
\tilde{x}(t) &\leftrightarrow x(-f) && \text{duality,}\\
x(t-\tau) &\leftrightarrow e^{-\mathrm{i}2\pi f\tau}\tilde{x}(f) && \text{time shift,}\\
e^{\mathrm{i}2\pi f_0 t}x(t) &\leftrightarrow \tilde{x}\,(f-f_0) && \text{frequency shift,}\\
x(t/T) &\leftrightarrow \tilde{x}(fT) && \text{scaling } (T>0),\\
\mathrm{d}x/\mathrm{d}t &\leftrightarrow \mathrm{i}2\pi f\tilde{x}(f) && \text{differentiation,}\\
\int_{-\infty}^{\infty} x(\tau)y(t-\tau)\mathrm{d}\tau &\leftrightarrow \tilde{x}(f)\tilde{y}(f) && \text{convolution,}\\
\int_{-\infty}^{\infty} x(\tau)y^*(\tau-t)\mathrm{d}\tau &\leftrightarrow \tilde{x}(f)\tilde{y}^*(f) && \text{correlation,}\\
x(t)y(t) &\leftrightarrow \int_{-\infty}^{\infty}\tilde{x}(u)\tilde{y}(f-u)\mathrm{d}u && \text{multiplication.}
\end{aligned}
$$

These properties can be used to derive the following very useful consequences.

Parseval's Theorem Suppose we have a function $z(t)$ that is the correlation of two other functions $x(t)$ and $y(t)$. By the definition of the inverse transform, we have that

$$z(0) = \int_{-\infty}^{\infty}\tilde{z}(f)\mathrm{d}f$$

Furthermore, by the definition of correlation, we have that

$$z(0) = \int_{-\infty}^{\infty}x(\tau)y^*(\tau)\mathrm{d}\tau$$

It follows from the correlation property of Fourier transforms that

$$\int_{-\infty}^{\infty}x(t)y^*(t)\mathrm{d}t = \int_{-\infty}^{\infty}\tilde{x}(f)\tilde{y}^*(f)\mathrm{d}f$$

This is known as Parseval's theorem.

Energy Equation and Spectral Density The following interesting corollary of Parseval's theorem follows by letting $y = x$:

$$\int_{-\infty}^{\infty}|x(t)|^2\ \mathrm{d}t = \int_{-\infty}^{\infty}|\tilde{x}(f)|^2\ \mathrm{d}f$$

This is known as the energy equation. It states that the energy of a function in the time domain is equal to the energy of the function in the frequency domain. The function $|\tilde{x}(f)|^2$ is called the (energy) spectral density of $x(t)$. It quantifies the energy contained in $x(t)$ per unit of frequency.

Orthogonality Finally, suppose $x(t)$ and $y(t)$ are orthogonal under the definition given by eq. (5). Then Parseval's theorem guarantees that the Fourier transforms of these functions are also orthogonal. In other words

$$\int_{-\infty}^{\infty}x(t)y^*(t)\mathrm{d}t = 0 \quad \text{if and only if} \quad \int_{-\infty}^{\infty}\tilde{x}(f)\tilde{y}^*(f)\mathrm{d}f = 0.$$

Let us look at an important example in relation to this orthogonality result. Consider the set of functions given by [29]

$$\theta_k(t) = e^{\mathrm{i}2\pi kt/T}\sqcap(t/T), \quad k\in\mathbb{Z}$$

We studied these earlier in the context of Fourier series representations of time-limited functions, and we have already shown that these functions are orthogonal to each other. Moreover, they collectively form an orthogonal basis for a function confined to the interval $[-T/2, T/2]$.

By using the scaling and frequency shift properties, we can calculate the Fourier transform of $\theta_k(t)$ to be

$$\tilde{\theta}_k(f) = T\,\mathrm{sinc}(fT - k), \quad k \in \mathbb{Z}$$

By Parseval's theorem, the functions in the set $\left\{\tilde{\theta}_k(f)\right\}$ are orthogonal. These

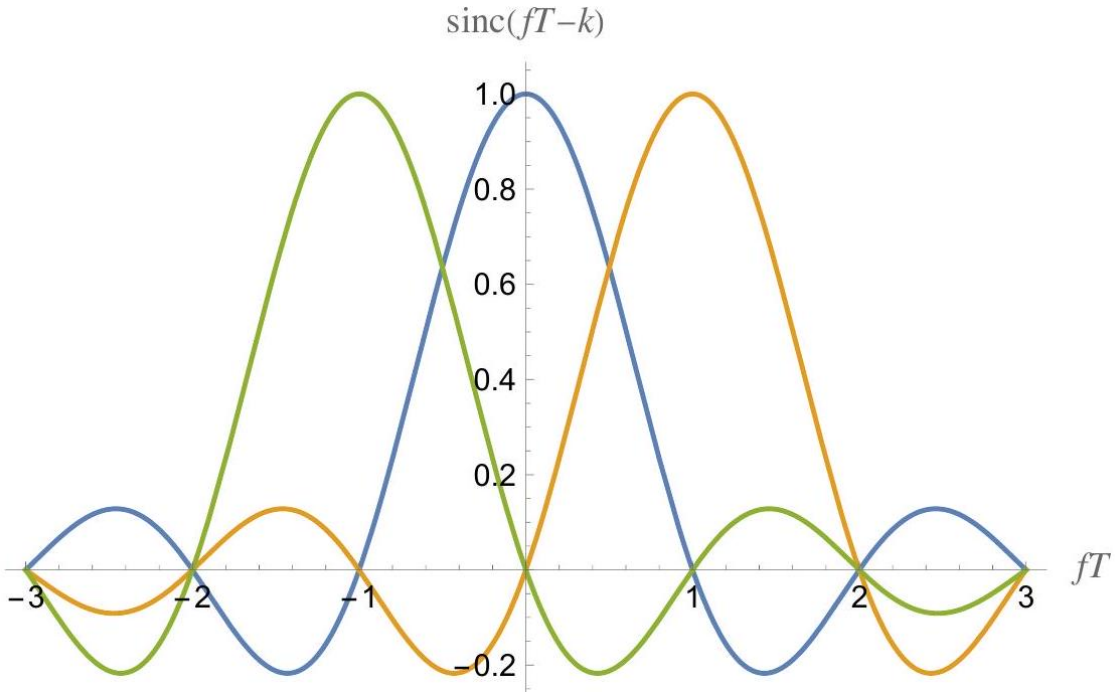


Figure 15: Illustration of $\mathrm{sinc}(fT), \mathrm{sinc}(fT-1)$, and $\mathrm{sinc}(fT+1)$.

are scaled sinc functions shifted by integer values.

Figure 15 depicts three of these orthogonal functions for $T = 1$. Note that for integer values of the argument, only one of these functions is nonzero. We will return to this observation below.

The DTFT and Band-Limited Waveforms

Above, we discussed a way to use a Fourier series to represent a time-limited function. This effectively required us to multiply the series that represents a periodic function by a rectangular pulse that is limited to the interval $[-T/2, T/2]$. This interval is not special. Indeed, we could easily shift the time window and obtain a similar result.

Now, let us consider the time-frequency dual of this problem. Specifically, suppose we have a band-limited function $\tilde{x}(f)$ that is nonzero only in the frequency interval $[-W, W]$. The discrete-time Fourier transform (DTFT) of $\tilde{x}(f)$ is defined as

$$\tilde{x}(f) = \sum_{k=-\infty}^{\infty} x_k e^{-\mathrm{i}2\pi k f/2W} \sqcap (f/2W)$$

Note that this series is basically the same as the Fourier series, but where
we have substituted f for $t, 2W$ for T, and x_k for $\tilde{x}_k$. We have also taken the conjugate of the exponential function. The coefficients are thus readily obtained via the equation

$$x_l = \frac{1}{2W} \int_{-W}^{W} \tilde{x}(f) e^{\mathrm{i}2\pi l f/2W}\, \mathrm{d}f.$$

[29] Here, $\mathbb{Z}$ is the set of integers.

The form of the coefficient integral resembles an inverse Fourier transform. Hence, the coefficients relate to the temporal properties of the function $\tilde{x}(f)$. As such, no tilde is used to embellish the DTFT coefficients.

In an analogous manner to the Fourier series, the DTFT provides a one-toone correspondence between a band-limited function and a sequence of numbers. Hence, we can use the DTFT to convert the analogue waveform source to an analogue sequence source. Because the waveform is band-limited, an infinite number of DTFT coefficients would be needed to perfectly encode the waveform. Again, we could truncate the DTFT to obtain a suitable approximation.

As an aside, a function can be time-limited or band-limited, but not both. This curious fact is actually a manifestation of the uncertainty principle, which relates to systems described by two so-called conjugate variables, which are variables related by a Fourier transform relation, time and frequency in this case. The uncertainty principle states that if one has certain knowledge of the value of one quantity, knowledge of the value of the other quantity must necessarily be absent. As two examples, consider the Fourier transform pairs

$$\sqcap(t) \leftrightarrow \operatorname{sinc}(f)$$

and

$$1 \leftrightarrow \delta(f)$$

where $\delta(\cdot)$ is the Dirac delta function. In the first case, the time-domain function is time-limited, and the transform is not limited in frequency. In the second case, the transform is band-limited, and the inverse transform is not
limited in time.

The Sampling Theorem

Now, let us return to the problem of representing an analogue waveform as a sequence. We defined the DTFT in eq. (7). We can equally write the DTFT as

$$\tilde{x}(f) = \sum_{k=-\infty}^{\infty} x_k \tilde{\phi}_k(f)$$

where

$$\tilde{\phi}_k(f) = e^{-\mathrm{i}2\pi k f/2W} \sqcap (f/2W)$$

Note that $\left\{\tilde{\phi}_k(f)\right\}$ is a set of orthogonal functions.
It is reasonable to compute the inverse transform of $\tilde{x}(f)$ in order to observe the properties of the waveform in the time domain. Assuming we can integrate each term in the series separately [30], the inverse transform can be expressed as

$$x(t) = \sum_{k=-\infty}^{\infty} x_k \phi_k(t)$$

where

$$\phi_k(t) = 2W \operatorname{sinc}(2Wt - k)$$

is the inverse Fourier transform of $\tilde{\phi}_k(f)$. Again, by Parseval's theorem and properties of the Fourier transform, the functions in the set $\{\phi_k(t)\}$ are orthogonal to one another.

To summarise, we have shown that a band-limited waveform can be represented as a series of time-shifted sinc functions, each weighted by a DTFT coefficient:

$$x(t) = \sum_{k=-\infty}^{\infty} 2W x_k \operatorname{sinc}(2Wt - k)$$

Notice that when $2Wt - k$ is an integer, only one of the sinc functions in the

series will be nonzero. Using this fact, we have that

$$x\left(\frac{l}{2W}\right) = 2Wx_l$$

and thus

$$x(t) = \sum_{k=-\infty}^{\infty} x\left(\frac{k}{2W}\right) \text{sinc}(2Wt - k)$$

Eq. (8) is the sampling equation. It shows that we can represent a band-limited waveform by first sampling that waveform at intervals of $T = 1/2$ W seconds (equivalently, at a rate of 2 W samples per second), and then constructing a set of shifted sinc functions, each weighted by the corresponding sample. We can also write the sampling equation in terms of the sampling interval T as follows:

$$x(t) = \sum_{k=-\infty}^{\infty} x(kT) \,\text{sinc}\left(\frac{t}{T} - k\right)$$

In practice, it is impossible to represent the waveform $x(t)$ perfectly via the sampling equation. This is primarily because real waveforms are causal; they must have a beginning, and our observation of the waveform must not depend upon its future structure. However, the sinc functions in the sampling equation are noncausal. To overcome this issue, we can truncate the sinc functions and delay them in time such that the waveform does, indeed, have a beginning. The sinc function decays like $1/t$, so we may have to implement a large window for the truncation and a correspondingly large delay, which could be undesirable. Nevertheless, similar methods are used in practice.

Aliasing (Review)

The DTFT-based presentation of the sampling theorem that is covered here may seem quite different to what you have learned in courses involving timefrequency analysis. Most students will be familiar with the model of sampling whereby a sampled signal is obtained by multiplying the analogue waveform
by an impulse train with spacing T, then using the multiplication property of Fourier transforms (see the list of properties above) to analyse the sampled signal in the frequency domain. This approach leads to the idea that sampling at less than a rate of $2W$ leads to aliasing.

Although this approach to studying the sampling problem is nice and simple, there are some drawbacks. The first drawback is with the model. The sampling process in practice cannot be represented through the use of an impulse train. Indeed, Dirac impulses are mathematical objects that cannot be replicated perfectly in a physical system. To overcome this problem, one could replace the impulse train with a pulse train. The pulse duration should be made to be short so as to replicate a sequence of impulses. The calculations that ensue are still straightforward, but the additional detail muddies the picture a little. Sklar and Harris (2021) present a nice, simple exposition of this topic; see Section 2.4.1 of that text for details.

The second drawback is that the impulse train model does not provide a direct link to orthogonal expansions of analogue waveforms. It does not fit in nicely with the study of time-limited waveforms, either. It stands alone and may leave one wondering if there is a more systematic approach to encoding analogue waveform sources. The Fourier series and DTFT framework we have outlined here provides such a systematic theory, from which practical methods can be developed (e.g., truncation and quantisation of coefficients). Finally, we will see that orthogonal expansions feature heavily in digital modulation and demodulation theory, and virtually all digital communication systems use some form of this theory in practice.

3.3 Channel Coding

[30] Technically, we have to show that the DTFT satisfies certain properties in order to integrate term by term when we compute the inverse Fourier transform. Not all functions satisfy these properties, but we will not concern ourselves with such matters in this course.

Chapter 4

Sampling Theory

Ideal Sampling

Chapter 5

Baseband Modulation and Demodulation

5.1 Baseband Modulation and Demodulation

In this section, we will develop a general framework for encoding bit strings in waveforms that can be transmitted over a physical channel (see Figure 6). For now, we will be concerned with a transmission that is limited to frequencies between zero and $W > 0$ Hz; hence, the encoding process is referred to as baseband modulation. The approach we will take is built on the concept of orthogonal expansions. We will also use the framework we develop to construct the inverse process of baseband demodulation, which is carried out at the receiver. Finally, we will give details of how a dispersive channel can introduce intersymbol interference (ISI), and we will derive Nyquist's criterion for achieving zero ISI when transmitting through such a channel.

First, we review vector spaces and inner products, and we describe a general method for constructing orthogonal expansions. We will soon see that this vector formalism will come in useful when designing and analysing modulation and demodulation techniques.

5.1.1 Vector Spaces and Inner Product Spaces

Recall that a vector space $\mathcal{V}$ is comprised of a set of elements called vectors that obey the following laws:

1. Addition laws: For every $\boldsymbol{v} \in \mathcal{V}$ and $\boldsymbol{u} \in \mathcal{V}$, there exists unique vector $\boldsymbol{v} + \boldsymbol{u} \in \mathcal{V}$ satisfies the following:
   (a) Commutativity: $\boldsymbol{v} + \boldsymbol{u} = \boldsymbol{u} + \boldsymbol{v}$;
   (b) Associativity: $\boldsymbol{v} + (\boldsymbol{u} + \boldsymbol{w}) = (\boldsymbol{v} + \boldsymbol{u}) + \boldsymbol{w}$ for every $\boldsymbol{w} \in \mathcal{V}$;
   (c) Zero: there exists a unique element $\boldsymbol{0} \in \mathcal{V}$ such that $\boldsymbol{0} + \boldsymbol{v} = \boldsymbol{v}$;
   (d) Negation: there exists a unique element $-\boldsymbol{v}$ with $\boldsymbol{v} + (-\boldsymbol{v}) = \boldsymbol{0}$.

2. Scalar multiplication laws: For each scalar α and every $\boldsymbol{v} \in \mathcal{V}$, there exists a vector $\alpha\boldsymbol{v} \in \mathcal{V}$ that satisfies the following:
   (a) Scalar associativity: $\alpha(\beta\boldsymbol{v}) = (\alpha\beta)\boldsymbol{v}$ for each scalar β;
   (b) Scalar identity: the unit scalar 1 satisfies $1\boldsymbol{v} = \boldsymbol{v}$.

5.2 Distributive laws:

(a) for all scalars α and all $\boldsymbol{v}, \boldsymbol{u} \in \mathcal{V}, \alpha(\boldsymbol{v} + \boldsymbol{u}) = \alpha\boldsymbol{v} + \alpha\boldsymbol{u}$;
(b) for all scalars α and β and all $\boldsymbol{v} \in \mathcal{V}, (\alpha + \beta)\boldsymbol{v} = \alpha\boldsymbol{v} + \beta\boldsymbol{v}$.

Let $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n$ be vectors in a vector space $\mathcal{V}$. If for any $\boldsymbol{w} \in \mathcal{V}$, we can write $\boldsymbol{w}$ as a linear combination of the vectors in $\{\boldsymbol{v}_i\}$, then we say that $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n$ spans $\mathcal{V}$. If, in addition, the vectors in $\{\boldsymbol{v}_i\}$ are linearly independent - i.e., it is not possible to express any of these vectors as a linear

combination of the others - then we say that $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n$ is a basis for $\mathcal{V}$. If $n < \infty$, then we say that $\mathcal{V}$ is finite dimensional, otherwise, we say that $\mathcal{V}$ is infinite dimensional.

In our quest to develop a theory of modulation that relies on the notion of orthogonality, we will require a way to quantify the angle between two vectors as well as the length of a vector. To do this, let us introduce the inner product operation. An inner product is a function of two vectors $\boldsymbol{v}, \boldsymbol{u} \in \mathcal{V}$. It is represented using the notation $\langle \boldsymbol{v}, \boldsymbol{u} \rangle$, and it satisfies the following laws:

1. Conjugate (Hermitian) symmetry: $\langle \boldsymbol{v}, \boldsymbol{u} \rangle = \langle \boldsymbol{u}, \boldsymbol{v} \rangle^*$;

2. Conjugate (Hermitian) bilinearity: $\langle \alpha \boldsymbol{v} + \beta \boldsymbol{u}, \boldsymbol{w} \rangle = \alpha \langle \boldsymbol{v}, \boldsymbol{w} \rangle + \beta \langle \boldsymbol{u}, \boldsymbol{w} \rangle$;

3. Positivity: $\langle \boldsymbol{v}, \boldsymbol{v} \rangle \geq 0$, with equality if and only if $\boldsymbol{v} = \boldsymbol{0}$.

A vector space equipped with an inner product is called an inner product space. For finite-dimensional complex vectors comprised of n elements, the inner product is defined as

$$\langle \boldsymbol{v}, \boldsymbol{u} \rangle = \sum_{i=1}^{n} v_i u_i^*$$

where v_i is the i th element of the n-tuple [31] $\boldsymbol{v} = (v_1, \ldots, v_n)$. This definition resembles the familiar dot product that you may have seen in basic courses on linear algebra. The notable difference is the appearance of the conjugate operator; although, some texts define the dot product for complex vectors in the same way.

We can obtain the length of a vector by utilising the formulae given above. More precisely, we call the length of a vector its norm, and we use the notation $\|\boldsymbol{v}\|$ to signify this quantity. In general, the norm is given by

$$\|\boldsymbol{v}\| = \sqrt{\langle \boldsymbol{v}, \boldsymbol{v} \rangle}$$

For the finite-dimensional case, we have

$$\|\boldsymbol{v}\| = \sqrt{\sum_{i=1}^{n} |v_i|^2}$$

The use of the norm allows us to express the inner product of two vectors in the following familiar form:

$$\langle \boldsymbol{v}, \boldsymbol{u} \rangle = \|\boldsymbol{v}\| \|\boldsymbol{u}\| \cos \theta_{\boldsymbol{v},u}$$

where $\theta_{\boldsymbol{v},\boldsymbol{u}}$ is the angle between $\boldsymbol{v}$ and $\boldsymbol{u}$. If the two vectors are perpendicular, the angle is $\pi/2$, and we see that $\langle \boldsymbol{v}, \boldsymbol{u} \rangle = 0$. In this case, $\boldsymbol{v}$ and $\boldsymbol{u}$ are said to be orthogonal. For orthogonal vectors, we can also write

$$\langle \boldsymbol{v}, \boldsymbol{u} \rangle = \sum_{i=1}^{n} v_i u_i^* = 0$$

As an example of two orthogonal vectors, we might choose $\boldsymbol{v} = (1, 1, 1, 1)$ and $\boldsymbol{u} = (1, 1, -1, -1)$. It may not be readily apparent that $\theta_{\boldsymbol{v},\boldsymbol{u}} = \pi/2$ in this example, but it is easy to see that these vectors are orthogonal by using the summation formula.

[31] An n-tuple is a group of n elements.

5.2.1 Orthonormal Expansions

In Section 2, we showed that periodic and time-limited functions can be represented by a Fourier series, which is an orthogonal expansion. We also showed that band-limited functions can be represented by a DTFT, which is another orthogonal expansion. We discussed how one might use orthogonal expansions to sample analogue waveforms prior to quantisation and discrete source encoding. The process of modulation can be thought of similarly, but instead of representing a continuous function by a discrete set of values, we will first map bit strings to discrete values and subsequently construct a continuous waveform by using these values as weights in an expansion.

Here, we will study the orthogonal expansion principle using the language of vectors. We will discuss projections of vectors, which will lead to the important Cauchy-Schwarz inequality. We will then discuss how to construct sets of (normalised) orthogonal vectors using the Gram-Schmidt process. We will conclude the section with an explanation of how we might generalise the vector space formalism to cover function spaces. This last step will present us with the tools we need to formally develop a theory of modulation.

Orthogonal and Orthonormal Expansions

We have defined the orthogonality of vectors as the case where their inner product is zero. Suppose we have a set of orthogonal vectors $\{\boldsymbol{v}_i\}$ that form a basis for the vector space $\mathcal{V}$. We can write an arbitrary vector $\boldsymbol{w} \in \mathcal{V}$ as a linear combination of these vectors, i.e.,

$$\boldsymbol{w} = \sum_i \alpha_i \boldsymbol{v}_i$$

for some scalars $\{\alpha_i\}$. This is an orthogonal expansion of $\boldsymbol{w}$.
Vectors that satisfy the orthogonality condition and also have unit length are called orthonormal. For any set of orthogonal vectors $\{\boldsymbol{v}_i\}$, we can obtain
a set of orthonormal vectors $\{\phi_i\}$ by defining

$$\boldsymbol{\phi}_i = \frac{1}{\|\boldsymbol{v}_i\|}\boldsymbol{v}_i, \quad \text{for all } i.$$

If $\phi_1, \phi_2, \ldots$ are a basis for the vector space $\mathcal{V}$, then we can write the orthonormal expansion of an arbitrary vector $\boldsymbol{w} \in \mathcal{V}$ as

$$\boldsymbol{w} = \sum_i \beta_i \boldsymbol{\phi}_i$$

for some scalars $\{\beta_i\}$. The only difference between an orthogonal expansion and an orthonormal expansion lies in the scale of each component vector.

As an example, consider the space $\mathbb{R}^n$, which is the space of n-dimensional real vectors. The standard basis for this space is

$$\begin{aligned}
\boldsymbol{e}_1 &= (1, 0, 0, \ldots, 0) \\
\boldsymbol{e}_2 &= (0, 1, 0, \ldots, 0) \\
&\vdots \\
&\vdots \\
\boldsymbol{e}_n &= (0, 0, 0, \ldots, 1)
\end{aligned}$$

Hence, we can write any vector $\boldsymbol{w} \in \mathbb{R}^n$ as

$$\boldsymbol{w} = \sum_i w_i \boldsymbol{e}_i = (w_1, \ldots, w_n)$$

We are usually more familiar with the simple notation used on the righthand side of this set of equations rather than the clunky summation formula. Yet, it is important to recognise that when we write the simple representation $\boldsymbol{w} = (w_1, \ldots, w_n)$, we are implicitly making use of the standard

orthonormal expansion for $\mathbb{R}^n$. This example also serves to illustrate the point that the number of elements contained in $\boldsymbol{w}$ is the same as the number of orthonormal basis vectors, which specifies the dimension of the space.

Projection

Recall that the projection of a vector $\boldsymbol{v}$ onto a vector $\boldsymbol{u}$ is given by

$$\boldsymbol{v}_{\parallel \boldsymbol{u}} = \left\langle \boldsymbol{v}, \frac{\boldsymbol{u}}{\|\boldsymbol{u}\|} \right\rangle \frac{\boldsymbol{u}}{\|\boldsymbol{u}\|}$$

Note that the projection only retains the direction of $\boldsymbol{u}$, not the magnitude. In this way, the projection of v onto a unit vector ϕ is given by

$$\boldsymbol{v}_{\parallel \phi} = \langle \boldsymbol{v}, \boldsymbol{\phi} \rangle \boldsymbol{\phi}.$$

Staying with the example of projecting onto a unit vector, the original vector $\boldsymbol{v}$ can be written as the sum of this projection and the remaining component, which we call $\boldsymbol{v}_{\perp \phi}$, i.e.,

$$\boldsymbol{v} = \boldsymbol{v}_{\parallel \phi} + \boldsymbol{v}_{\perp \phi}$$

Rewriting the remaining component in terms of the other two, we can show that it is orthogonal to ϕ, which justifies the use of the $\perp$ subscript:

$$\begin{aligned}
\langle \boldsymbol{v}_{\perp \phi}, \boldsymbol{\phi} \rangle &= \langle \boldsymbol{v} - \boldsymbol{v}_{\parallel \boldsymbol{\phi}}, \boldsymbol{\phi} \rangle \\
&= \langle \boldsymbol{v}, \boldsymbol{\phi} \rangle - \langle \boldsymbol{v}_{\parallel \phi}, \boldsymbol{\phi} \rangle \\
&= \langle \boldsymbol{v}, \boldsymbol{\phi} \rangle - \langle \langle \boldsymbol{v}, \boldsymbol{\phi} \rangle \boldsymbol{\phi}, \boldsymbol{\phi} \rangle \\
&= \langle \boldsymbol{v}, \boldsymbol{\phi} \rangle - \langle \boldsymbol{v}, \boldsymbol{\phi} \rangle \|\boldsymbol{\phi}\|^2 \\
&= \langle \boldsymbol{v}, \boldsymbol{\phi} \rangle - \langle \boldsymbol{v}, \boldsymbol{\phi} \rangle \\
&= 0
\end{aligned}$$

The Cauchy-Schwarz Inequality

The results on projections and orthonormality lead directly to the celebrated Cauchy-Schwarz inequality. This inequality will be extremely useful when we design a general rule for demodulating digital waveforms later.

Theorem 3.1 (Cauchy-Schwarz inequality). Let $\boldsymbol{v}$ and $\boldsymbol{u}$ be two vectors in an inner product space. Then

$$|\langle \boldsymbol{v}, \boldsymbol{u} \rangle| \leq \|\boldsymbol{v}\| \|\boldsymbol{u}\|$$

Equality is met when $\boldsymbol{v}$ is proportional to $\boldsymbol{u}$.

Proof. Suppose $\boldsymbol{u} \neq \boldsymbol{0}$, since the inequality is immediate otherwise. When $\boldsymbol{v}$ is proportional to $\boldsymbol{u}$, equality in the relation can be readily observed.

Now, suppose $\boldsymbol{v}$ and $\boldsymbol{u}$ are not proportional. By the Pythagorean theorem, we know that

$$\|\boldsymbol{v}\|^2 = \left\|\boldsymbol{v}_{\parallel \boldsymbol{u}} + \boldsymbol{v}_{\perp \boldsymbol{u}}\right\|^2 = \left\|\boldsymbol{v}_{\parallel \boldsymbol{u}}\right\|^2 + \|\boldsymbol{v}_{\perp \boldsymbol{u}}\|^2$$

This is just the vector version of $c^2 = a^2 + b^2$ for a right triangle. Now, since the norm of a vector is nonnegative, we can write

$$\|\boldsymbol{v}\|^2 \geq \left\|\boldsymbol{v}_{\parallel \boldsymbol{u}}\right\|^2 = \left\| \left\langle \boldsymbol{v}, \frac{\boldsymbol{u}}{\|\boldsymbol{u}\|} \right\rangle \frac{\boldsymbol{u}}{\|\boldsymbol{u}\|} \right\|^2 = \frac{|\langle \boldsymbol{v}, \boldsymbol{u} \rangle|^2}{\|\boldsymbol{u}\|^4} \|\boldsymbol{u}\|^2 = \frac{|\langle \boldsymbol{v}, \boldsymbol{u} \rangle|^2}{\|\boldsymbol{u}\|^2}$$

which is the inequality stated in the theorem.

The Gram-Schmidt Process

Until now, our treatment of orthonormal expansions has been on the understanding that an orthonormal basis exists. This may seem like a big assumption. However, the Gram-Schmidt process can be used to find such a basis.

The process works in the following manner. Suppose we wish to construct an n-dimensional orthonormal basis for the n-dimensional vector space $\mathcal{V}$. We start with n linearly independent vectors $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n \in \mathcal{V}$. Note that these are not necessarily orthogonal. We normalise the first vector to obtain the unit vector

$$\boldsymbol{\phi}_1 = \frac{1}{\|\boldsymbol{v}_1\|}\boldsymbol{v}_1$$

This vector spans the one-dimensional subspace $\mathcal{V}_1 \subseteq \mathcal{V}$ spanned by $\boldsymbol{v}_1$. Now, we find a vector that is orthogonal to this subspace by projecting $\boldsymbol{v}_2$ onto ϕ_1 and subtracting from $\boldsymbol{v}_2$, which gives

$$(\boldsymbol{v}_2)_{\perp\mathcal{V}_1} = \boldsymbol{v}_2 - \langle \boldsymbol{v}_2, \boldsymbol{\phi}_1 \rangle \boldsymbol{\phi}_1$$

This vector is then normalised to obtain

$$\boldsymbol{\phi}_2 = \frac{1}{\left\|(\boldsymbol{v}_2)_{\perp\mathcal{V}_1}\right\|} (\boldsymbol{v}_2)_{\perp\mathcal{V}_1}$$

The two orthonormal vectors span the subspace $\mathcal{V}_2 \subseteq \mathcal{V}$ spanned by $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$

The process is repeated to find all orthonormal vectors that together span $\mathcal{V}$. Specifically, the k th orthogonal vector is obtained by projecting $\boldsymbol{v}_k$ onto the subspace $\mathcal{V}_{k-1} \subseteq \mathcal{V}$ and subtracting from $\boldsymbol{v}_k$:

$$(\boldsymbol{v}_k)_{\perp\mathcal{V}_{k-1}} = \boldsymbol{v}_k - \sum_{i=1}^{k-1} \langle \boldsymbol{v}_k, \boldsymbol{\phi}_i \rangle \boldsymbol{\phi}_i$$

This is then normalised to obtain

$$\boldsymbol{\phi}_k = \frac{1}{\left\|(\boldsymbol{v}_k)_{\perp\mathcal{V}_{k-1}\|}\right\|} (\boldsymbol{v}_k)_{\perp\nu_{k-1}}$$

Inner Product Spaces of Functions

All results developed above have pertained to the abstract notion of a vector space. Students typically find these concepts to be most easily understood by considering examples in the n-dimension real or complex spaces, denoted by $\mathbb{R}^n$ and $\mathbb{C}^n$, respectively. For modulation, we wish to develop a theory of orthonormal expansions where functions are used as basis vectors.

Remember, the term vector is a relatively abstract concept here. A vector is just an element of a vector space, which is defined using the laws given previously. There is no reason that we cannot ask whether some classes of functions, viewed as vectors, might obey these laws. Indeed, it may be a helpful exercise to replace the vector notation $\boldsymbol{v}$ with the function $v(t)$ in the vector space axioms given previously and verify that the axioms hold. A specific class of functions that can be shown to satisfy these axioms is the class of functions with finite energy, i.e., $\int_{-\infty}^{\infty} |v(t)|^2 \, \mathrm{d}t < \infty$.

To begin our development of the theory of orthonormal expansions for function spaces, we will require a definition of the inner product for functions. All other results developed above will follow relatively easily once this definition is established. Specifically, we choose the inner product of two functions $v(t)$ and $u(t)$ to be

$$\langle v, u \rangle = \int_{-\infty}^{\infty} v(t) u^*(t) \mathrm{d}t$$

It is straightforward to see that the first two laws governing inner products (symmetry and bilinearity) are satisfied by this integral formula. The third law (positivity) is satisfied for well-behaved functions, but there are some technical issues that may be encountered, which we will not concern ourselves with in this course (see Section 5.2.3 of Gallager (2008) for details).

Regarding the matter of notation, observe that we omit the independent variable t from the left-hand side of the inner product expression when referring to functions. This is because the inner product operates on the functions, themselves (i.e., the dependent variables); the specific letter used for the independent variable is irrelevant. Some texts continue to use the vector notation $\langle \boldsymbol{v}, \boldsymbol{u} \rangle$ for functions. We will soon show that this is equally valid notation, so one should not get hung up on details.

The definition of the norm of a function follows immediately from the inner product integral:

$$\|v\| = \sqrt{\int_{-\infty}^{\infty} |v(t)|^2 \, \mathrm{d}t}$$

Furthermore, two functions $v(t)$ and $u(t)$ are said to be orthogonal if their inner product is zero:

$$\langle v, u \rangle = \int_{-\infty}^{\infty} v(t) u^*(t) \mathrm{d}t = 0$$

Note that this is exactly the definition that we gave in eq. (5).

As we saw with Fourier series and DTFTs, we can often write a function $w(t)$ as an orthogonal expansion, i.e.,

$$w(t) = \sum_i \alpha_i v_i(t)$$

where all pairs of distinct functions in the set $\{v_i(t)\}$ satisfy the orthogonality criterion. If, in addition, each of these functions has unit norm, then the summation formula given above constitutes an orthonormal expansion. The Fourier series is an orthogonal expansion, but not an orthonormal expansion.

We can project one function onto another just as we described for vectors, but with the new inner product and norm definitions. As a result, the CauchySchwarz inequality applies to functions $v(t)$ and $u(t)$ that belong to an inner product space and is written as

$$\left| \int_{-\infty}^{\infty} v(t) u^*(t) \mathrm{d}t \right| \leq \sqrt{\int_{-\infty}^{\infty} |v(t)|^2 \, \mathrm{d}t} \sqrt{\int_{-\infty}^{\infty} |u(t)|^2 \, \mathrm{d}t}$$

Also, if we are given a set of linearly independent functions, we can apply the Gram-Schmidt process to obtain a set of orthonormal functions, which in turn form the basis of a function (inner product) space and can be used to represent any function contained in that space through an orthonormal expansion.

To conclude this discussion on orthonormal expansions, we tie up the slightly fuzzy idea that many functions are vectors [32]. For any function $v(t)$ in some inner product space $\mathcal{V}$, the Gram-Schmidt process ensures we can write it as an orthonormal expansion

$$v(t) = \sum_{i=1}^{n} v_i \phi_i(t)$$

for some orthonormal basis $\{\phi_1(t), \ldots, \phi_n(t)\}$. Note that this function is finite dimensional. If the upper limit on the summation were infinity, then the function would be infinite dimensional. Similarly, for another function $u(t) \in \mathcal{V}$, we can write the orthonormal expansion

$$u(t) = \sum_{i=1}^{n} u_i \phi_i(t)$$

[32] Not all functions are elements of vector spaces. Gallager (2008) treats this subject in detail from a measure theoretic perspective. The end result, however, is that we can almost always neglect the details in practice.

The same basis functions are used here; only the coefficients of the expansion are different. This is because both functions belong to the same space, which is spanned by $\{\phi_1(t), \ldots, \phi_n(t)\}$.

The inner product of these two functions can be written as

$$\begin{aligned}
\langle v, u \rangle &= \int_{-\infty}^{\infty} v(t)u^*(t)\mathrm{d}t \\
&= \int_{-\infty}^{\infty} \left(\sum_{i=1}^{n} v_i \phi_i(t) \right) \left(\sum_{j=1}^{n} u_j \phi_j(t) \right)^* \mathrm{d}t \\
&= \int_{-\infty}^{\infty} \sum_{i=1}^{n} \sum_{j=1}^{n} v_i u_j^* \phi_i(t) \phi_j^*(t) \mathrm{d}t \\
&= \sum_{i=1}^{n} \sum_{j=1}^{n} v_i u_j^* \int_{-\infty}^{\infty} \phi_i(t) \phi_j^*(t) \mathrm{d}t \\
&= \sum_{i=1}^{n} \sum_{j=1}^{n} v_i u_j^* \langle \phi_i, \phi_j \rangle \\
&= \sum_{i=1}^{n} v_i u_i^* \\
&= \langle \boldsymbol{v}, \boldsymbol{u} \rangle
\end{aligned}$$

where we can write the vectors $\boldsymbol{v}$ and $\boldsymbol{u}$ in the traditional form:

$$\boldsymbol{v} = (v_1, \ldots, v_n) \quad \text{and} \quad \boldsymbol{u} = (u_1, \ldots, u_n)$$

So, we have shown that the summation and integral forms of the inner product, and thus all associated corollaries, are equivalent in the sense that an analogue
waveform can always be represented by the coefficients of its orthonormal expansion given an appropriate orthonormal basis.

We now have all of the tools we need to encode discrete symbols in analogue waveforms. Indeed, we will show in the next section that the process of modulation is, in effect, reciprocal to source encoding; we will begin with a collection of bit strings and use the theory of orthonormal expansions to create analogue waveforms, which can then be transmitted over a channel to an intended receiver.

5.2.2 Pulse Amplitude Modulation (PAM)

Suppose we have a string of bits that we wish to encode in an analogue waveform. A simple approach would be to invoke the sampling theorem directly, but in reverse. To this end, we could map the k th bit b_k contained in an arbitrary bit sequence to a signal amplitude x_k. For example, we might define the mapping

$$0 \mapsto -1 \quad \text{and} \quad 1 \mapsto +1$$

so that $x_k \in \{-1, +1\}$ for all k. Using the sampling theorem, we can then assign these signal values as weights in the orthogonal expansion

$$x(t) = \sum_{k=-\infty}^{\infty} x_k \operatorname{sinc}\left(\frac{t}{T} - k \right)$$

In this way, the set $\{x_k\}$ represent the values of the band-limited function $x(t)$ sampled at intervals of T seconds. Hence, we could recover the sequence $\{x_k\}$, and thus $\{b_k\}$, directly from the samples $\{x(kT)\}$.

Most digital baseband modulation schemes are based on a similar principle as described above. The simplest of these schemes is known as pulse amplitude modulation (PAM). To describe the

basic idea of this method, suppose we have a sequence of bits generated by a source encoder. In a pulse amplitude modulator, each group of m bits is mapped to a signal [33] in an M-ary

set $\mathcal{A} = \{a_1, \ldots, a_M\}$, where $M = 2^m$. These signals are used as weights in an expansion of the form

$$x(t) = \sum_{k=-\infty}^{\infty} x_k p(t - kT)$$

where $x_k \in \mathcal{A}$ for all k and $p(t)$ is the basic pulse waveform. This basic waveform can also be regarded as the impulse response of a pulse shaping filter by defining the impulse train

$$u(t) = \sum_{k=-\infty}^{\infty} x_k \delta(t - kT)$$

and writing $x(t)$ as

$$x(t) = (u * p)(t)$$

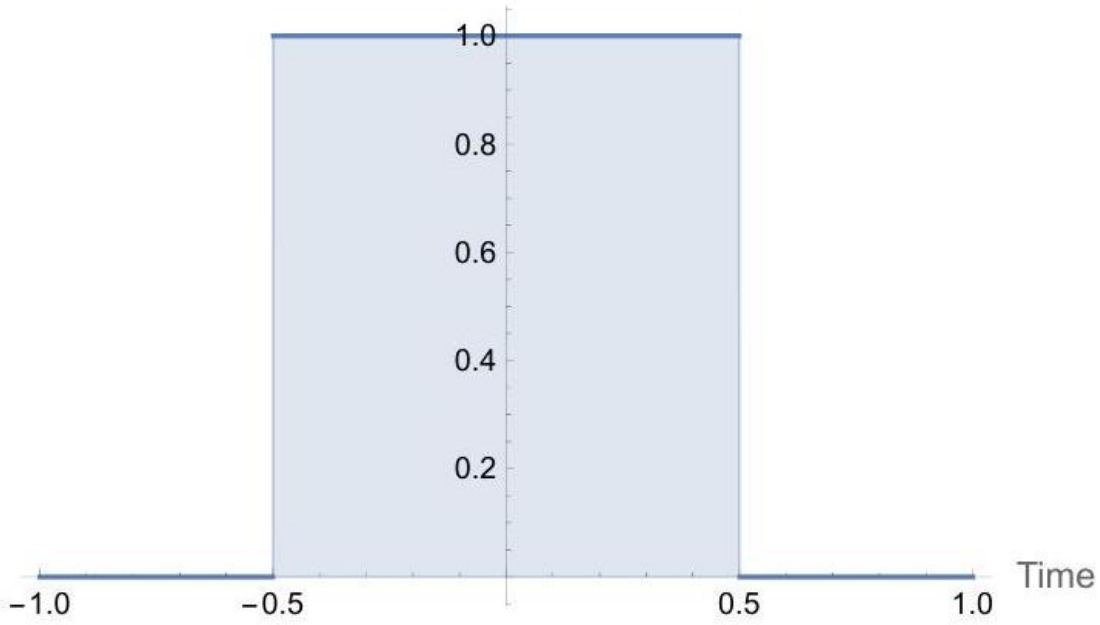
The signal rate of this scheme is $R_\mathrm{s} = 1/T$. Since $m = \log M$ bits are encoded every T seconds, the bit rate is $R_\mathrm{b} = mR_\mathrm{s} = m/T$ bits per second.

The pulse function $p(t)$ can be chosen such that $\{p(t - kT)\}$ is an orthonormal basis for some function space, in which case eq. (9) is an orthonormal expansion. It is also possible to choose the pulse function such that the set $\{p(t - kT)\}$ is not orthogonal at all. In this course, we will primarily be concerned with the case of orthogonal pulse sets.

The most basic choice of pulse function is the rectangular pulse

$$p(t) = \square(t/T).$$

In this case, the functions in $\{p(t - kT)\}$ are clearly pairwise orthogonal, since they do not overlap in time. They are not orthonormal, however, since $\|p(t)\| = \sqrt{T}$. Obviously, we could easily redefine the pulse such that we have process.



(a) Rectangular pulse.

(b) "Half-sine" pulse.

Figure 16: Examples of basic pulses used in PAM and related modulation schemes.

an orthonormal basis if we wish:

[33] We use the term "signal" to describe the values that bit strings are mapped to during the modulation

$$p(t) = \frac{1}{\sqrt{T}} \sqcap (t/T)$$

Many other choices exist for the pulse function. As an example of a pulse used in practice, devices built according to the IEEE 802.15.4a standard [34] utilise a "half-sine" pulse

$$p(t) = \cos(\pi t/T) \sqcap (t/T)$$

The rectangular and "half-sine" pulses are illustrated in Figure 16.
Notice that the theory of PAM modulated waveforms has been developed here under the assumption that the fundamental pulse is centred at $t = 0$. Hence, the half-sine pulse has been expressed as a cosine [35]. Furthermore, we have considered infinite-dimensional orthogonal expansions, i.e., we assumed that an infinite sequence of bits is mapped to an infinite sequence of signals. All of this implies that, technically, any linear system through which the modulated waveform passes would be non-causal. However, ignoring this issue for now makes the exposition clearer and the mathematics easier. It may help to think

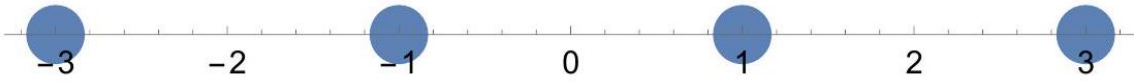


Figure 17: Constellation diagram for PAM with four signals and $d = 2$.
of $t = 0$ simply as a reference point in time, i.e., the time that an observer focuses on when viewing the waveform as a whole.

Signal Constellations

The standard PAM signal set is given by

$$\mathcal{A} = \left\{ -\frac{d}{2}(M-1), \ldots, -\frac{d}{2}, \frac{d}{2}, \ldots, \frac{d}{2}(M-1) \right\}$$

where d can be seen to be the distance between neighbouring signals. The concept of distance will be important when we explore the error probability of modulation schemes later. Signal sets such as the one given above are called constellations. The reason for using this term is that signal sets are often plotted on an Argand diagram [36] in order to easily illustrate their properties. An example for PAM with $d = 2$ and $M = 4$ is illustrated in Figure 17.

Energy per Signal

In a typical communication system, it is important to be able to quantify the average energy per signal at the modulator. This will later help us to analyse the signal-to-noise ratio (SNR) at the receiver.

The bit strings output from any properly designed source encoder will appear to be random if observed over a sufficiently long period of time. Hence, the signals encoded in the modulator will be random as well. Thus, we can
write the modulated PAM signal as

$$X(t) = \sum_{k=-\infty}^{\infty} X_k p(t - kT)$$

[34] This standard is the basis of Zigbee, WirelessHART, and 6LoWPAN specifications, which are used to achieve wireless communication in Internet of Things applications, such as industrial sensor networks.
[35] In the IEEE 802.15.4a standard, the fundamental pulse is actually a sine waveform that occupies the interval $[0, T)$. Specifically, it can be written as $p(t) = \sin(\pi t/T) \sqcap (t/T - 1/2)$, which is a little clunky owing to the shift in the rectangular pulse.
[36] Many passband modulation schemes use complex-valued signals. We will cover this topic later in the course.

where each coefficient in the set $\{X_k\}$ is an rv drawn from the constellation $\mathcal{A}$ with some probability. The waveform $X(t)$ is a random process. We will review some properties of random processes shortly, but for now, our goal is to compute the average energy per signal from this process.

Recall that the energy of a waveform is given by the norm of that waveform squared. Let us define the $2n$-dimensional partial waveform in relation to $X(t)$ to be

$$X^{2n}(t) = \sum_{k=-n+1}^{n} X_k p(t - kT)$$

It follows that we can define the average energy per signal to be

$$E_{\mathrm{s}} = \lim_{n\to\infty} \frac{1}{2n} \left\| X^{2n}(t) \right\|^2$$

Assuming that the pulse function is chosen such that the expression for $X(t)$ given above is an orthonormal expansion, we can simplify this equation for E_{s} as follows:

$$\begin{aligned}
E_{\mathbf{s}} &= \lim_{n\to\infty} \frac{1}{2n} \int_{-\infty}^{\infty} \left(X^{2n}(t)\right)^2 \, \mathrm{d}t \\
&= \lim_{n\to\infty} \frac{1}{2n} \int_{-\infty}^{\infty} \sum_{k=-n+1}^{n} \sum_{l=-n+1}^{n} X_k X_l p(t - kT) p(t - lT) \mathrm{d}t \\
&= \lim_{n\to\infty} \frac{1}{2n} \sum_{k=-n+1}^{n} \sum_{l=-n+1}^{n} X_k X_l \int_{-\infty}^{\infty} p(t - kT) p(t - lT) \mathrm{d}t \\
&= \lim_{n\to\infty} \frac{1}{2n} \sum_{k=-n+1}^{n} X_k^2.
\end{aligned}$$

Now, if we assume the distribution of the random variable X_k is the same for all k, and that all variables are independent, then the law of large numbers
ensures that the sample average of the random variable tends to the expected value as the number of samples grows large. In other words, we have

$$E_{\mathrm{s}} = \lim_{n\to\infty} \frac{1}{2n} \sum_{k=-n+1}^{n} X_k^2 = \mathrm{E}\left[X_k^2\right]$$

which, it should be noted, is independent of the index k since the rvs $\{X_k\}$ are identically distributed.

The result we have derived here agrees nicely with our goal of computing the average energy per signal. We have viewed the signals $\{X_k\}$ as rvs, and we have arrived at an expression that describes the average energy of one of these rvs. Although we treated an infinite-dimensional waveform here, we could have considered a finite-dimensional waveform and used the expected value in the definition of the average energy per signal rather than relying on its appearance through the law of large numbers. In both cases, the result is the same.

Similarly, we could have considered an expansion in which the time-shifted pulses are not orthogonal. In this case, it still makes sense to define the average energy per symbol as the expected value of X_k^2 to be consistent with the other scenarios. In all cases, it is prudent to normalise the energy of the pulse function that multiplies X_k so that it has unit energy [37].

For PAM, this energy is given by

$$E_{\mathrm{s}} = \sum_{i=1}^{M} p_i a_i^2$$

where p_i is the probability of signal a_i occurring and $a_i \in \mathcal{A}$ for all i. If all signals in $\{X_k\}$ are drawn uniformly, this simplifies to

$$E_{\mathbf{s}} = \frac{1}{M} \sum_{i=1}^{M} a_i^2$$

It is possible to prove by induction that

$$E_{\mathrm{s}} = \frac{d^2\left(M^2-1\right)}{12} = \frac{d^2\left(2^{2m}-1\right)}{12}$$

for the standard PAM constellation given above.
Notice that for fixed distance d, the energy per signal increases exponentially in the number of bits represented by the signal, which is worth bearing in mind when thinking of practical implementations. If we turn the equation around and solve for the number of bits per signal as a function of the energy-to-squared-distance ratio, we have

$$m = \frac{1}{2}\log\left(1+\frac{12E_{\mathrm{s}}}{d^2}\right) \qquad \text{bits per signal}$$

which looks suspiciously like Shannon's capacity formula for an AGN channel.
In such a channel, we would want to separate signals by a distance that is proportional to the standard deviation σ of the noise in order to maintain the possibility of distinguishing between signals at the receiver. Letting $d = \alpha\sigma$ for some proportionality constant α, we can rewrite the equation for m as

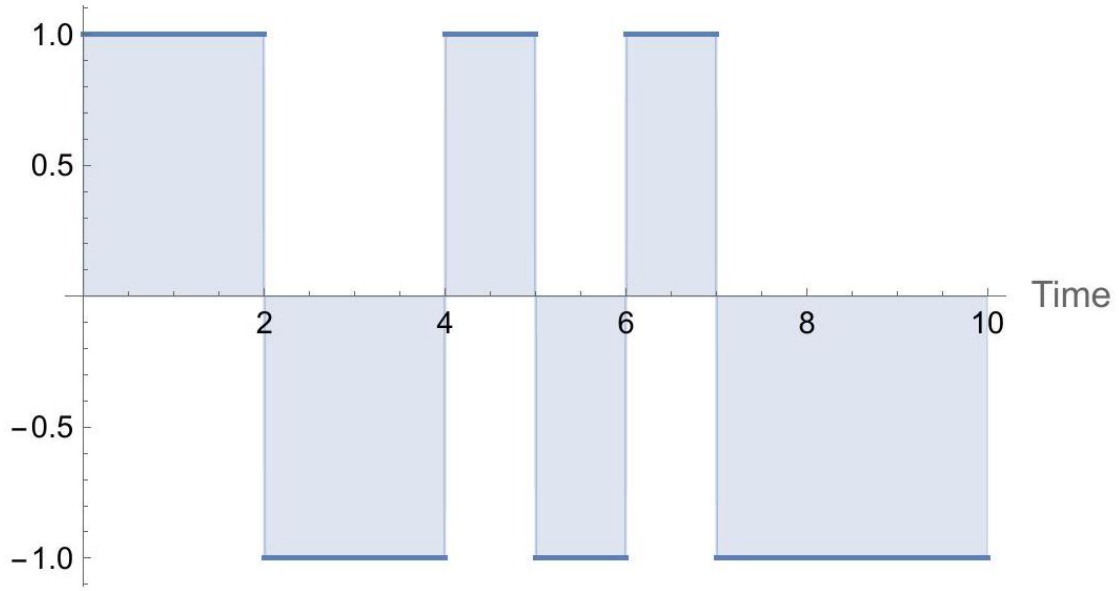
$$m = \frac{1}{2}\log\left(1+\frac{12}{\alpha^2}\frac{E_{\mathrm{s}}}{\sigma^2}\right) \qquad \text{bits per signal}$$

which, indeed, does resemble the celebrated $\frac{1}{2}\log(1+\text{ SNR }$) formula, where the SNR term is effectively E_{s}/σ^2.

We will return to noise and AGN channels later in the course. For now, this brief excursion served only to link Shannon's rather abstract theoretical result to a practical modulation scheme.

Examples of PAM Waveforms

The simplest of PAM waveforms was illustrated in Figure 5 and is reproduced here for convenience. This is a binary antipodal PAM waveform with a rectangular pulse. This particular modulation technique is sometimes referred



Encoded string: 121
Figure 18: Example of a binary antipodal PAM waveform with a rectangular pulse.

[37] Sometimes, it may be more useful to normalise the pulse so that it has unit power. It does not really matter what one does as long as all energy is properly accounted for in an analysis of the signals and waveform.

to as polar non-return-to-zero (NRZ) coding.
In general, the use of a pair of pulses to represent binary data is known as line coding. Grant and Glover (2009) [38] and Sklar and Harris (2021) [39] give numerous examples of line codes. Upon consulting these texts, you will notice that there is not really an excepted naming convention for line codes. Although, in general, return-to-zero (RZ) codes are those for which the basic pulse shape departs from zero at the beginning of a bit interval and returns to zero during the bit interval, whereas NRZ pulses depart from zero at the beginning of the interval and do not return to zero until the end. The purpose of choosing a pulse that returns to zero during a bit interval is that timing recovery at the receiver (i.e., synchronisation) is made much easier than for NRZ coding.

Most line coding schemes can be modelled as we have outlined above. A basic pulse shape is chosen, and a binary signal set (usually proportional to $\{0, 1\}$ or $\{-1, +1\}$) is used to map source encoded bits to PAM signals.

To encode more information in a PAM waveform, and thus to increase the rate of communication, one must employ larger signal sets. As alluded to earlier, the number of elements M in these signal sets is typically a power of

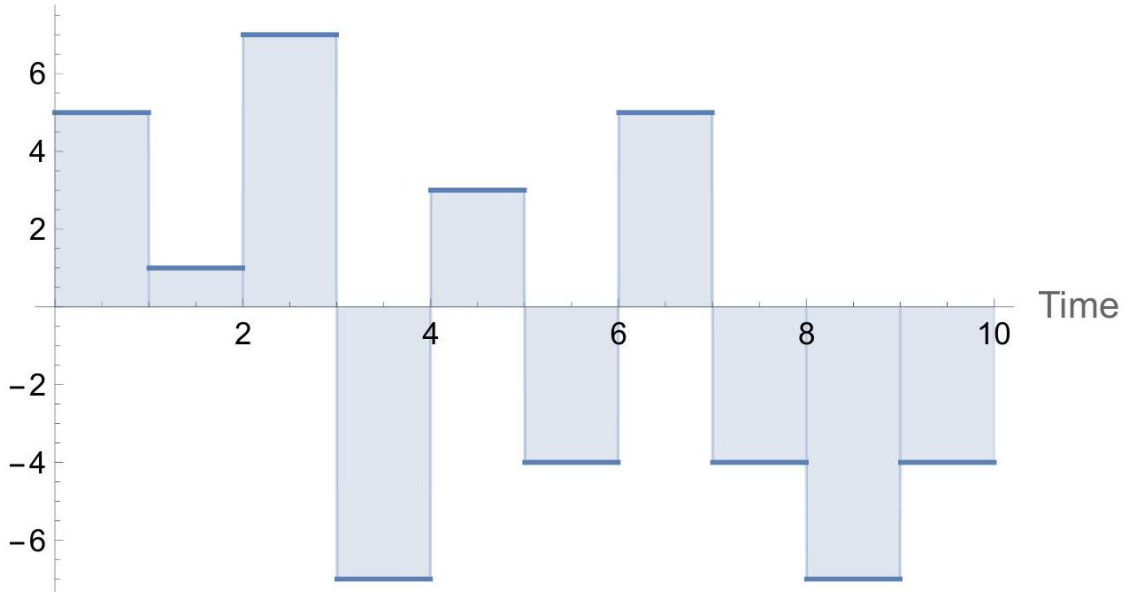


Figure 19: Example of an M-ary PAM waveform with $M = 8$. Three bits are encoded in each pulse. The constellation is $\mathcal{A} = \{-7, -5, -3, -1, 1, 3, 5, 7\}$.
two, and hence the number of bits that can be encoded in a single pulse is $m = \log M$. Such schemes are referred to as M-ary PAM modulation. An example is illustrated in Figure 19.

PAM Spectrum

Deterministic Waveforms An important part of modulation theory and communication system design is the characterisation of signal spectrum. For a given PAM waveform, this is easily accomplished by taking the Fourier transform of eq. (9), which yields

$$\tilde{x}(f) = \sum_{k=-\infty}^{\infty} x_k e^{-\mathrm{i}2\pi f k T} \tilde{p}(f)$$

where we have invoked the time shift property of Fourier transforms. Suppose only a finite number of the signals in $\{x_k\}$ are nonzero. Then $x(t)$ is a finiteenergy waveform, and we can thus characterise the energy spectral density of $x(t)$ as the squared-magnitude of its Fourier transform:

[38] See Section 6.4.
[39] See Section 2.8.2.

$$|\tilde{x}(f)|^2 = \left| \sum_{k=-\infty}^{\infty} x_k e^{-\mathrm{i}2\pi f k T} \right|^2 |\tilde{p}(f)|^2$$

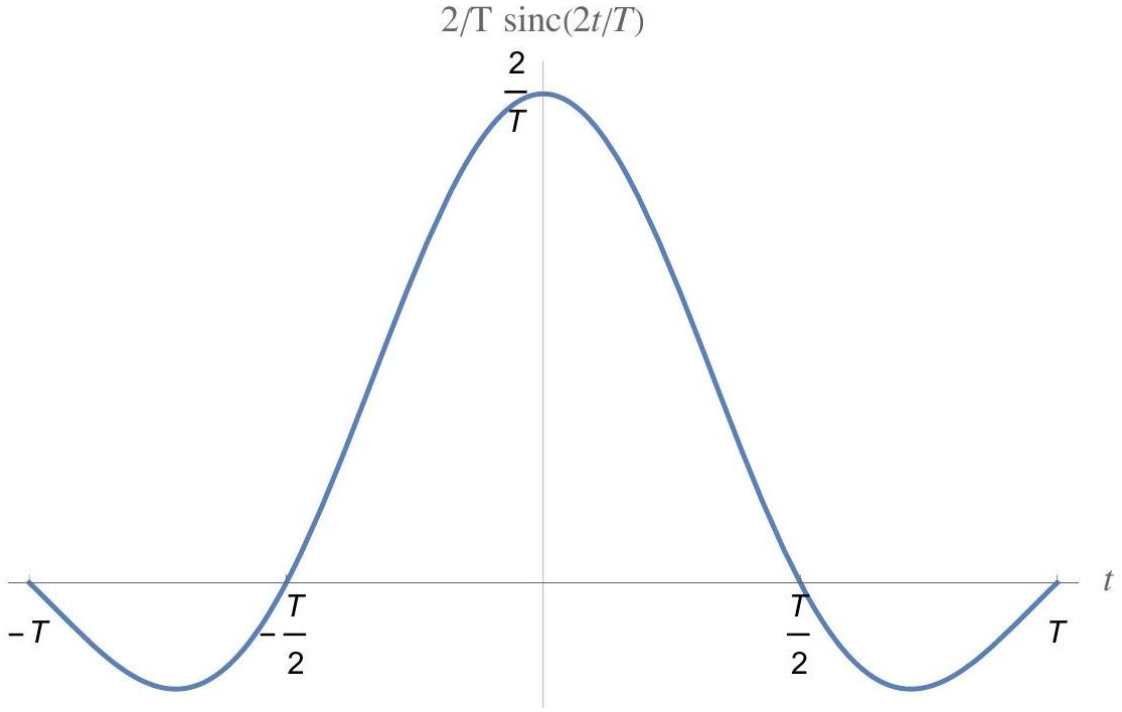


Figure 20: Illustration of $\frac{2}{T}\,\mathrm{sinc}(2t/T)$.

Notice that the shape of the spectrum is determined by the pulse $p(t)$. If the pulse is band-limited, then the PAM waveform will be band-limited. For example, let $T = 1/W$ be the time between the two zero crossings of the pulse

$$p(t) = \frac{2}{T}\,\mathrm{sinc}(2t/T)$$

that are nearest to the origin (Figure 20) [40]. Then

$$\tilde{p}(f) = \Pi\left(\frac{f}{2W}\right)$$

by the Fourier duality property, and we see that $x(t)$ is band-limited to $-W \leq f \leq W$.

The occupation of the frequencies from zero to W makes sense, but the idea that a signal can occupy negative frequencies may seem odd. This is an artefact of Fourier analysis. We will return to this topic a little later.

Random Waveforms In our discussion of the average energy per signal, we viewed the PAM signals as rvs. In general, digital communication system

modelling takes this probabilistic approach, because engineers are concerned with performance measures like the probability of detection error.

When we treat the PAM signals as rvs, the PAM waveform can be written as

$$X(t) = \sum_{k=-\infty}^{\infty} X_k p(t - kT)$$

In this case, we cannot study the signal spectrum by taking a Fourier transform, since there is not a well-defined notion of the transform of a random process. Despite this difficulty, it is possible

[40] Observe that $\{p(t - kT)\}$ is an orthogonal set by Parseval's theorem, as discussed in the context of the sampling theorem.

to analyse the spectrum of a random pulse sequence, i.e., a random PAM waveform. Detailed calculations are given below. These are rather involved, so we skip to the end here and state the key result:

The shape of the spectrum of a random PAM waveform, which is characterised by its power spectral density [41], is proportional to the energy spectral density $|\tilde{p}(f)|^2$ corresponding to the pulse function $p(t)$.

Observe that the pulse function $p(t)$ plays the same role in shaping the spectrum of both deterministic and random PAM waveforms. For this reason, engineers devote careful consideration to the design of $p(t)$ when specifying a new system. We have already seen the rectangular pulse and the half-sine pulse, which are both fundamental pulse shapes in practice. We will come across another important pulse shape a little later.

5.3 † Derivation of Power Spectral Density for Random Waveforms

We begin with a brief review of some properties of random processes. Recall that the autocorrelation function of a random process $X(t)$ is defined as

$$\mathrm{R}_X(t, t+\tau) = \mathrm{E}[X(t)X(t+\tau)]$$

Here, the independent variable t represents the time at which the function is evaluated. The variable τ signifies the time difference, or lag, for which the correlation is evaluated relative to time t.

A random process is said to be stationary if its distribution is not a function of time, i.e., it remains constant as time increases. Most physical process are nonstationary. For example, the distribution of daytime temperatures in Oxford changes as the seasons evolve. But, when viewed over an appropriate time scale, most random processes modelled on physical events resemble stationary processes, which are much easier to work with mathematically.

If the autocorrelation function is stationary, then it is not a function of the time variable t, i.e.,

$$\mathrm{R}_X(t, t+\tau) = \mathrm{R}_X(0, \tau)$$

For the rest of this discussion, we will simply write $\mathrm{R}_X(\tau)$ when stationary is satisfied. In this case, the power spectral density of $X(t)$ is defined as the Fourier transform

$$\mathrm{S}_X(f) = \int_{-\infty}^{\infty} \mathrm{R}_X(\tau) e^{-\mathrm{i}2\pi f\tau} \mathrm{d}\tau$$

Let us reflect on terminology for a moment. The function $\mathrm{R}_X(\tau)$ is called the "autocorrelation function", because it evaluates the correlation between the random process $X(t)$ and itself for a given time lag τ. The term "power spectral density", however, may seem a little more abstract, given that it is simply defined as the Fourier transform of the autocorrelation function. To understand the etymology, first recall that an ergodic random process is defined as one for which the time average of the process is equal to the ensemble average (i.e., the expected value). If we were able to integrate a random process, we could define the time-averaged autocorrelation function as

$$\overline{\mathrm{R}}_X(\tau) = \lim_{T\to\infty} \frac{1}{T} \int_{-T/2}^{T/2} X(t)X(t+\tau)\mathrm{d}t$$

from whence we see that the time-average power is

$$\bar{P}_S = \overline{\mathrm{R}}_X(0) = \mathrm{R}_X(0)$$

where the second equality follows from ergodicity. But, the definition of the power spectral density tells us that

[41] Since the Fourier transform of a random process is not well defined, the energy spectral density of a random PAM waveform does not exist. Instead, we must consider the Fourier transform of the autocorrelation function of the process, which is defined as the power spectral density.

$$\mathrm{R}_X(0) = \int_{-\infty}^{\infty} \mathrm{S}_X(f)\mathrm{d}f$$

which is the inverse Fourier transform of the power spectral density of $X(t)$ evaluated at $\tau = 0$. This equation reveals that the function $\mathrm{S}_X(f)$ must quantify the power of $X(t)$ per unit of frequency; hence, the name "power spectral density" follows naturally.

To characterise the spectrum of a random PAM waveform $X(t)$, it seems that we should calculate the autocorrelation function of $X(t)$ and compute its Fourier transform to obtain the power spectral density. However, this is not possible, because, as we shall see, the autocorrelation function depends on t as well as τ. In other words, $X(t)$ is not stationary.

The autocorrelation function of $X(t)$ can be manipulated as follows:

$$\begin{aligned}
\mathrm{R}_X(t, t+\tau) &= \mathrm{E}[X(t)X(t+\tau)] \\
&= \mathrm{E}\left[\left(\sum_{k=-\infty}^{\infty} X_k p(t-kT)\right)\left(\sum_{j=-\infty}^{\infty} X_j p(t+\tau-jT)\right)\right] \\
&= \mathrm{E}\left[\sum_{k=-\infty}^{\infty}\sum_{j=-\infty}^{\infty} X_k X_j p(t-kT)p(t+\tau-jT)\right] \\
&= \sum_{k=-\infty}^{\infty}\sum_{j=-\infty}^{\infty} \mathrm{E}\left[X_k X_j\right] p(t-kT)p(t+\tau-jT) \\
&= \sum_{k=-\infty}^{\infty}\sum_{l=-\infty}^{\infty} \mathrm{E}\left[X_k X_{k+l}\right] p(t-kT)p(t+\tau-(k+l)T)
\end{aligned}$$

In these calculations, we have used the fact that the expectation operator is linear, and hence the expected value of a sum is the sum of expected values. Furthermore, in the last line, we have made the substitution $j = k + l$ for the inner summation index.

Let us focus on the function $\mathrm{E}\left[X_k X_{k+l}\right]$ for the moment. This is the autocorrelation function of the discrete sequence $\{X_k\}$, and we use the notation $\mathrm{R}_{\{X_k\}}(l)$ to represent it (assuming the discrete signal process is stationary). To illustrate how to calculate this function, suppose the elements in $\{X_k\}$ are independent and identically distributed (i.i.d.) binary rvs that take on values in the set $\{0, 1\}$ with equal probability. Then for $l = 0$, we have

$$\mathrm{E}\left[X_k^2\right] = \frac{1}{2} \times 1 + \frac{1}{2} \times 0 = \frac{1}{2}$$

For any other $l \neq 0$, we can write

$$\mathrm{E}\left[X_k X_{k+l}\right] = \mathrm{E}\left[X_k\right]\mathrm{E}\left[X_{k+l}\right] = \left(\frac{1}{2} \times 1 + \frac{1}{2} \times 0\right)^2 = \frac{1}{4}$$

where the two expectations split in the first step because the rvs X_k and X_{k+l} are independent when $l \neq 0$. Hence, for this example, we have that

$$\mathrm{R}_{\{X_k\}}(l) = \begin{cases} 1/2, & l = 0 \\ 1/4, & l \neq 0 \end{cases}$$

Returning to the question of stationarity, we observe that eq. (10) depends on both t and τ. But, we also note that eq. (10) is periodic in t. To see this, simply substitute $t + T$ for t and note that, due to the infinite summation limits, we recover the original function. Random processes that have periodic autocorrelation functions are known as cyclostationary processes. Although we cannot simply take the Fourier transform of the autocorrelation function of a cyclostationary process to obtain the power spectral density, we can use Fourier analysis to characterise the spectrum of such processes quite easily.

First, we write $\mathrm{R}_X(t, t+\tau)$ as a Fourier series

$$\mathrm{R}_X(t, t+\tau) = \sum_{k=-\infty}^{\infty} \mathrm{R}_{X,k}(\tau) e^{\mathrm{i}2\pi kt/T}$$

where

$$\mathrm{R}_{X,l}(\tau) = \frac{1}{T} \int_{-T/2}^{T/2} \mathrm{R}_X(t, t+\tau) e^{-\mathrm{i}2\pi lt/T} \, \mathrm{d}t$$

The Fourier series coefficients capture the power spectral density over a period of the autocorrelation function. The l th coefficient is a function of τ only and is known as the lth cyclic autocorrelation function of $X(t)$. The power spectral density of the cyclostationary process is defined as the Fourier transform of the 0th cyclic autocorrelation function.

Observe that the 0th cyclic autocorrelation function, which is given by

$$\mathrm{R}_{X,0}(\tau) = \frac{1}{T} \int_{-T/2}^{T/2} \mathrm{R}_X(t, t+\tau) \mathrm{d}t$$

is nothing more than the time average of the periodic autocorrelation function over a single period. Other higher-order cyclic autocorrelation functions are similarly defined, but where a phase rotation is applied during the averaging process.

Let us average the central period of the autocorrelation of $X(t)$, which we gave in eq. (10):

$$\begin{aligned}
\mathrm{R}_{X,0}(\tau) &= \frac{1}{T} \int_{-T/2}^{T/2} \sum_{k=-\infty}^{\infty} \sum_{l=-\infty}^{\infty} \mathrm{R}_{\{X_k\}}(l) p(t-kT) p(t+\tau-(k+l)T) \mathrm{d}t \\
&= \frac{1}{T} \sum_{l=-\infty}^{\infty} \mathrm{R}_{\{X_k\}}(l) \sum_{k=-\infty}^{\infty} \int_{-T/2}^{T/2} p(t-kT) p(t+\tau-(k+l)T) \mathrm{d}t \\
&= \frac{1}{T} \sum_{l=-\infty}^{\infty} \mathrm{R}_{\{X_k\}}(l) \sum_{k=-\infty}^{\infty} \int_{-T/2-kT}^{T/2-kT} p(t) p(t+\tau-lT) \mathrm{d}t
\end{aligned}$$

In the last line, we have made use of the substitution $t \leftarrow t - kT$. Observe that the integrand in this line of the equation does not depend on k, and the integration region is the interval $\mathcal{I}_k = [-T/2 - kT, T/2 - kT)$. Collectively, the intervals $\{\mathcal{I}_k\}$ partition the real line, i.e., they form a disjoint set that covers the reals. Thus, for each k, the integral captures the area under the curve $p(t)p(t+\tau-lT)$ in the region $\mathcal{I}_k$, and the sum of these integrals captures
the area under the curve over the entire real line. Hence, we can write

$$\mathrm{R}_{X,0}(\tau) = \frac{1}{T} \sum_{l=-\infty}^{\infty} \mathrm{R}_{\{X_k\}}(l) \int_{-\infty}^{\infty} p(t) p(t+\tau-lT) \mathrm{d}t$$

Let us define the correlation function of the pulse to be

$$\mathrm{C}_p(\tau) = \int_{-\infty}^{\infty} p(t) p(t+\tau) \mathrm{d}t$$

Notice that this function differs from the time-average autocorrelation in that we do not multiply by $1/T$ before taking the limit of the integration region to the whole real line. Thus, we implicitly assume the correlation of $p(t)$ with itself has finite energy. For rectangular pulses and other well-behaved pulses used in practice, this is always the case.

Using the definition of the correlation function, we can rewrite the expression for $\mathrm{R}_{X,0}(\tau)$ as

$$\mathrm{R}_{X,0}(\tau) = \frac{1}{T} \sum_{l=-\infty}^{\infty} \mathrm{R}_{\{X_k\}}(l) \mathrm{C}_p(\tau - lT)$$

Observe that this equation can be written in terms of a convolution integral:

$$\mathrm{R}_{X,0}(\tau) = \frac{1}{T}\int_{-\infty}^{\infty}\left(\sum_{l}\mathrm{R}_{\{X_k\}}(l)\delta(t-lT)\right)\mathrm{C}_p(\tau-t)\mathrm{d}t$$

We are now in a position to compute the Fourier transform of the Oth cyclic autocorrelation function to obtain the power spectral density of the random PAM waveform. Taking the transform of the convolution above gives

$$\begin{aligned}\mathrm{S}_X(f) &= \frac{1}{T}\sum_{l=-\infty}^{\infty}\mathrm{R}_{\{X_k\}}(l)e^{-\mathrm{i}2\pi flT}|\tilde{p}(f)|^2 \\ &= \frac{1}{T}\,\mathrm{S}_{\{X_k\}}(f)|\tilde{p}(f)|^2\end{aligned}$$

where, in the first line, we have made use of the convolution, correlation, and conjugation properties of Fourier transforms, and in the second line, we have simply captured the fact that the sum in the first line is the power spectral density of the discrete random PAM signal process, i.e.,

$$\mathrm{S}_{\{X_k\}}(f) = \sum_{l=-\infty}^{\infty}\mathrm{R}_{\{X_k\}}(l)e^{-\mathrm{i}2\pi flT}$$

PAM Demodulation

Upon transmission of a (deterministic) PAM waveform

$$x(t) = \sum_{k=-\infty}^{\infty} x_k p(t-kT)$$

in a practical system, it passes through a noisy channel that may or may not also cause temporal dispersion in the original signal waveform. At the receiver, the signal is passed through a demodulator, the aim of which is to recover accurate estimates of the signals $\{x_k\}$ that constitute the weights in the PAM waveform. The output of the first stage of the demodulator should be a sequence of analogue values. The second stage, known as detection, then maps these values to the most likely transmitted signals. Finally, these signals can be mapped to the bit strings they represent. Figure 21 provides an illustration of these processes.

Non-Dispersive Channels We will first consider the simple case where the channel does not cause temporal dispersion. Furthermore, we will assume the channel does not attenuate the signal. If it did, we could easily correct for the attenuation at the receiver through a gain control stage, and the following discussion of demodulation would still be valid. Finally, to aid presentation, we will neglect noise; we will incorporate noise into the discussion in Section 5.

In our simple model, the received waveform at the input to the demodulator is just the transmitted PAM waveform $x(t)$. This waveform is passed through

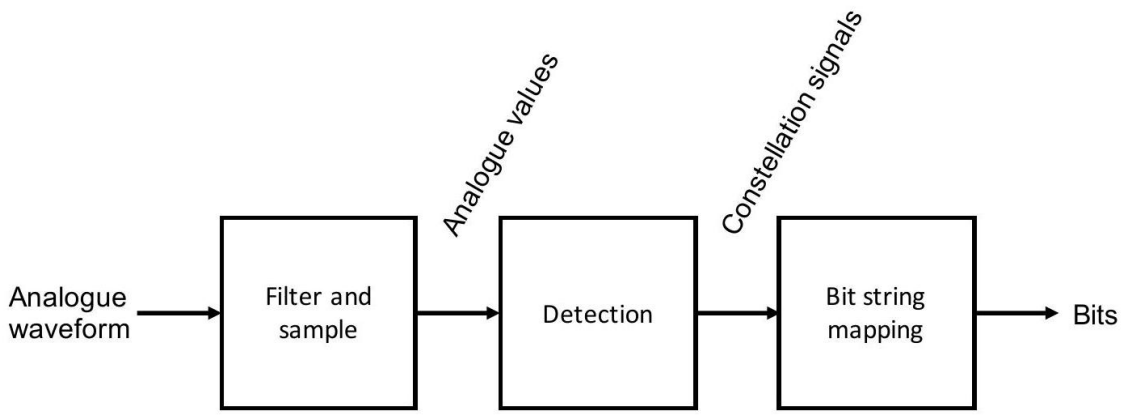


Figure 21: Block diagram of demodulator processes.
a linear filter with impulse response $q(t)$ to obtain

$$\begin{aligned} y(t) &= \int_{-\infty}^{\infty} x(\tau)q(t-\tau)\mathrm{d}\tau \\ &= \sum_{k=-\infty}^{\infty} x_k \int_{-\infty}^{\infty} p(\tau - kT)q(t-\tau)\mathrm{d}\tau \end{aligned}$$

Note that the first line of this equation is just the convolution of the PAM waveform with the filter impulse response, and the second line is a weighted sum of convolutions of the shifted pulses and the filter response.

The output of the filter is sampled at time $t = lT$ for each $l \in \mathbb{Z}$. We can write the l th sample as

$$y_l = y(lT) = \sum_{k=-\infty}^{\infty} x_k \int_{-\infty}^{\infty} p(\tau - kT)q(lT-\tau)\mathrm{d}\tau$$

Ideally, y_l would be a perfect replica of x_l. In particular, it would not contain interference from the other signals ..., $x_{l-1}, x_{l+1}, \ldots$, which is known as intersymbol interference (ISI). If $\{p(t-kT)\}$ forms an orthonormal basis for the PAM waveform, then we can simply choose the filter response to match the pulse function, i.e., $q(t) = p(-t)$. This would yield the estimate

$$y_l = x_l$$

as required. When the demodulator filter response is matched to the pulse in this way, the demodulation process is known as matched-filter demodulation.

The question remains as to whether the matched-filter demodulator yields a good estimate of the transmitted signal sequence $\{x_k\}$ in a noisy channel. We will return to this question in Section 5 through an analysis of the signal-tonoise ratio (SNR) of the noisy analogue signal output from the demodulator.

The Nyquist Criterion We observe from eq. (11) that the output of the demodulator filter can be written as [42]

$$y(t) = \sum_{k=-\infty}^{\infty} x_k g(t-kT)$$

where

$$g(t) = \int_{-\infty}^{\infty} p(t-\tau)q(\tau)\mathrm{d}\tau.$$

This should make sense. Recall that we were able to express the PAM waveform as the convolution of an impulse train and the pulse function $p(t)$, and the demodulator filtering operation invokes another convolution. And, of course, convolutions are associative [43].

We stated earlier that ISI can be avoided if we choose the pulse $p(t)$ and the filter response $q(t)$ such that $g(0) = 1$ and $g(lT) = 0$ for all integers $l \neq 0$. Demodulated waveforms that satisfy this property are called ideal Nyquist waveforms. We showed that when a matched filter is used in the demodulator, and the pulse sequence is comprised of a set of orthonormal pulse functions, then the ideal Nyquist condition is met.

Two immediate examples of pulse sequences that yield ideal Nyquist waveforms are the set of time-shifted rectangular pulses and the set of time-shifted sinc functions. In the first case, the pulses are time-limited, but they are not band-limited, since the Fourier transform of a rectangular pulse is a sinc func-

tion, which decays slowly (like $1/f$). In the second case, the pulses are not time-limited, although they are band-limited.

[42] This follows from the change of variables $\tau \leftarrow t - \tau$.

[43] Let $x(t), y(t)$, and $z(t)$ all be functions of a single variable. The associativity property of convolution ensures that $((x * y) * z)(t) = (x * (y * z))(t)$

It is important in practical systems to design ideal Nyquist waveforms that are nearly band-limited and nearly time-limited. Or, at the very least, the waveforms should decay much faster than $1/t$ in time and $1/f$ in frequency. The Nyquist criterion gives us a rule that we can use to find such pulses.

Theorem 3.2 (Nyquist criterion). Let $\tilde{g}(f) = \tilde{p}(f)\tilde{q}(f)$ denote the Fourier transform of $g(t) = (p * q)(t)$, where $p(t)$ is the fundamental pulse of a PAM waveform with signal rate $1/T$ and $q(t)$ is the impulse response of the demodulator filter. Then the demodulated PAM waveform is ideal Nyquist if and only if $\tilde{g}(f)$ satisfies the "Nyquist criterion", which is defined as

$$\sum_{k=-\infty}^{\infty} \tilde{g}\left(f + \frac{k}{T}\right) = T$$

A proof of this theorem is given at the end of this section. For now, let us explore its consequences. Firstly, although it seems a little abstract, one can imagine that many pulses and filter responses satisfy the Nyquist criterion and, thus, yield ideal Nyquist waveforms. Crucially, if we specify either the pulse or the filter response, but not both, then we can use the Nyquist criterion to determine the other. Furthermore, we can state the following corollary of the theorem, which we alluded to above but make precise here [44].

Corollary 3.3 (Orthonormal Shifts). Let $p(t)$ be defined such that $\tilde{g}(f) = |\tilde{p}(f)|^2$ satisfies the Nyquist criterion. Then $\{p(t-kT)\}$ is an orthonormal set. Conversely, suppose $p(t)$ is chosen such that $\{p(t-kT)\}$ is an orthonormal set. Then $|\tilde{p}(f)|^2$ satisfies the Nyquist criterion.

We can also use the theorem to explore the relationship between the signal rate $1/T$ and the bandwidth of the waveform. Suppose we wish to limit the baseband bandwidth of the transmitted PAM waveform to half of the signal

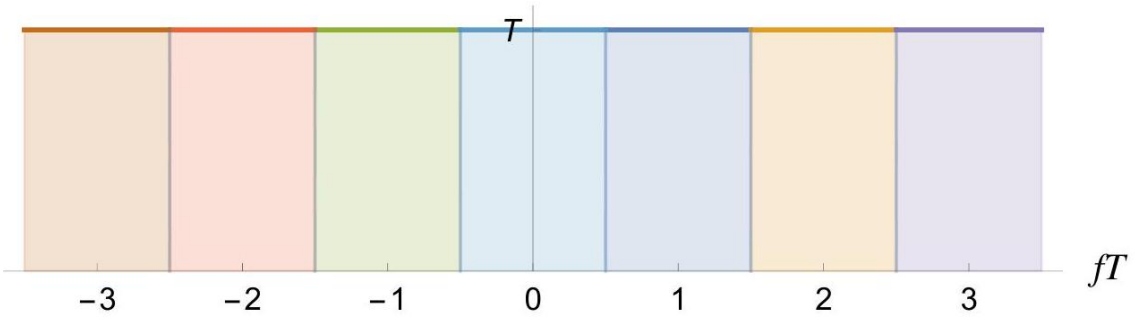


Figure 22: An example illustrating the case where the Nyquist criterion is met when signalling at a rate equal to twice the baseband bandwidth ($2W = 1/T$). Each rectangular function depicts $\tilde{g}(f - k/T) = T \sqcap (ft - k)$ for integer k.

rate, i.e., $W = 1/2T$. To achieve this goal, we could theoretically design the pulse such that $\tilde{p}(f) = 0$ outside of the interval $[-1/2T, 1/2T]$. For example, the sinc pulse

$$p(t) = \text{sinc}\left(\frac{t}{T}\right) \quad \leftrightarrow \quad \tilde{p}(f) = T \sqcap (fT)$$

achieves this condition. The function $\tilde{g}(f) = \tilde{p}(f)\tilde{q}(f)$ is, of course, also confined to the interval $[-1/2T, 1/2T]$, which has a width of $1/T = 2W$. Referring to the Nyquist criterion, the shifted functions $\{\tilde{g}(f - k/T)\}$ do not overlap in frequency in this scenario, but they meet exactly at their left/right boundaries. Consequently, each function must equal the constant value T over its frequency band. We can conclude that when the demodulated waveform is band-limited to half of the signal rate, there is only one possible design of the demodulated pulse frequency response that yields an ideal Nyquist (i.e., ISI free) waveform:

$$\tilde{g}(f) = T \sqcap (fT)$$

[44] The proof of the corollary is straightforward and is omitted for brevity. See Section 6.3.2 of Gallager (2008) for details.

See Figure 22 for an illustration. This result suggests that if we do, in fact, choose the pulse $p(t)$ to be a sinc function, then the demodulator filter would not filter at all (the frequency response would be flat). This is clearly impractical, since noise introduced at the receiver would pass through to the detector at all frequencies and cause problems.

Let us now consider the case where the baseband bandwidth is confined to be less than half of the signal rate, i.e., $W < 1/2T$. Referring to the Nyquist

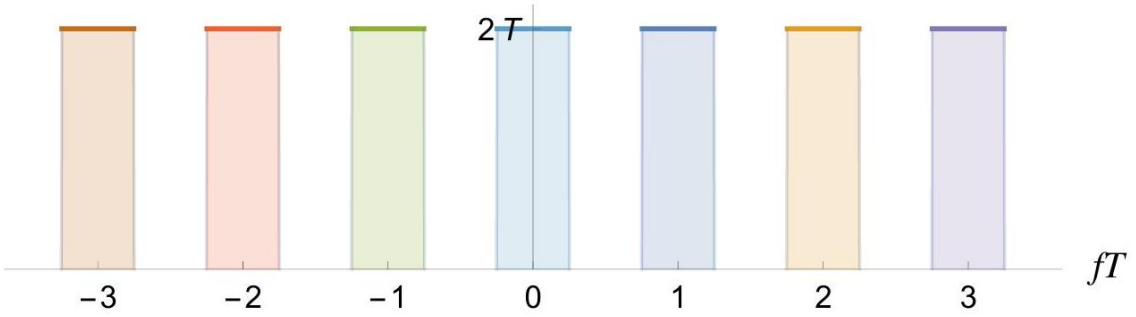


Figure 23: An example illustrating the case where the Nyquist criterion cannot be met when signalling at a rate equal to four times the baseband bandwidth ($W = 1/4T$). Each rectangular function depicts $\tilde{g}(f - k/T) = 2T \sqcap (2ft - 2k)$ for integer k.

criterion, we see that there is no combination of pulse and filter response that yields ideal Nyquist waveforms in this case, since the shifted functions $\{\tilde{g}(f - k/T)\}$ do not overlap in frequency. This situation is depicted in Figure 23.

Finally, we analyse the case where the baseband bandwidth is greater than half of the signal rate, i.e., $W > 1/2T$. Many pulse/filter designs exist that satisfy the Nyquist criterion in this scenario. A popular choice that is used in practice is the raised cosine transform, which is given by

$$\tilde{g}(f) = \begin{cases} T, & 0 \leq |f| \leq (1-\beta)/2T \\ T\cos^2\left[\frac{\pi T}{2\beta}\left(|f| - \frac{1-\beta}{2T}\right)\right], & (1-\beta)/2T \leq |f| \leq (1+\beta)/2T \\ 0, & |f| \geq (1+\beta)/2T \end{cases}$$

where $\beta \in [0, 1]$ is a design parameter. See Figure 24 for an illustration of how this transform satisfies the Nyquist criterion.

The parameter β is called the rolloff factor. It specifies how quickly the transform decays to zero around $f = 1/2T$. When $\beta = 0$, the transform is a rectangular pulse, and the baseband bandwidth is $W = 1/2T$. When $\beta = 1$, the bandwidth is extended to the signal rate $W = 1/T$, and the decay is much more gradual. In many practical applications, rolloff factors are chosen to be in the region of 0.1 to 0.2 .

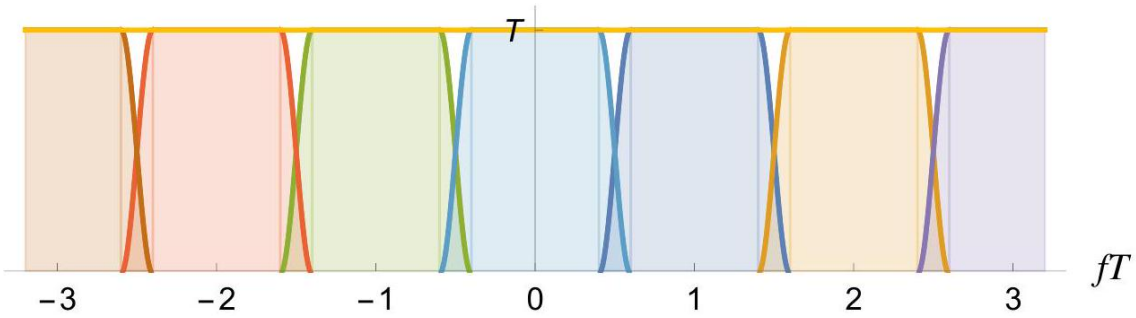


Figure 24: An example illustrating the case where the Nyquist criterion is met when signalling at a rate equal to approximately 1.66 times the baseband bandwidth ($W = 3/5T$). Each function is a raised cosine transform shifted by an integer multiple of $1/T$. The rolloff factor is $\beta = 0.2$,

which yields the bandwidth stated above. The solid horizontal yellow line is the sum of all raised cosine transforms shown, which illustrates that the Nyquist criterion is satisfied.

The inverse Fourier transform of the raised cosine transform is

$$g(t) = \text{sinc}\left(\frac{t}{T}\right)\frac{\cos(\pi\beta t/T)}{1-4\beta^2t^2/T^2}$$

Again, notice that when $\beta = 0$, we recover the sinc pulse. Also, we observe that the pulse is not strictly time-limited, but it decays much faster than the sinc pulse (like $1/t^3$) for $\beta > 0$. Hence, the raised cosine transform leads us to our goal of achieving ISI-free communication using band-limited and (approximately) time-limited waveforms.

As a final note, we must highlight the fact that the raised cosine result does not specify what the pulse $p(t)$ and the filter response $q(t)$ should be. Following the matched-filter principle, however, we can simply choose these functions such that their transforms are the square root of the raised cosine transform. More specifically, a popular design is the following:

$$|\tilde{p}(f)| = |\tilde{q}(f)| = \sqrt{\tilde{g}(f)}$$

where $\tilde{g}(f)$ is the raised cosine transform here. Note that the phases of the pulse and filter transforms are not defined here. However, the matched-filter principle suggests that we should select $\tilde{q}(f) = \tilde{p}^*(f)$ so that we achieve the
required time reversal in the the filter impulse response.

Dispersive Channels Let us turn our attention to dispersive channels. Recall that an linear time-invariant (LTI) system causes temporal dispersion, in general, and the output of the channel (the input to the demodulator in this case) is modelled as a convolution between the input waveform $x(t)$ and the channel impulse response (CIR) $h(t)$ (see Section 1.4). Now, passing the channel output through the demodulator filter yields the concatenated convolution

$$y(t) = (q * h * x)(t)$$

But, convolution operations are associative. Hence, in this case, we can treat

$$c(t) = (q * h)(t)$$

as an "effective" demodulation filter response. In doing so, we can apply the Nyquist criterion in the way that was described above.

The difference here is that we have

$$g(t) = (c * p)(t) = (q * h * p)(t)$$

Typically, the CIR $h(t)$ will need to be estimated or measured in a practical system so that we know how to reverse the dispersion that is caused, and thus recover the transmitted message. Hence, $h(t)$ is given; it cannot be altered easily. The filter design problem must, therefore, take $h(t)$ into account. This is, in theory, easily done. Observe that

$$\tilde{g}(f) = \tilde{q}(f)\tilde{h}(f)\tilde{p}(f)$$

by the repeated use of the convolution property of Fourier transforms. If, for example, we wish for $\tilde{g}(f)$ to be a raised cosine transform, which would satisfy the Nyquist criterion, then we simply need to design $\tilde{q}(f)$ and $\tilde{p}(f)$ to be the square root of the raised cosine transform divided by the square root of the
channel frequency response $\tilde{h}(f)$, being sure to satisfy the conjugate phase property that yields the required time reversal in the filter impulse response as well.

† Proof of Theorem 3.2 Here, we provide a proof of the Nyquist criterion for achieving zero ISI communication. Recall that a demodulated waveform is ideal Nyquist if $g(0) = 1$ and $g(lT) = 0$ for $l \neq 0$ where $g(t) = (p * q)(t)$. We can write $g(t)$ as the inverse transform

$$g(t) = \int_{-\infty}^{\infty} \tilde{g}(f)e^{\mathrm{i}2\pi ft}\,\mathrm{d}f$$

such that

$$\begin{aligned} g(lT) &= \int_{-\infty}^{\infty} \tilde{g}(f) e^{\mathrm{i}2\pi f l T} \, \mathrm{d}f \\ &= \sum_{k=-\infty}^{\infty} \int_{\frac{2k-1}{2T}}^{\frac{2k+1}{2T}} \tilde{g}(f) e^{\mathrm{i}2\pi f l T} \, \mathrm{d}f \\ &= \sum_{k=-\infty}^{\infty} \int_{-1/2T}^{1/2T} \tilde{g}\left(f + \frac{k}{T}\right) e^{\mathrm{i}2\pi f l T} \, \mathrm{d}f \\ &= \int_{-1/2T}^{1/2T} \sum_{k=-\infty}^{\infty} \tilde{g}\left(f + \frac{k}{T}\right) e^{\mathrm{i}2\pi f l T} \, \mathrm{d}f \end{aligned}$$

where, in the next-to-last line, we have made the substitution $f \leftarrow f - k/T$ and made use of the fact that $e^{\mathrm{i}2\pi k n} = 1$ for all integers k and l.

Let

$$\tilde{h}(f) = \sum_{k=-\infty}^{\infty} \tilde{g}\left(f + \frac{k}{T}\right)$$

Clearly, this is a periodic function with period $1/T$. Hence, we can write the Fourier series (with variable f)

$$\tilde{h}(f) = \sum_{l=-\infty}^{\infty} h_l e^{\mathrm{i}2\pi f l T}$$

where

$$h_l = T \int_{-1/2T}^{1/2T} \tilde{h}(f) e^{-\mathrm{i}2\pi f l T} \, \mathrm{d}f$$

Comparing with the expression we derived above for $g(lT)$, we see that the condition for achieving an ideal Nyquist waveform can be stated as

$$h_l = T g(-lT) = \begin{cases} T, & l = 0 \\ 0, & l \neq 0 \end{cases}$$

Substituting into the Fourier series for $\tilde{h}(f)$, we see that $\tilde{h}(f) = T$. Or, stated as in the theorem, we must have

$$\sum_{k=-\infty}^{\infty} \tilde{g}\left(f + \frac{k}{T}\right) = T$$

for the ideal Nyquist condition to be achieved.

Chapter 6

Passband Modulation and Demodulation

We covered the fundamental concepts and techniques related to modulation and demodulation in the previous section. Where that discussion was focused on baseband modulation, the discussion in this section will relate to passband modulation. The basic idea of passband modulation is that a baseband waveform is translated in frequency to a higher frequency band. This is a relatively simple process, but it will enable us to develop more advanced modulation schemes that are used in practical systems today.

6.0.1 Frequency Conversion

The translation of a waveform in frequency is known as frequency conversion. If the process relates to the shifting of a waveform to occupy bandwidth centred at a higher frequency - e.g., shifting from baseband to passband - it is referred to as up-conversion. Translation in the opposite direction is known as downconversion.

Consider a band-limited baseband waveform $x(t)$ with transform $\tilde{x}(f)$ that is zero for $|f| > W$. This waveform can be up-converted to frequency f_c simply by multiplying by $\cos(2\pi f_c t)$. The resulting passband waveform and its Fourier transform are given by (see Figure 25)

$$x_p(t) = x(t)\cos(2\pi f_c t) \quad \leftrightarrow \quad \tilde{x}_p(f) = \frac{1}{2}\tilde{x}(f - f_c) + \frac{1}{2}\tilde{x}(f + f_c).$$

The transform is most easily computed by expression the cosine as a sum of complex exponentials and using the frequency shift property of Fourier transforms. The frequency f_c is often referred to as the carrier frequency or the centre frequency of the transmission. Note that we can also represent the passband waveform as

$$x_p(t) = \Re\left\{x(t)e^{i2\pi f_c t}\right\}.$$

To recover the baseband waveform at the receiver, the passband wave-

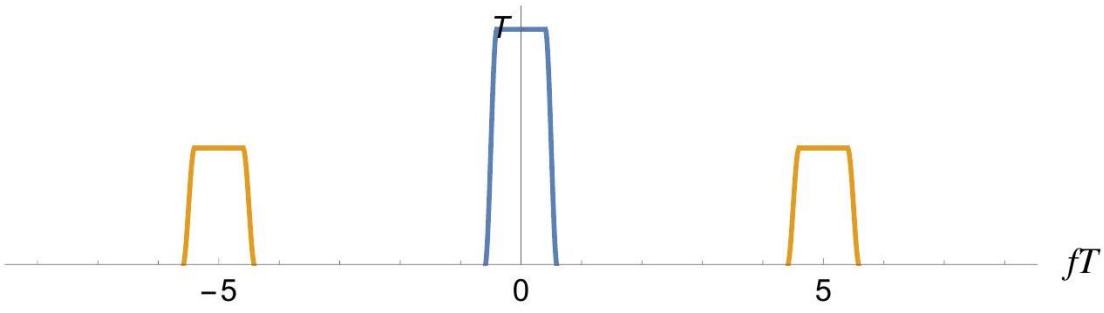


Figure 25: Illustration of a baseband raised cosine transform (blue) being upconverted to the centre frequency $f_c = 5/T$ (orange).

form is simply multiplied by the same cosine (scaled by a factor of two this time), and the resulting waveform is passed through a low-pass filter to remove high frequency components. Neglecting noise and other channel effects, the waveform prior to filtering is given by

$$\begin{aligned}\hat{x}_p(t) &= 2x_p(t)\cos(2\pi f_c t) \\ &= 2x(t)\cos^2(2\pi f_c t) \\ &= x(t)\left(1+\cos(4\pi f_c t)\right)\end{aligned}$$

After filtering, the component centred at frequency $2f_c$ is removed, and we recover $x(t)$.

In most of the literature on communication systems, frequency conversion is viewed as part of the modulation process. The filtering operation alluded to above is the same as the demodulator filtering operation discussed in the previous section when designed properly (as a low-pass filter). Frequency conversion is the last procedure to be carried out at the transmitter after the baseband waveform has been encoded, and it is the first procedure to be executed at the receiver. Again, this is a form of layering. If done correctly, we do not have to consider the effects of frequency conversion when designing a baseband modulation scheme (with one exception, which will be discussed shortly).

This approach to frequency conversion is known as double sideband modulation. For a standard PAM baseband waveform, this form of up-conversion is inefficient, since all information is contained in one sideband of the waveform.

This follows from the fact that a PAM signal (and thus a PAM waveform) is real valued. And, of course, the Fourier transforms of real-valued waveforms are conjugate symmetric about $f = 0$ (see the conjugation property of Fourier transforms on page 35). By relaxing this real-valued restriction to the signal constellation, we can design much more efficient modulation schemes, as we will discuss in Section 4.2.

6.0.2 Quadrature Amplitude Modulation (QAM)

Suppose we allow our signal constellation to contain complex elements:

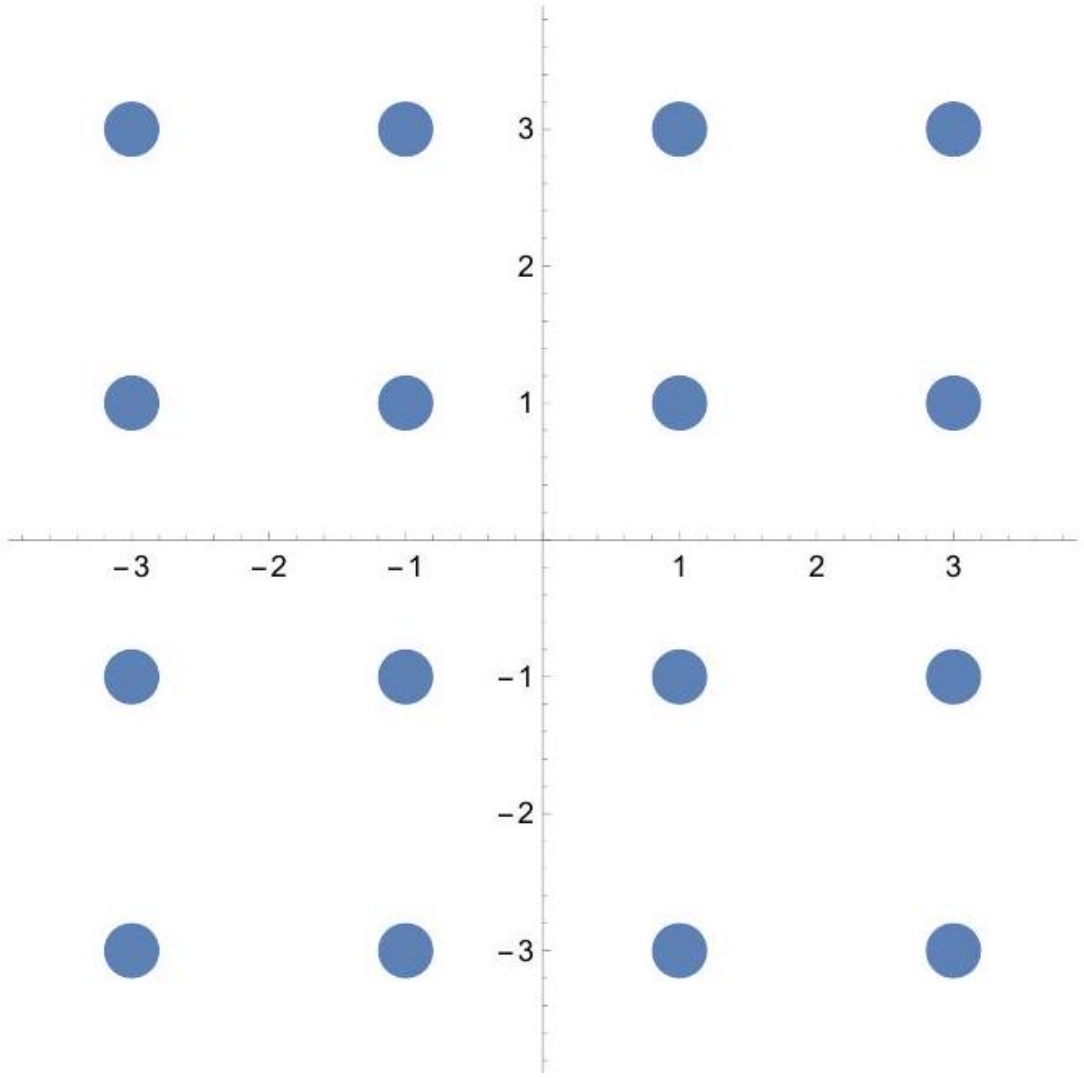
$$\mathcal{A} = \{a_1 + \mathrm{i}b_1, \ldots, a_M + \mathrm{i}b_M\}$$

where $a_i, b_i \in \mathbb{R}$ for all i. The baseband waveform

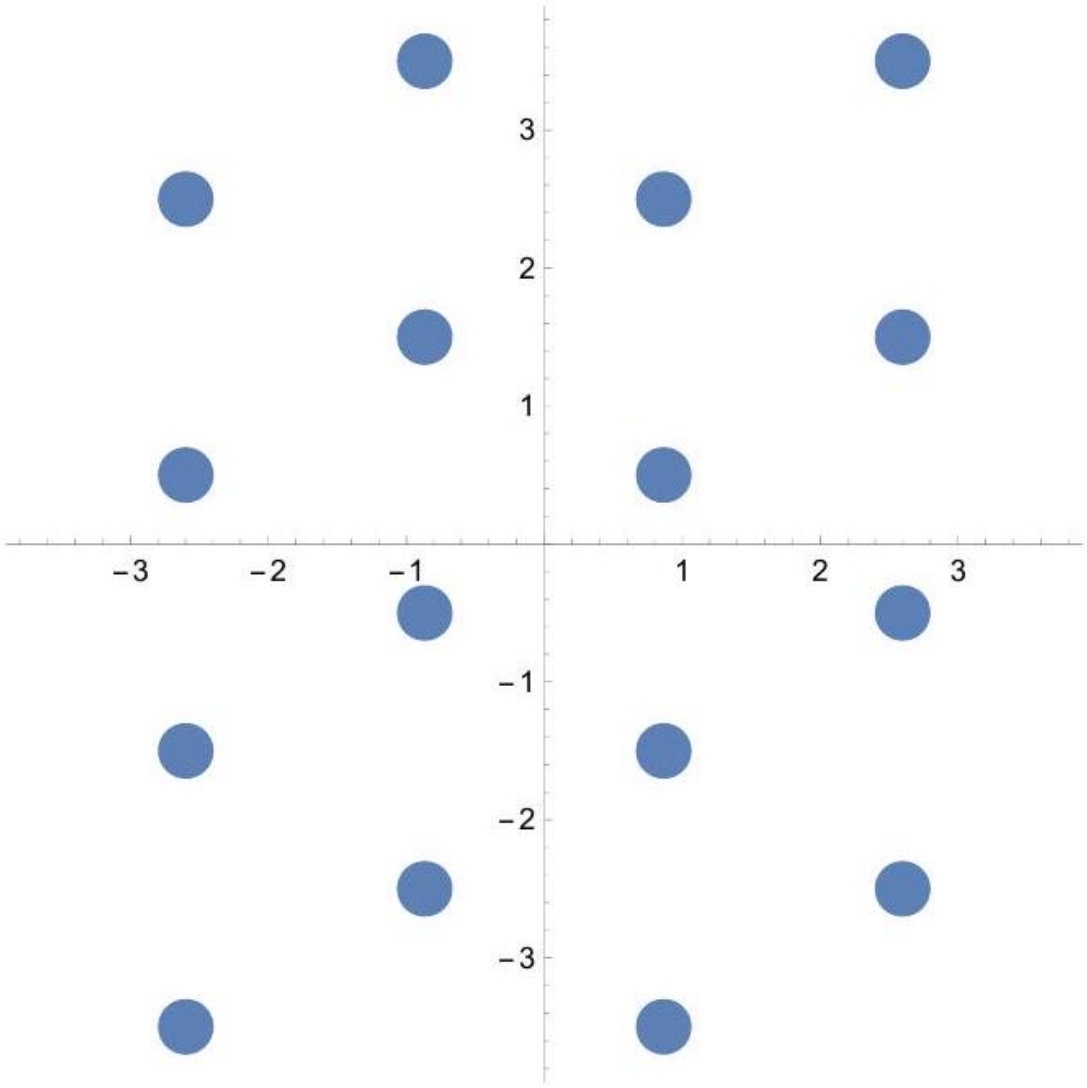
$$x(t) = \sum_{k=-\infty}^{\infty} x_k p(t - kT)$$

in this case would be complex as well. The process of encoding complex-valued signals in a waveform is, in general, referred to as quadrature amplitude modulation (QAM). The name is derived from the idea that, by using both real and imaginary components of signals, we are effectively overlaying two PAM signal sets. The real component of a QAM signal is known as the in-phase component. The imaginary component is called the quadrature component. An example of a standard QAM constellation that is constructed from two 4 -ary PAM constellations can be seen in Figure 26a. Four bits can be encoded in each of these signals.

Not all QAM constellations are constructed easily from two PAM constellations. An example of a constellation that does not readily adhere to such a construction is shown in Figure 26b. In fact, hexagonal QAM constellations such as this perform a little better than standard square lattice constellations when the number of signals M is large. Most practical applications encode less than ten bits per QAM signal, so $M \leq 512$, and the square lattice con-



(a) Standard 16-QAM constellation.



(b) Hexagonal 16-QAM constellation.

Figure 26: Examples of QAM constellations for $M = 16$.

stellations tend to work quite well.

Obviously, we cannot transmit a complex waveform. However, we can use the properties of cosine and sine functions to encode the real and imaginary parts of the signals $\{x_k\}$ in a passband waveform and recover these signals at the receiver.

As with passband PAM modulation, we can represent the passband QAM waveform as

$$x_{\mathfrak{p}}(t) = \Re\left\{x(t)e^{\mathrm{i}2\pi f_c t}\right\}$$

The difference here is that

$$x(t) = \Re\{x(t)\} + \mathrm{i}\Im\{x(t)\}$$

Thus, we have

$$x_{\mathfrak{p}}(t) = \Re\{x(t)\}\cos(2\pi f_c t) - \Im\{x(t)\}\sin(2\pi f_c t)$$

This expression gives us a clue as to how to encode QAM signals in a passband waveform. Specifically, we construct two separate baseband PAM waveforms: one for the real part of the QAM signal sequence and one for the imaginary part. We then up-convert the two waveforms separately by multiplying

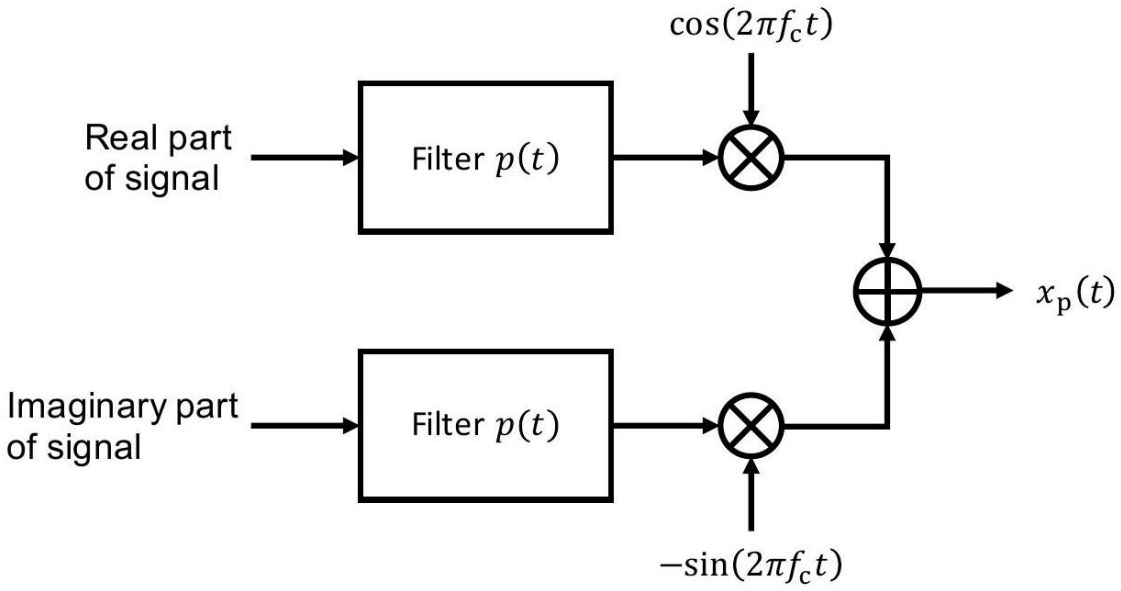


Figure 27: Block diagram of QAM modulator.

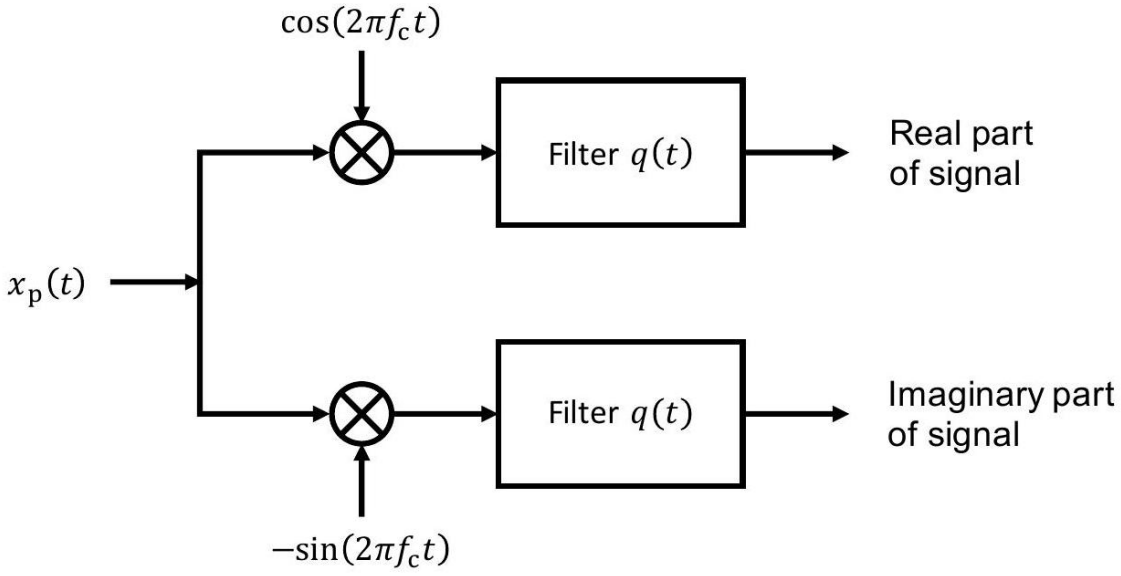


Figure 28: Block diagram of QAM demodulator.

by $\cos(2\pi f_c t)$ and $-\sin(2\pi f_c t)$, respectively, and add the resulting passband waveforms (Figure 27).

To recover the real and imaginary signals at the receiver, we perform downconversion and filtering. This is done over two branches in parallel - the in-phase (cosine) branch and the quadrature (sine) branch - as shown in Figure 28. To demodulate the real component, we multiply by $\cos(2\pi f_c t)$ and then filter the signal. To demodulate the imaginary component, we multiply by $-\sin(2\pi f_c t)$ and filter. Again, in a proper design, the filter would be matched to the pulse function $p(t)$.

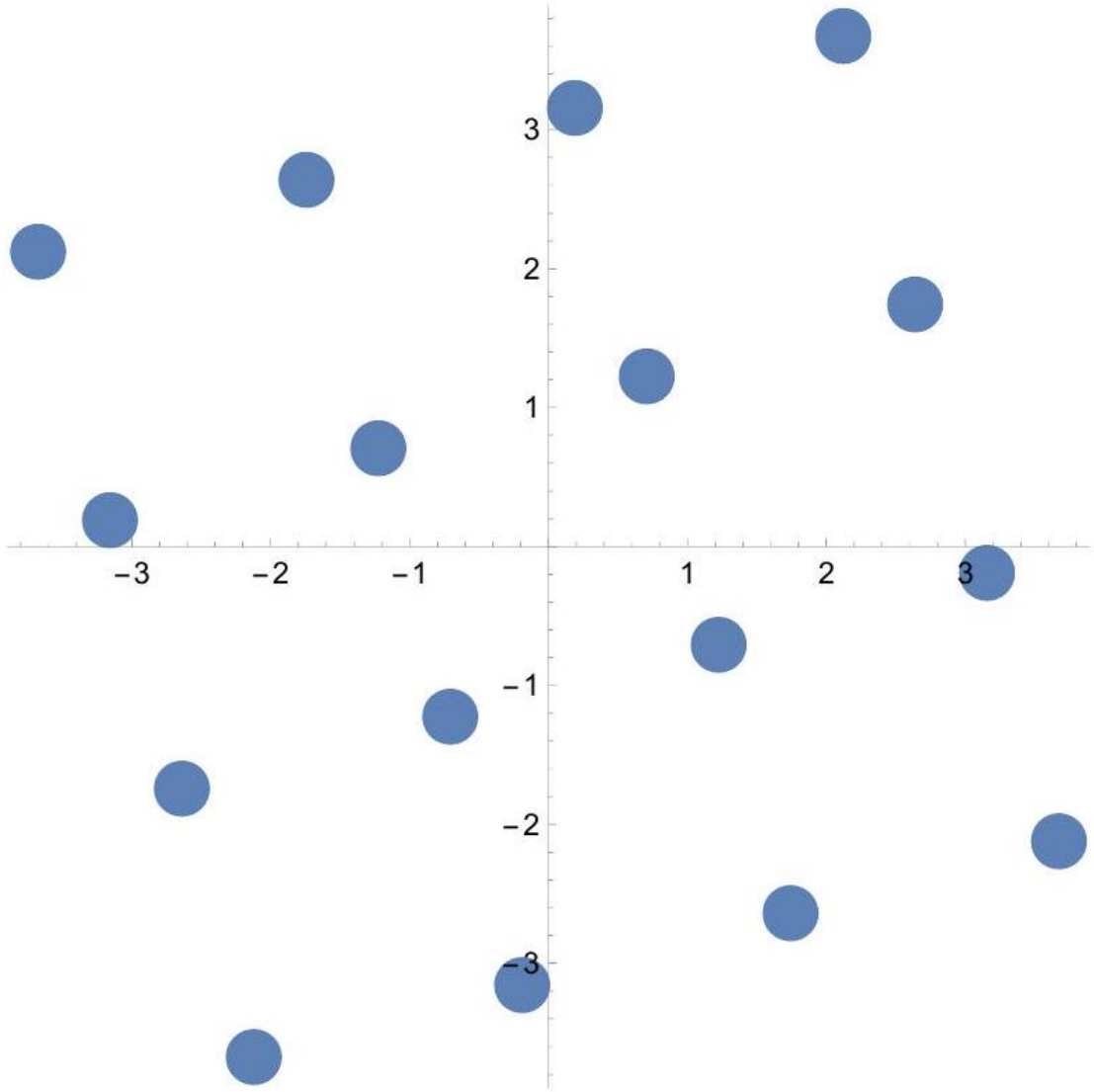


Figure 29: A rotated QAM constellation resulting from a carrier phase offset of $\pi/12$ radians between the transmitter and the receiver.

Phase Recovery It is possible, and often probable in practice, that the time delay induced by transmission over a physical channel will yield imperfect phase alignment at the receiver. This causes the local oscillators at the receiver to contain a phase offset. The process of down-conversion then leads to crosstalk between the real and imaginary signals. It is easiest to view this crosstalk through the lens of the constellation diagram. Effectively, if the phase offset changes slowly with time, then the constellation at the output of the demodulator (ignoring noise) is a rotated version of the constellation employed at the transmitter (Figure 29). In such cases, the phase offset can be easily corrected by first transmitting training signals that are known to both the transmitter and the receiver and solving for this offset.

Alternatively, if the phase offset changes more quickly with time, a phaselocked loop circuit can be implemented to estimate and correct the offset. This circuit resembles a simple feedback control loop with unit gain feedback and a phase-lag compensator followed by a voltage-controlled oscillator in the forward path.

6.0.3 Frequency Shift Keying (FSK)

Where PAM and QAM signals encode information in the amplitudes and phases of the transmitted pulse waveforms, it is also possible to encode information in the frequency band over which a signal

is transmitted. Suppose we have M such bands. Then by selecting one of them for transmitting a pulse during a given time period T, we can encode $m = \log M$ bits at a rate of $R_\mathrm{b} = m/T$ bits per second. This modulation technique is known as frequency-shift keying (FSK).

The baseband model for an FSK waveform is basically the same as for PAM and QAM, but where the standard signal set is defined as

$$\mathcal{A} = \left\{1, e^{\mathrm{i}2\pi f_\mathrm{d} t}, e^{\mathrm{i}2\pi(2f_\mathrm{d})t}, \ldots, e^{\mathrm{i}2\pi((M-1)f_\mathrm{d})t}\right\}$$

Notice that the signals in this case are functions of time. They are complex sinusoids with frequencies that are integer multiples of the fundamental frequency deviation f_d. Hence, we can write the baseband FSK waveform as

$$\begin{aligned} x(t) &= \sum_{k=-\infty}^{\infty} x_k(t) p(t - kT) \\ &= \sum_{k=-\infty}^{\infty} e^{\mathrm{i}2\pi s_k f_\mathrm{d} t} p(t - kT) \end{aligned}$$

where $s_k \in \{0, \ldots, M-1\}$ indicates the frequency shift for the k th signal.
To characterise the spectrum of $x(t)$, we can be easily compute its Fourier transform by using the frequency and time shift properties along with linearity:

$$\tilde{x}(f) = \sum_{k=-\infty}^{\infty} e^{-\mathrm{i}2\pi(f - s_k f_\mathrm{d})kT} \tilde{p}\,(f - s_k f_\mathrm{d})$$

For the k th pulse, observe that the spectrum is shaped by the shifted pulse transform $\tilde{p}\,(f - s_k f_\mathrm{d})$. The complex exponential factor only alters the phase of

6.1 FSK Transform

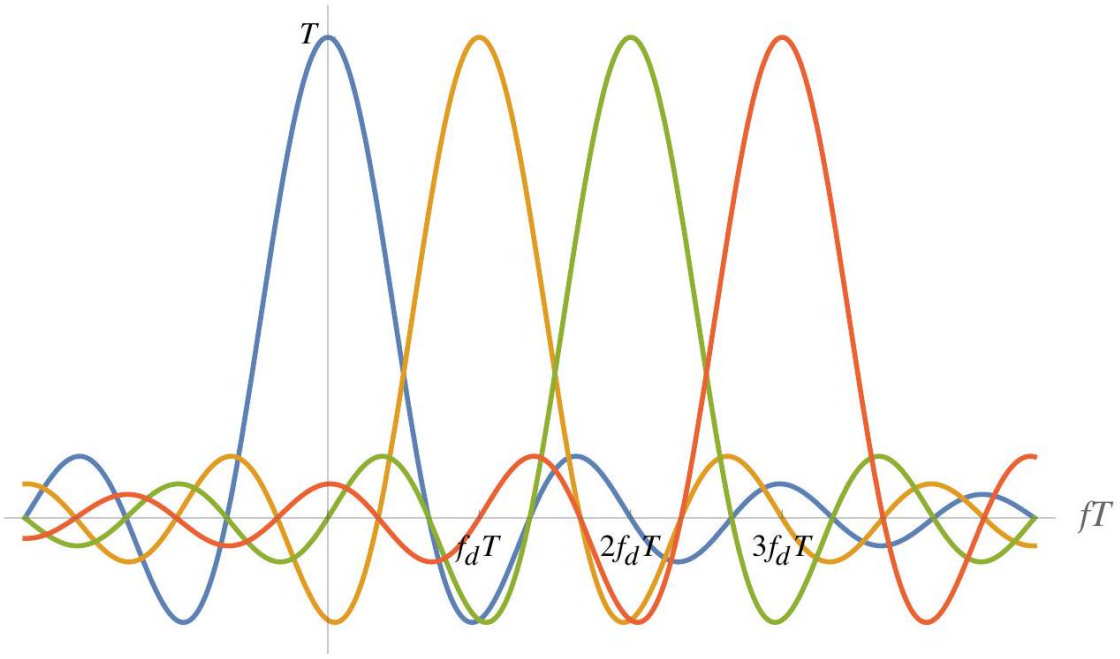


Figure 30: Illustration of the spectrum of an FSK waveform with $M = 4$ and $f_\mathrm{d} = 3/2T$ with rectangular pulse $p(t) = \sqcap(t/T)$. The baseband pulse has bandwidth $W = 2/T$ (measured to the first zero crossings of the sinc function), and the total bandwidth runs from $f = -1/T$ to $f = 11/2T$, which is the first zero crossing to the right of the line of symmetry for the pulse centred

at the highest frequency.
the transform. Hence, if the pulse is approximately band-limited to the interval $[-W, W]$, say, then the waveform is band-limited to the interval $[-W, W+ \ (M-1)f_{\mathrm{d}}]$, as illustrated in Figure 30. However, only a portion of this bandwidth is in use at any given time. Referring again to Figure 30, only one of the sinc functions shown would be active during a given signal interval T.

Although we have presented FSK as a baseband modulation scheme, FSK waveforms are typically only used in passband applications owing to the fact that individual pulses are, effectively, passband waveforms. Fortunately, frequency conversion of an FSK waveform simply translates the waveform in the usual way. Hence, we can apply the systematic framework developed above to characterise passband FSK waveforms.

As a final comment about FSK, one may wonder how demodulation can occur if the signals, themselves, are functions of time. Many different practical demodulator designs for FSK signals exist. However, we observe from eq. (12)
that if f_{d} is chosen such that the exponentially-weighted pulses are orthogonal for all constellation signals, then an M-branch demodulator filter can be employed, with each branch being matched to a given constellation signal. We will return to these ideas in the next section.

6.1.1 Orthogonal Frequency-Division Multiplexing (OFDM)

Most modern high-rate digital communication systems employ a modulation scheme known as orthogonal frequency-division multiplexing (OFDM). Having developed a theory of modulation based on orthonormal expansions, we will see that the step from PAM, QAM, and FSK to OFDM is almost obvious. Nevertheless, although OFDM was invented in the 1960s, it did not receive widespread use until the early 2000s.

The relatively recent uptake of OFDM is primarily due to the fact that it is difficult to implement in practice if modern digital signal processing (DSP) techniques are not utilised. In the latter half of the 20th century, these techniques were being developed. By the time Wi-Fi was standardised, they were ubiquitous, and their application in OFDM transceivers was well understood. Today, OFDM is used in fibre broadband networks, cable broadband networks, cellular networks, terrestrial digital video broadcast (DVB) systems, digital audio broadcast (DAB) systems, and Wi-Fi networks.

OFDM Waveform

As the name suggests, an OFDM waveform multiplexes data during a single transmission by using different frequency bands simultaneously. This is one aspect where OFDM differs from FSK, which utilises different frequency bands at different times. Indeed, FSK encodes information in the sequence of bands that are transmitted on, whereas OFDM typically encodes information on each available frequency band using a PAM or QAM waveform.

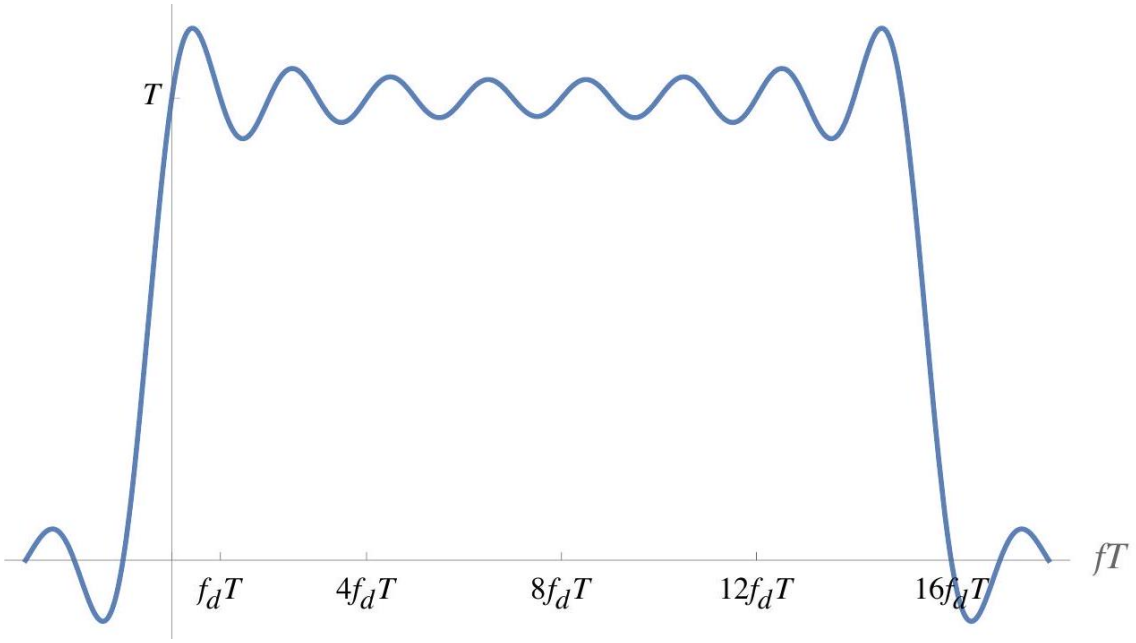


Figure 31: The Fourier transform of an example OFDM waveform with $n = 16$ subcarriers.

In its most basic form, the baseband OFDM waveform can be modelled as

$$x(t) = \sum_{k=-\infty}^{\infty} \sum_{l=0}^{n-1} x_{kl} e^{\mathrm{i}2\pi l l_{\mathrm{d}} t} p(t - kT)$$

Notice that the waveform during the k th symbol period is a superposition of signals that occupy different frequencies spaced apart by f_{d}Hz. The Fourier transform of this waveform looks very similar to the transform shown in Figure 30 when the pulse $p(t) = \Pi(t/T)$, but where the different sinc functions superimpose since all bands are utilised simultaneously. This is shown in Figure 31 .

The parameter n in the OFDM waveform model given above is the number of bands that are used. The centre frequency of each band is called a subcarrier. This terminology is adopted due to the fact that most applications in which OFDM is employed are passband applications, and thus the waveform described here would be up-converted to a carrier frequency f_{c} that is much higher than the baseband bandwidth. Frequency conversion of OFDM waveforms can be accomplished in the standard way, so we will not explicitly discuss
passband OFDM waveforms here.
The only other variables in eq. (13) that we have not covered are the signals $\{x_{kl}\}$. As noted above, these are typically drawn from a PAM or a QAM constellation. Different data is encoded on each subcarrier, i.e., for each index l. Consequently, for an M-ary constellation, the rate of this OFDM transmission is

$$R_{\mathrm{b}} = \frac{n \log M}{T} \quad \text{bits per second}$$

This factor of n increase in bit rate relative to comparable PAM, QAM, and FSK schemes simply comes from the multiplexing nature of the transmission. The benchmark schemes do not multiplex data. Of course, multiplexing comes with the drawback that the utilised bandwidth is n times larger than the benchmarks. The higher bandwidth used by OFDM justifies its label as a broadband modulation scheme in some texts.

Orthogonality In OFDM systems, the frequency deviation f_{d}, more commonly called the subcarrier spacing, and the pulse function $p(t)$ are chosen such that the exponentially weighted pulses $\left\{e^{\mathrm{i}2\pi l f_{\mathrm{d}} t} p(t - kT)\right\}$ are orthogonal. The pulse shape is ideally the rectangular pulse

$$p(t) = \sqcap(t/T)$$

When $p(t)$ is a rectangular pulse, the choice of f_d can be found easily by computing the inner product of two exponentially weighted pulses. Let us consider the fundamental pulses to make the notation simple. We define the k th exponentially weighted pulse as

$$w_k(t) = e^{\mathrm{i}2\pi k f_\mathrm{d} t} \sqcap (t/T)$$

The inner product of two pulses is given by

$$\langle w_k, w_l \rangle = \int_{-\infty}^{\infty} \left(e^{\mathrm{i}2\pi k f_\mathrm{d} t} p(t)\right) \left(e^{-\mathrm{i}2\pi l f_\mathrm{d} t} p(t)\right) \mathrm{d}t = \int_{-\infty}^{\infty} e^{-\mathrm{i}2\pi (l-k) f_\mathrm{d} t} \sqcap (t/T) \mathrm{d}t$$

This is just the Fourier transform of the rectangular pulse evaluated at frequency $f = (l-k)f_\mathrm{d}$. Hence, the inner product is

$$\langle w_k, w_l \rangle = T \,\mathrm{sinc}\left((l-k) f_\mathrm{d} T\right).$$

Clearly, when $k = l$, the inner product is T, regardless of the value of the subcarrier spacing f_d. The smallest value of f_d that yields orthogonality can be seen to be

$$f_\mathrm{d} = \frac{1}{T}$$

This is the minimum subcarrier spacing that yields orthogonal FDM, and it tends to be the choice that is most often used in practice. Of course, one could choose $f_\mathrm{d} = u/T$ or any positive integer u and still achieve orthogonality. But choices of $u > 1$ would expand the bandwidth that is used without providing a benefit in terms of bit rate. Nevertheless, such choices are made in some wireless networks to allow different transmissions to occur simultaneously without interfering with one another. For example, in a two-user network, a subcarrier spacing of $f_\mathrm{d} = 2/T$ may be used, where one user's transmission is offset from the other user's transmission by $1/T$ Hz. This technique is known as orthogonal frequency-division multiple access (OFDMA). A version of OFDMA is currently utilised in 4G and 5G cellular networks.

Demodulation in Non-Dispersive Channels As shown above, OFDM is an orthogonal modulation scheme. In fact, it is doubly orthogonal, since the two groups of superimposed signals transmitted in separate time intervals are orthogonal, and the signals in each group are, themselves, orthogonal. Consequently, assuming the channel is not dispersive, demodulation at the receiver can be accomplished with the aid of the matched filter principle. In this case, however, we should have n separate matched filters, each one matched to a complex exponential at a different frequency, as shown in Figure 32. The waveform at the output of each filter is then sampled at integer multiples of the pulse duration T, just as for PAM, QAM, and FSK.

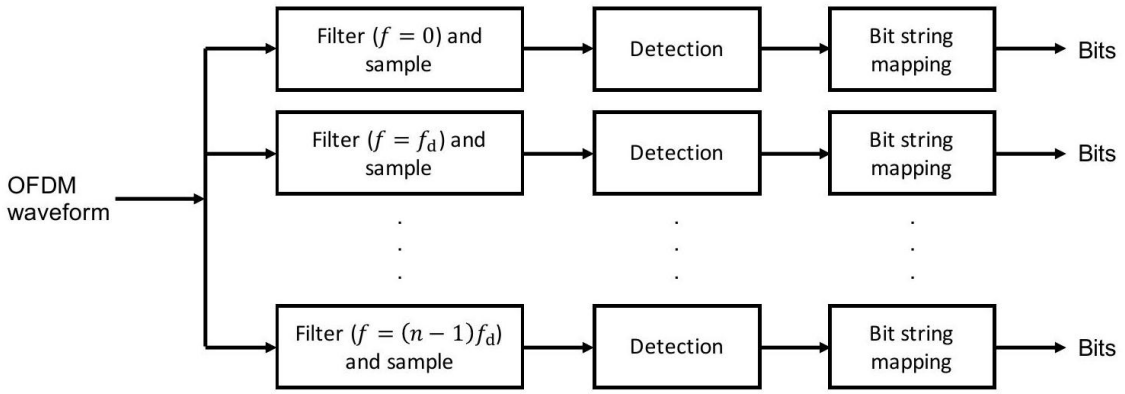


Figure 32: Block diagram depicting the basic structure of a conceptual OFDM demodulator. Actual demodulator implementations are rather more elegant and make use of modern DSP theory and algorithms.

Demodulation in Dispersive Channels When the channel causes temporal dispersion, the situation is a little different than for the other modulation schemes we discussed. OFDM is a block transmission scheme, since groups of signals are superimposed during a given pulse period T and

are transmitted simultaneously. In standard OFDM parlance, the ISI induced by a dispersive channel is referred to as interblock interference (IBI).

Modern OFDM systems are implemented with the goal of eliminating IBI through simple processing that is independent of the CIR. Several methods exist that achieve this aim. All of them rely on the basic idea of inserting a so-called guard interval between subsequent blocks, i.e., superimposed signal intervals, at the transmitter and processing these in some way at the receiver.

The simplest form of the guard interval is just a silent period that is long enough to allow for the dispersion (i.e., the tail of the waveform-CIR convolution) to die down. This method is known as zero padding. At the receiver, the dispersion contained in the zero padding interval is either discarded or used in a different way to facilitate the demodulation process.

The method that is most often used in practice is known as the cyclic prefix. In this technique, a portion of each superimposed signal, say αT with $0 < \alpha < 1$, is copied and appended to the beginning of that signal interval. In doing so, the total signal interval duration is increased to $(1 + \alpha)T$ seconds, and thus the bit rate is decreased (because no new information is added).

Again, at the receiver, the cyclic prefix is discarded before further processing takes place.

It is the insertion and removal of the guard interval that makes OFDM such a useful and simple mechanism for achieving high-rate communication. It can be shown that the cyclic prefix and zero padding methods actually facilitate a simple OFDM implementation based on the fast Fourier transform, which is a computationally inexpensive way to perform a discrete Fourier transform. Transmitter and receiver processing is made easy, and, crucially, ISI can be dealt with in the demodulator without needing to estimate the CIR and design a filter matched to that response. In mobile applications, the CIR may change regularly. Hence, the more traditional method of mitigating ISI by directly utilising the Nyquist criterion can be impractical in these situations.

Chapter 7

Noisy Channels

7.1 Introduction

It has been pointed out that all communication channels are noisy. It was this fundamental observation that Shannon was interested in when he developed his mathematical theory of communication. However, we ignored noise when we formulated a theory of modulation. That theory was based on the notion of orthogonal expansions, and we will soon see that the same framework can be applied to study noisy systems. Ultimately, we will need to understand how noisy demodulated signals can be mapped back to constellation signals in the optimal manner. Before we can embark upon that discussion, we must establish a firm model of noisy channels.

7.1.1 Additive Noise Channel

Recall that when a waveform $x(t)$ passes through an LTI channel with additive noise, we can model the input to the demodulator as

$$v(t) = (h * x)(t) + z(t) = \int_{-\infty}^{\infty} x(\tau)h(t-\tau)\mathrm{d}\tau + z(t)$$

Here, the function $z(t)$ is a sample noise waveform, and $h(t)$ is the channel impulse response (CIR). If the channel is not dispersive, then the received waveform at the input to the demodulator is just

$$v(t) = x(t) + z(t)$$

Notice that the models given above might refer to baseband or passband waveforms. To simplify matters, we will focus on the baseband case in what follows. Figure 33 provides a pictorial representation of the additive noise channel described by eq. (14).

For QAM, the modulated waveform can be written as the orthonormal

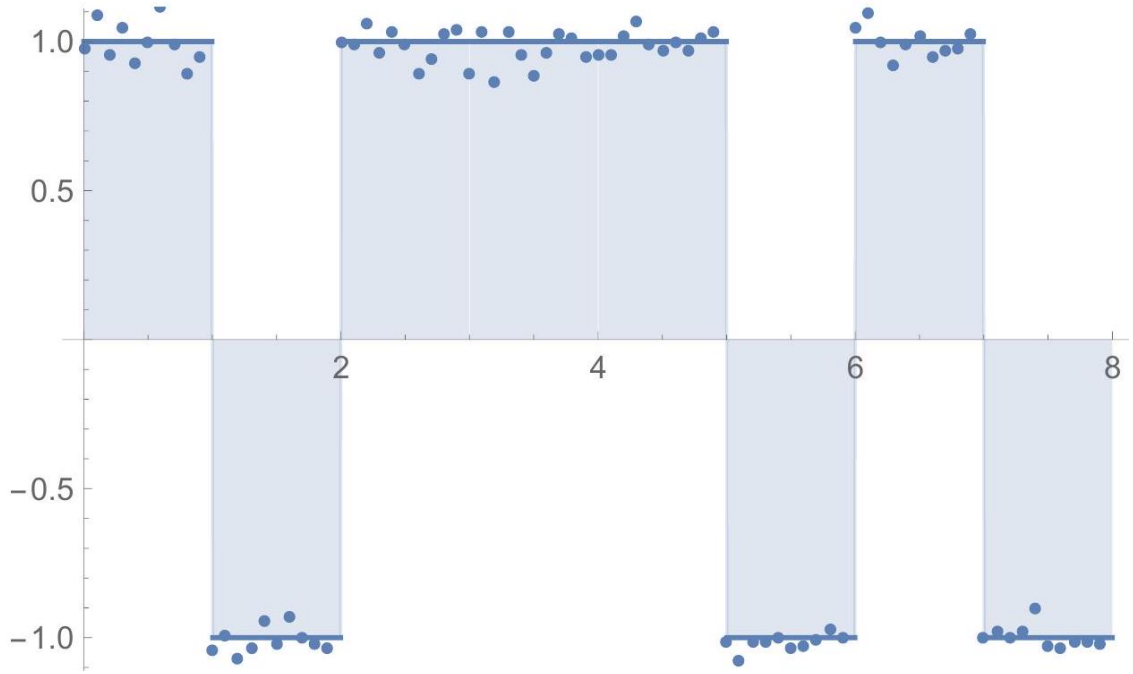


Figure 33: An illustration of a noisy signal. A PAM waveform is shown along with samples of the noisy channel output.
expansion [45]

$$x(t) = \sum_{k=-\infty}^{\infty} x_k \phi_k(t)$$

In our development of modulation theory, we adopted scaled, time-shifted Nyquist pulses for the orthonormal functions $\{\phi_k(t)\}$. For now, however, let us keep the discussion more general.

This expression for $x(t)$ is an infinite-dimensional expansion. In practice, of course, the signal will be finite dimensional. So, we will refine our definition of the QAM waveform to be

$$x(t) = \sum_{k=1}^{n} x_k \phi_k(t)$$

where n is the dimension of $x(t)$.
We can project the noise waveform $z(t)$ onto the same basis that spans the space that $x(t)$ belongs to. But, this will not necessarily capture all components of the waveform. In general, the noise function will belong to a higher

dimensional space than the QAM waveform. Hence, we can write the noise as

$$z(t) = \sum_{k=1}^{n} z_k \phi_k(t) + z_\perp(t)$$

where

$$z_k = \langle z, \phi_k \rangle = \int_{-\infty}^{\infty} z(t) \phi_k^*(t) \mathrm{d}t$$

and $z_\perp(t)$ is the component of $z(t)$ that is orthogonal to the space spanned by the basis functions $\{\phi_k(t)\}$.

It turns out, however, that we do not need to worry about $z_\perp(t)$ if we design the demodulator filter properly. Remember that this filter should be matched to the waveform basis functions. In our study of modulation, these basis functions were just the pulses shifted by integer multiples of the signal interval T. The matched filter effectively computes the inner product of the received,

[45] FSK and OFDM can be modelled similarly.

noisy waveform and each basis function. In doing so, orthogonality between the basis functions and $z_\perp(t)$ will ensure this component of the noise will be filtered out.

To illustrate this clearly, consider a non-dispersive channel and write the matched filter output sampled at time lT as

$$\begin{aligned} y_l &= y(lT) \\ &= \langle v, \phi_l \rangle \\ &= \langle x, \phi_l \rangle + \sum_{k=1}^{n} z_k \langle \phi_k, \phi_l \rangle + \langle z_\perp, \phi_l \rangle \\ &= x_l + z_l \end{aligned}$$

Observe that the matched filtering and subsequent sampling processes convert the noisy analogue waveform channel to a noisy analogue sequence channel. This is a huge simplification and an important step towards being able to determine the most likely transmitted signal vector $\boldsymbol{x} = (x_1, \ldots, x_n)$, which we will discuss in Section 6.

The result given here can be extended to the case where channels cause dispersion. However, the details are subtle and would not add much to enhance the discussion at this point. Hence, they are omitted.

One more crucial observation that can be made through the analysis given above is that the matched filter removes a significant amount of noise by projecting the received waveform onto a smaller vector space. As a very simple example of this concept, consider the case where a single PAM signal is transmitted over T seconds at passband. In a practical system, the baseband signal is real, but noise will occupy both the real and imaginary baseband channels, since there will be both in-phase and quadrature components in the passband waveform [46]. Of course, the noise that affects the quadrature component of the waveform has no bearing on the detector performance. The demodulator filter is matched to the in-phase PAM pulse waveform, and all noise on the quadrature channel is filtered out.

7.1.2 Random Noise Processes

The description of the noise waveform above assumed a specific function $z(t)$. In practice, noise is random. Sometimes, the noise will be minimal. At other times, it may severely alter the transmitted waveform. In Section 6 , we will be concerned with analysing how well systems can cope with noise. For that analysis, we will require a probabilistic model of a communication system. As far as noise is concerned, this suggests we should view the noise waveform as a random process $Z(t)$. Hence, for a given transmitted waveform, the additive noise in the channel will result in the random process

$$V(t) = x(t) + Z(t)$$

being passed to the input of the demodulator filter.

We know how to characterise random processes, in general. Here, we will consider some basic features of noise processes. First, we will always be

concerned with zero-mean processes, i.e.,

$$\mathrm{E}[Z(t)] = 0$$

for all t. This restriction does not narrow the discussion unnecessarily. After all, if we were interested in non-zero-mean processes, we could simply add a constant (deterministic) mean to $Z(t)$.

For a zero-mean process, we are usually interested in characterising the correlation properties of the process via the autocorrelation function. For the noise process, this is just

$$\mathrm{R}_Z(t_1, t_2) = \mathrm{E}[Z(t_1) Z(t_2)]$$

[46] As this is a PAM waveform, the quadrature signal component is ideally zero, but noise will still lie in the quadrature vector space.

Recall that a random process is stationary if the joint distribution of a set of (random) samples of the process does not depend on the actual sample times, but only on the differences between the sample times. A zero-mean process is said to be wide-sense stationary (WSS) if its autocorrelation only depends on the time difference $t - \tau$, i.e., $\mathrm{R}_Z(t_1, t_2) = \mathrm{R}_Z(t_1 - t_2, 0)$. If $Z(t)$ is widesense stationary, then we can compute its power spectral density by taking the Fourier transform

$$\mathrm{S}_Z(f) = \int_{-\infty}^{\infty} \mathrm{R}_Z(\tau) e^{-\mathrm{i}2\pi f\tau} \mathrm{d}\tau$$

where we have replaced the cumbersome notation $\mathrm{R}_Z(t_1 - t_2, 0)$ with $\mathrm{R}_Z(\tau)$ and have used $\tau = t_1 - t_2$ to denote the difference in sample times.

By applying the orthonormal expansion model discussed for deterministic noise waveforms above, we can write a random noise process as

$$Z(t) = \sum_{k=-\infty}^{\infty} Z_k \phi_k(t)$$

where $\{Z_k\}$ is a set of noise rvs. Note that these rvs make $Z(t)$ a random process. Indeed, the functions $\{\phi_k(t)\}$ are deterministic.

Let us analyse the mean and autocorrelation function in light of this orthonormal expansion formulation. For the mean, we have

$$\mathrm{E}[Z(t)] = \mathrm{E}\left[\sum_{k=-\infty}^{\infty} Z_k \phi_k(t)\right] = \sum_{k=-\infty}^{\infty} \mathrm{E}[Z_k]\, \phi_k(t) = 0$$

where we have assumed the series converges in order to facilitate averaging term by term. Since this is an orthonormal expansion, we must have that $\mathrm{E}[Z_k] = 0$ for all k. If this condition were not true, this would imply that at least one function in the set $\{\phi_k(t)\}$ could be expressed as a linear combination of the others, which would violate the orthonormality assumption.

The autocorrelation function is given by

$$\begin{aligned}\mathrm{R}_Z(t_1, t_2) &= \mathrm{E}\left[\left(\sum_{k=-\infty}^{\infty} Z_k \phi_k(t_1)\right)\left(\sum_{l=-\infty}^{\infty} Z_l \phi_l(t_2)\right)\right] \\ &= \sum_{k=-\infty}^{\infty} \sum_{l=-\infty}^{\infty} \mathrm{E}[Z_k Z_l]\, \phi_k(t_1)\, \phi_l(t_2).\end{aligned}$$

Now, let us assume that $\{Z_k\}$ are i.i.d. with variance $\mathrm{E}\left[Z_k^2\right] = \sigma^2$ for all k. Then it follows that

$$\mathrm{E}[Z_k Z_l] = \mathrm{E}[Z_k]\, \mathrm{E}[Z_l] = 0, \quad k \neq l$$

and we can write

$$\mathrm{R}_Z(t_1, t_2) = \sigma^2 \sum_{k=-\infty}^{\infty} \phi_k(t_1)\, \phi_k(t_2)$$

Band-Limited Noise Processes

Consider the specific random process defined by

$$Z(t) = \sum_{k=-\infty}^{\infty} Z_k \,\mathrm{sinc}\left(\frac{t}{\Delta} - k\right)$$

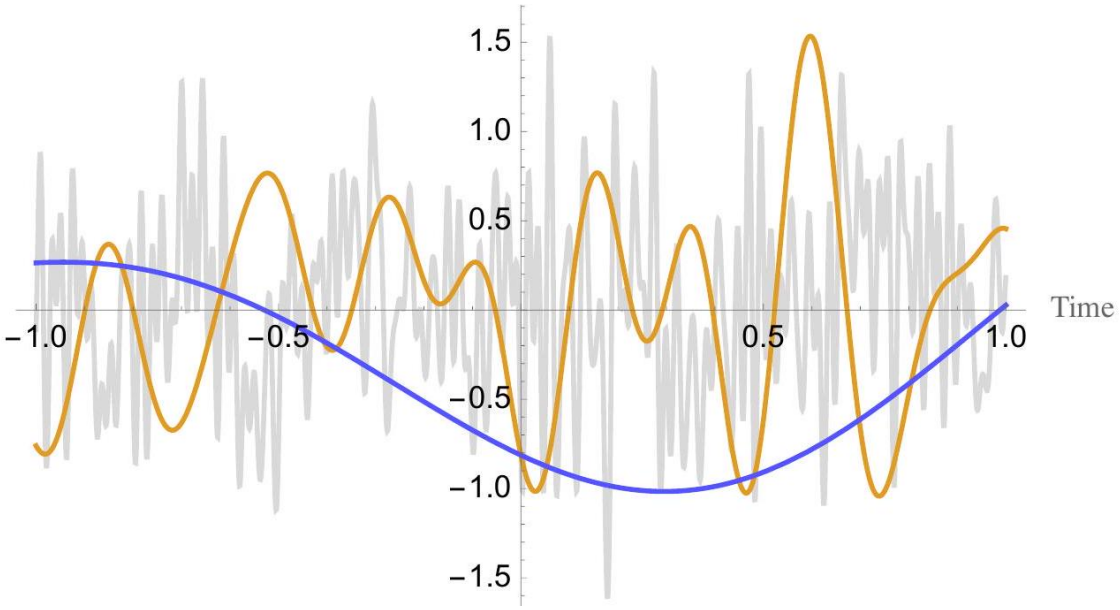


Figure 34: Examples of band-limited noise processes. Blue: $\Delta = 1.0$. Orange: $\Delta = 0.1$. Light gray: $\Delta = 0.01$. As Δ decreases, the correlation between nearby samples of $Z(t)$ also decreases.

where $\{Z_k\}$ are zero-mean and i.i.d. with variance σ^2, and Δ signifies the fundamental time shift of the sinc functions. It is important to recognise that the signal interval T is not necessarily equal to Δ, and the rvs $\{Z_k\}$ are not necessarily the same as the weights of the general expansion studied in the previous section. Each sample waveform $z(t)$ of this random process is band-limited to the interval $[-1/2\Delta, 1/2\Delta]$Hz. Note that this expansion is orthogonal. Figure 34 shows three sample waveforms corresponding to the noise process defined in eq. (15).

We claim that $Z(t)$ is a band-limited process. However, to properly study the spectrum of the process, we must first prove that $Z(t)$ is WSS (if, indeed, it is) and then compute its power spectral density. The autocorrelation function of $Z(t)$ is

$$\mathrm{R}_Z(t_1, t_2) = \sigma^2 \sum_{k=-\infty}^{\infty} \operatorname{sinc}\left(\frac{t_1}{\Delta} - k\right) \operatorname{sinc}\left(\frac{t_2}{\Delta} - k\right)$$

Now, recall the sampling equation for a function $u(t)$ band-limited to $[-1/2\Delta, 1/2\Delta]$ Hz :

$$u(t) = \sum_{k=-\infty}^{\infty} u(k\Delta) \operatorname{sinc}\left(\frac{t}{\Delta} - k\right)$$

Let us specify $u(t) = \operatorname{sinc}\left((t_1 - t)/\Delta\right)$. Then, remarkably,

$$\begin{aligned}
\operatorname{sinc}\left(\frac{t_1 - t_2}{\Delta}\right) &= u(t_2) \\
&= \sum_{k=-\infty}^{\infty} u(k\Delta) \operatorname{sinc}\left(\frac{t_2}{\Delta} - k\right) \\
&= \sum_{k=-\infty}^{\infty} \operatorname{sinc}\left(\frac{t_1}{\Delta} - k\right) \operatorname{sinc}\left(\frac{t_2}{\Delta} - k\right)
\end{aligned}$$

Comparing with our expression for $\mathrm{R}_Z(t_1, t_2)$ above, we see that the autocorrelation function can be written as

$$\mathrm{R}_Z(t_1, t_2) = \sigma^2 \operatorname{sinc}\left(\frac{t_1 - t_2}{\Delta}\right)$$

This function is only dependent on the difference in the sampling times t_1 and t_2. Hence, $Z(t)$ is WSS.

The power spectral density of this band-limited process now follows easily by computing the Fourier transform

$$\mathrm{S}_Z(f) = \sigma^2 \int_{-\infty}^{\infty} \mathrm{sinc}(\tau/\Delta) e^{-\mathrm{i}2\pi f\tau} \mathrm{d}\tau = \sigma^2 \Delta \sqcap (f\Delta)$$

Thus, the random process $Z(t)$ is band-limited to $[-1/2\Delta, 1/2\Delta]$Hz.

White Noise Processes

If the time-shifted sinc functions in the orthogonal expansion of the bandlimited noise process are closely spaced, then the bandwidth of the noise can be large. Also, observe that the power spectral density is flat over the bandwidth of the process. In practice, if Δ is sufficiently small, then we can view the band-limited process as a white noise process.

A random process $Z(t)$ is defined as a white noise process if its autocorrelation function satisfies

$$\mathrm{R}_Z(\tau) \propto \delta(\tau)$$

As indicated by this function, adjacent samples of the process are uncorrelated, no matter how closely spaced the samples are to each other. The power spectral density of a white noise process is a constant.

In reality, no white noise process exists. To see why, let us assume $Z(t)$ is a stationary, ergodic white noise process, and let us attempt to calculate the average power of $Z(t)$. This is given by

$$\mathrm{E}\left[Z(t)^2\right] = \mathrm{R}_Z(0) \propto \delta(0)$$

which is infinite.

Although white noise processes seem to violate the laws of physics, they can be very useful computational tools. If Δ is small enough - in particular, if $\Delta \ll T$, where T is the signal interval of a modulated waveform - then we can approximate the band-limited noise process discussed above by a white noise process. This will certainly make calculations simpler moving forward.

As a final note about the white noise process, its power spectral density is conventionally said to be

$$\frac{N_0}{2} \text{ watts per Hz}$$

The parameter N_0 is defined by the power of the noise process. From our discussion of the band-limited process, it must be the case that

$$\frac{N_0}{2} = \sigma^2 \Delta$$

where σ^2 is the power of the band-limited noise process (see eq. (16) and set $t_1 = t_2$). As Δ decreases, the baseband bandwidth $W = 1/2\Delta$ of the process increases, but the white noise power spectral density is constant. Thus, it must be the case that the power of the band-limited noise process increases linearly
with bandwidth:

$$\sigma^2 = W N_0$$

Filtering Noise Processes

Consider the additive noise channel with output waveform

$$V(t) = x(t) + Z(t).$$

In this model, we view the noise waveform as a random process; the transmitted waveform is deterministic. The received message $V(t)$ is then passed through a matched filter to facilitate

demodulation. We have thoroughly covered what happens to the function $x(t)$. Here, we will focus our attention on the filtered noise process.

We might write the noise at the output of the matched filter as[47]

$$\langle Z, \phi_l \rangle = \int_{-\infty}^{\infty} Z(t)\phi_l^*(t)\mathrm{d}t$$

However, we should take care when doing so. Consider the case where $Z(t)$ is a white noise process. Technically, $Z(t)$ would be discontinuous everywhere with probability one [48]. Trying to integrate such a function using the standard tools that we are familiar with would lead to confusion.

The good news is that these potential issues associated with filtering noise processes can be overcome with some reasonable approximations and a bit of care. In the end, we can usually assume the noise process is white and proceed with the calculus of filtering in a fairly standard way. Before jumping to this end, let us first consider the filtering process more closely.

Assume that $Z(t)$ is more physically representative than a white noise process. If we consider a band-limited process, the lth filtered random noise

sample can be written as

$$

$$\begin{aligned}\langle Z, \phi_l \rangle &= \int_{-\infty}^{\infty} Z(t)\phi_l^*(t)\mathrm{d}t \\ &= \int_{-\infty}^{\infty} \left(\sum_{k=-\infty}^{\infty} Z_k \phi_k(t) \right) \phi_l^*(t)\mathrm{d}t \\ &= \int_{-\infty}^{\infty} \left(\sum_{k=1}^{n} Z_k \phi_k(t) + Z_\perp(t) \right) \phi_l^*(t)\mathrm{d}t\end{aligned}$$

$$where$$

$\{\phi_k(t)\}$ are band-limited orthonormal basis functions. Now, assuming we can integrate each term separately and sum the result, we have

\left\langle Z, \phi_{l}\right\rangle=\sum_{k=1}^{n} Z_{k}\left\langle\phi_{k}, \phi_{l}\right\rangle+\left\langle Z_{\perp}, \phi_{l}\right\rangle=Z_{l}

$$

Hence, the output of the matched filter is just an rv in the set that defines the noise process expanded in the basis of the signal waveform. Furthermore, even if the noise expansion is infinite dimensional, the fact that the matched filter projects the received waveform onto the space spanned by the transmitted waveform basis functions means we can effectively ignore most of the noise rvs. The l th filtered sample of the received waveform can now be written as

$$Y_l = x_l + Z_l$$

Zero-Mean White Gaussian Noise Process

We now have the foundation for a simple probabilistic model of the output of the matched filter. To make this model more useful, we must specify the distribution of $\{Z_k\}$ in whatever basis we are interested in. In Section 1.4, we provided a physical argument for modelling $Z(t)$ as a Gaussian process. Here, we define such a process and give some details about the filtered noise rvs that will help us to analyse receiver performance later. The following three succinct definitions lead directly to the notion of a Gaussian process.

[47] Remember that the matched-filter effectively projects the waveform onto a basis function.

[48] This can be verified by recognising that $\mathrm{R}_Z(\tau) \propto \delta(\tau)$ implies noise samples spaced apart by an infinitesimal amount will be completely uncorrelated; and, if the support of the noise process is continuous, there is a zero probability of $Z(t)$ holding its value beyond a single instant in time.

Definition 5.1 (Gaussian Distribution). An rv Z is said to be Gaussian distributed with mean μ and variance σ^2, abbreviated $Z \sim \mathcal{N}\left(\mu, \sigma^2\right)$, if the probability of observing the event $\{Z \leq z\}$ is given by

$$\mathrm{P}(Z \leq z) = \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{z} e^{-(t-\mu)^2/2\sigma^2} \, \mathrm{d}t$$

In this case, the probability density function of Z is given by

$$f_Z(z) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(z-\mu)^2/2\sigma^2}, \quad z \in \mathbb{R}$$

Definition 5.2 (Jointly Gaussian). The rvs $Z_1, \ldots, Z_n$ are said to be jointly Gaussian if for any real constants $\alpha_1, \ldots, \alpha_n$, the linear combination $Z = \alpha_1 Z_1 + \cdots + \alpha_n Z_n$ is Gaussian distributed.

Definition 5.3 (Gaussian Process). A random process $Z(t)$ is called a Gaussian process if for all possible $t_1, \ldots, t_q$ and each positive integer q, the rvs $Z\left(t_1\right), \ldots, Z\left(t_q\right)$ are jointly Gaussian.

In many communication systems, noise added to the received waveform is modelled as being a zero-mean white Gaussian process. This is the so-called white Gaussian noise (WGN) model, sometimes called the additive white Gaussian noise (AWGN) model. For this model, the following are true:

$$\begin{aligned}
\mathrm{E}[Z(t)] &= 0 && \text{(stationarity);} \\
\mathrm{R}_Z(\tau) &= \frac{N_0}{2}\delta(\tau) \Leftrightarrow \mathrm{S}_Z(f) = \frac{N_0}{2} && \text{(white noise)} \\
Z_l &= \langle Z, \phi_l \rangle \sim \mathcal{N}\left(0, N_0/2\right) && \text{(filtered Gaussian)} \\
f_{Z_{l_1},\ldots,Z_{l_n}}\left(z_{l_1}, \ldots, z_{l_n}\right) &= \prod_{i=1}^{n} f_{Z_{l_i}}\left(z_{l_i}\right) && \text{(independence).}
\end{aligned}$$

The first property illustrates the process is WSS, since the mean is zero for all t. The second property is a mathematical convenience, as discussed earlier. In practical systems, it is typically assumed the bandwidth of the noise is far greater than that of the transmitted waveform, with the noise power spectral
density being fairly constant across this smaller bandwidth. Thus, the white noise property follows.

The third property will be very useful in the next section when we compute the probability of detection error. We can think of the inner product as a linear combination of Gaussian rvs weighted by the basis function $\phi_l(t)$ at each value of t. By the definition of the Gaussian process, it follows that Z_l is Gaussian distributed.

In the last line of the properties given above, the function on the left is the joint probability density function of n noise rvs indexed from l_1 to l_n. It is understood that all of these indices are unique. The notation suggests that these indices can take any values subject to this restriction. Hence, the line effectively states that the joint density is the product of the marginal densities, which implies the rvs are independent.

7.1.3 Signal-to-Noise Ratio

We conclude this section on noisy channels with a discussion of the receiver signal-to-noise ratio (SNR). Consider the WGN channel described by the filter output equation

$$Y_l = x_l + Z_l$$

When analysing the SNR, the noise is always taken to be a random variable. This is due to the assumption that the filtered noise samples are independent and cannot be influenced deterministically by careful hardware design. The signal x_l is assumed to be deterministic here, but we will also consider the case where it is an rv, i.e., X_l.

For a deterministic PAM waveform where the time-shifted pulses $\{p(t- lT)\}$ form an orthonormal basis, the SNR of the l th filtered signal is given by the signal power divided by the average noise power [49]

$$\mathrm{SNR}_l = \frac{x_l^2}{\mathrm{E}\left[Z_l^2\right]} = \frac{2x_l^2}{N_0}$$

When we view the PAM waveform as a WSS random process constructed from the rvs $X_1, X_2, \ldots$, then the SNR is

$$\mathrm{SNR} = \frac{\mathrm{E}\left[X_l^2\right]}{\mathrm{E}\left[Z_l^2\right]} = \frac{2E_\mathrm{s}}{N_0}$$

The average energy per signal E_s was discussed in Section 3.3.2. Note that we have omitted the signal index l here due to stationarity.

We can write a similar expression for QAM waveforms. The only difference is that the baseband signals are complex values, and WGN affects both of these values independently [50]. We label the noise rv that affects the real component $Z_{k,r} \sim \mathcal{N}(0, N_0/2)$ and the noise rv that affects the imaginary component $Z_{k,i} \sim \mathcal{N}(0, N_0/2)$ for all k. For a deterministic QAM waveform, the SNR of the l th filtered signal is then

$$\mathrm{SNR}_l = \frac{|x_l|^2}{\mathrm{E}\left[Z_{l,r}^2 + Z_{l,\mathrm{i}}^2\right]} = \frac{|x_l|^2}{N_0}$$

And, for the random QAM waveform, the SNR is

$$\mathrm{SNR} = \frac{\mathrm{E}\left[|x_l|^2\right]}{\mathrm{E}\left[Z_{l,\mathrm{r}}^2 + Z_{l,\mathrm{i}}^2\right]} = \frac{E_\mathrm{s}}{N_0}$$

Note that the factor of two is missing relative to the real-valued PAM waveform.

Optimality of the Matched Filter In Section 3.3.5, we showed that matching the demodulator filter response to the transmitted pulse waveform leads to a simple demodulation process, whereby the received waveform is projected onto a given basis, thus isolating the corresponding signal. If the transmitted pulses are time-shifted orthonormal functions, then the Nyquist criterion is satisfied (see Corollary 3.3). We did not, however, show that matched filtering

is optimal in the sense that it maximised the SNR. Here, we tie up this loose end by illustrating that the matched filter design leads to the highest possible SNR at the output of the demodulator filter.

We consider a deterministic PAM waveform $x(t)$ and zero-mean WGN $Z(t)$ at the receiver. The basic AWGN model is adopted:

$$\hat{Y}_l = \hat{x}_l + \hat{Z}_l$$

The "hat" notation is used to signify that the signal and noise samples at the output of the demodulator are not necessarily the same as the respective signal and noise quantities that constitute the weights in the orthonormal expansions of $x(t)$ and $Z(t)$. After all, we have not yet shown that the filter should be matched to the transmit pulse.

Let $q(t)$ denote the demodulator filter impulse response, and let $p(t)$ be the transmit pulse waveform. The SNR corresponding to the lth signal at the output of the demodulator filter is

$$\mathrm{SNR}_l = \frac{\hat{x}_l^2}{\mathrm{E}\left[\hat{Z}_l^2\right]}$$

where

$$\hat{x}_l = \int_{-\infty}^{\infty} x(\tau) q(lT - \tau) \mathrm{d}\tau$$

[49] Do not become confused about whether these quantities specify energy or power. The difference is only a constant (the signal period), which can easily be incorporated for a given model if needed.

[50] We will not concern ourselves with the intricacies of complex Gaussian rvs here. Suffice to say that the real and imaginary components of a QAM waveform can be viewed separately, and independent noise is added to each component. See Gallager (2008) for more details on Gaussian random variables and vectors.

with

$$x(t) = \sum_{k=-\infty}^{\infty} x_k p(t - kT)$$

denoting the sequence of time-shifted weighted pulses and

$$\hat{Z}_l = \int_{-\infty}^{\infty} Z(\tau) q(lT - \tau) \mathrm{d}\tau$$

We will now ignore ISI and focus only on the lth signal. The energy of this signal is

$$\hat{x}_l^2 = \left(\int_{-\infty}^{\infty} x_l p(t - lT) q(lT - t) \mathrm{d}t \right)^2$$

which is just the squared inner product of $p(t - lT)$ and $q(lT - t)$ weighted by x_l^2. By the Cauchy-Schwarz inequality (see Theorem 3.1 on page 47), we can bound this energy as follows:

$$\begin{aligned} \hat{x}_l^2 &\leq x_l^2 \int_{-\infty}^{\infty} (p(t - lT))^2 \ \mathrm{d}t \int_{-\infty}^{\infty} (q(lT - t))^2 \ \mathrm{d}t \\ &= x_l^2 \int_{-\infty}^{\infty} (p(t))^2 \ \mathrm{d}t \int_{-\infty}^{\infty} (q(t))^2 \ \mathrm{d}t \end{aligned}$$

The upper bound is met with equality if $q(t) \propto p(-t)$. Thus, it seems that we should, indeed, match the filter response to the transmitted pulse waveform. If the time-shifted pulse waveforms are orthogonal to one another, then our choice to ignore ISI seems justified and will not affect this result.

Before fully concluding that the matched filter approach maximises SNR, we still need to consider the filtered noise. We have

$$\mathrm{E}\left[\hat{Z}_l^2\right] = \mathrm{E}\left[\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} Z(t) Z(\tau) q(lT - t) q(lT - \tau) \mathrm{d}\tau \mathrm{d}t \right]$$

and since $\mathrm{E}[Z(t)Z(\tau)] = \frac{N_0}{2} \delta(t - \tau)$, we can write

$$\mathrm{E}\left[\hat{Z}_l^2\right] = \frac{N_0}{2} \int_{-\infty}^{\infty} (q(lT - t))^2 \ \mathrm{d}t = \frac{N_0}{2} \int_{-\infty}^{\infty} (q(t))^2 \ \mathrm{d}t$$

Now, writing the SNR expression out in full, we have

$$\mathrm{SNR}_l \leq \frac{x_l^2 \int_{-\infty}^{\infty} (p(t))^2 \ \mathrm{d}t \int_{-\infty}^{\infty} (q(t))^2 \ \mathrm{d}t}{\frac{N_0}{2} \int_{-\infty}^{\infty} (q(t))^2 \ \mathrm{d}t} = \frac{2x_l^2}{N_0}$$

where the last equality follows by assuming the pulse energy is one. Consequently, we do indeed see that the matched filter maximises the SNR at the output of the demodulator filter.

7.2 Detection and Performance

In previous sections, we fully explored the theory of modulation and matched filtering in the demodulator. We have not yet discussed how to optimally map filtered samples back to signals and, ultimately, to bit strings (see Figure 35). This mapping is effectively an estimation problem. Given the samples output from the matched filter, we wish to estimate the most likely transmitted signal sequence. In most communications literature, this problem is referred to as detection for historical reasons [51].

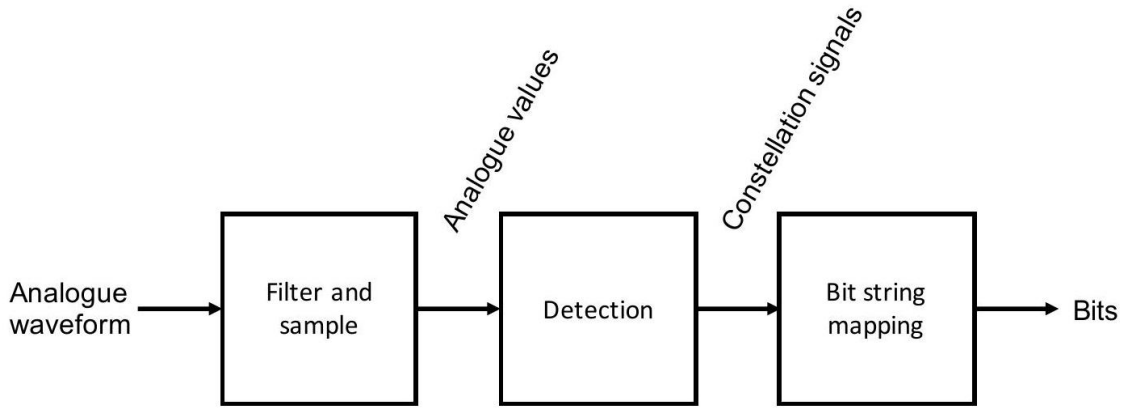


Figure 35: Block diagram of demodulator processes.

In this section, we will develop the optimal theory of detection for some of the modulation schemes we discussed earlier. We will also analyse the detector performance by quantifying the probability of symbol error for several techniques commonly used in practice.

7.2.1 Maximum A Posteriori Probability (MAP) Rule

Consider the AWGN channel model

$$y = x + z$$

where x is a scalar transmitted signal, which is unknown at the receiver, and z is a sample from a zero-mean WGN process with variance $N_0/2$ per real dimension. In other words, if x is a PAM signal, the rv Z that gives rise to the observed noise z is a real-valued Gaussian rv with variance $N_0/2$, and if x is a QAM signal, $Z = Z_\mathrm{r} + \mathrm{i}Z_\mathrm{i}$ is a complex-valued Gaussian rv, where Z_r and Z_i are each real-valued Gaussian with variance $N_0/2$. To simplify matters, we will focus on the real-valued model.

Suppose x can take any value in the M-ary constellation $\mathcal{A}$. The aim of the detector is to determine the most likely transmitted signal out of all signals in $\mathcal{A}$ given only the observation Y. A more mathematical way of putting this is that the detector seeks the estimate $\hat{x}$ that maximises the a posteriori probability that x was transmitted given that y was received. This detection rule is succinctly captured by the following statement:

$$\hat{x} = \arg\max_{x \in \mathcal{A}} \mathrm{P}(X = x \mid Y = y)$$

Eq. (17) is known as the maximum a posteriori probability (MAP) detection rule. Notice that the statement perfectly captures what we intuitively see as the best estimate that a detector can give for the transmitted signal x. Also, observe that the MAP rule converts a given sample of the AWGN model to a system containing rvs described by

$$Y = X + Z$$

The a posteriori probability $\mathrm{P}(X = x \mid Y = y)$, which we will sometimes write as $\mathrm{P}(x \mid y)$ when the meaning is clear, is the conditional probability that the signal x was transmitted given that the sample y is observed the output of the matched filter.

To solve eq. 17, the detector would need to cycle through all possible values of $x \in \mathcal{A}$ and evaluate $\mathrm{P}(x \mid y)$ for each one. It would assign $\hat{x}$ the signal that yields the maximum of these probabilities.

To evaluate $\mathrm{P}(x \mid y)$, we make use of Bayes' rule. Specifically, we can rewrite the MAP detection rule as

[51] The inconsistent terminology probably arises from radar research in which the fundamental goal is to detect the presence or absence of an object. Radar detection techniques are closely related to communication detection methods. Research on radar dominated communications research for a significant period around World War II.

$$\hat{x} = \arg\max_{x \in \mathcal{A}} \frac{f_{Y|X}(y \mid x)\mathrm{P}(x)}{f_Y(y)}$$

Here, $\mathrm{P}(x)$ is the a priori probability that the transmitted signal was x, $f_Y(y)$ is the probability density function of Y evaluated at the value y, and $f_{Y|X}(y \mid x)$ is the conditional probability density function of Y given X evaluated at y with $X = x$. Density functions are used instead of probabilities in this form of the MAP rule, because Y is a continuous rv owing to the fact that the WGN rv Z is continuous.

This version of the MAP rule can be simplified further by recognising that $f_Y(y)$ does not depend on x. In effect, it is a constant that affects each evaluation of the a posteriori probability equally. Removing it clarifies the rule:

$$\hat{x} = \arg\max_{x \in \mathcal{A}} f_{Y|X}(y \mid x)\mathrm{P}(x)$$

Now, we see that the rule only depends on the conditional density that the receiver observes y given x was transmitted and the a priori probability that x was conveyed in the first place.

If all signals in $\mathcal{A}$ are transmitted with equal probability, we can omit $\mathrm{P}(x)$ from the MAP rule evaluation and only base our decision on the likelihood function $f_{Y|X}(y \mid x)$. In this case, the MAP detector is sometimes called the maximum likelihood (ML) detector.

7.2.2 MAP Detection for Binary Antipodal Signals

Let $\mathcal{A} = \{-a, a\}$, where $a \in \mathbb{R}$ signifies the amplitude of the chosen signal. This constellation represents binary antipodal signals. Here, we will first discuss the MAP detection rule for such signals and subsequently analyse the probability that a binary antipodal signal is detected in error when transmitted through a non-dispersive WGN channel.

The MAP detection rule is given in the previous section. By specialising
that rule to the binary antipodal case, we see that the following binary decision rule arises:

$$f_{Y|X}(y \mid a)\mathrm{P}(a) \underset{\hat{x}=a}{\overset{\hat{x}=-a}{\lessgtr}} f_{Y|X}(y \mid -a)\mathrm{P}(-a)$$

or, alternatively,

$$\frac{f_{Y|X}(y \mid a)}{f_{Y|X}(y \mid -a)} \underset{\hat{x}=a}{\overset{\hat{x}=-a}{\lessgtr}} \frac{\mathsf{P}(-a)}{\mathsf{P}(a)}.$$

When transmission occurs through an AWGN channel, it is often easier to work with the natural logarithms of the left-hand and right-hand sides of these relations:

$$\ln\left(\frac{f_{Y|X}(y \mid a)}{f_{Y|X}(y \mid -a)}\right) \underset{\hat{x}=a}{\overset{\hat{x}=-a}{\lessgtr}} \ln\left(\frac{\mathsf{P}(-a)}{\mathsf{P}(a)}\right).$$

When both signals are equally probable a priori, this simplifies to

$$\ln\left(\frac{f_{Y|X}(y \mid a)}{f_{Y|X}(y \mid -a)}\right) \underset{\hat{x}=a}{\overset{\hat{x}=-a}{\lessgtr}} 0.$$

Detection in WGN Channels

Let us now develop the discussion for WGN channels. Given the transmitted signal $X = x$, the observed sample can be written as

$$Y = x + Z$$

This implies $Y \sim \mathcal{N}(x, N_0/2)$. Hence, we can write the conditional density

$$f_{Y|X}(y \mid x) = \frac{1}{\sqrt{\pi N_0}} e^{-(y-x)^2/N_0}$$

One can think of the MAP detector as simply evaluating the conditional Gaussian densities corresponding to the two antipodal signals having been transmitted, weighting them by the a priori probabilities, and selecting the signal that yields highest result (see Figure 36).

We can obtain a bit more insight about how MAP detection works by substituting the conditional densities into the binary decision rule above. This

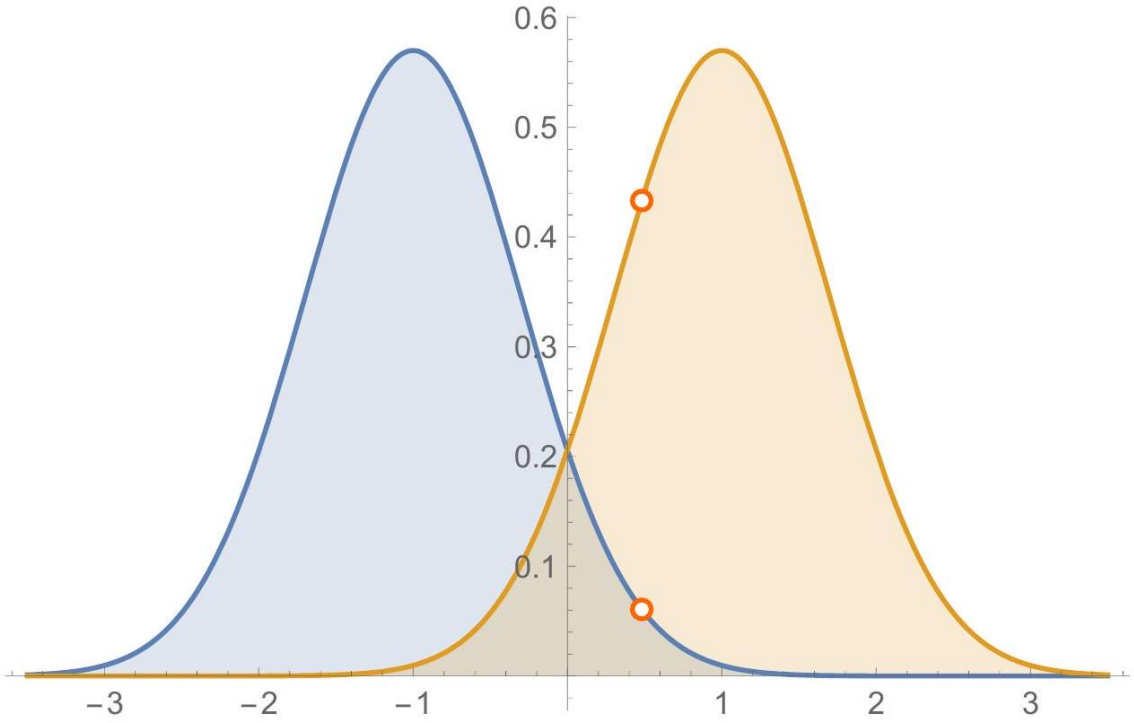


Figure 36: Example of MAP detection for an AWGN channel with $N_0/2 \simeq 0.5$. The antipodal constellation is $\mathcal{A} = \{-1, +1\}$, and it is assumed the signals are equally probable a priori. The filtered sample at the receiver is $y \simeq 0.5$, and the two conditional densities are shown in blue and orange. The orange density corresponding to the case where $X = +1$ is clearly larger at

the sampled value, and hence the detector determines that +1 was transmitted. If, in fact, the transmitter sent -1 , then a detection error would have occurred in this instance.
gives us the following MAP rule for binary antipodal signals:

$$-(y-a)^2/N_0 + (y+a)^2/N_0 \underset{\hat{x}=a}{\overset{x}{=}} -a\ b$$

where we have defined $b = \ln(\mathrm{P}(-a)) - \ln(\mathrm{P}(a))$ for simplicity. Expanding the quadratics and simplifying yields the rule

$$y \underset{\hat{x}=a}{\overset{\hat{x}=-a}{\lessgtr}} \frac{4a}$$

Effectively, this rule tells us that the received sample should be mapped to the positive signal if it is far enough to the right on the real line, and it should be mapped to the negative signal if it is not. When a and $-a$ are equally probable, the value that demarcates the two decision regions is the origin (i.e., $b = 0$). Otherwise, the decision boundary is $bN_0/4a$. As the SNR increases, the decision boundary moves closer to the origin regardless of the balance of probabilities for the two signals in $\mathcal{A}$. The decision regions and the decision boundary for the equiprobable example depicted in Figure 36 can be seen clearly as indicated by the intersection of the conditional densities.

Performance in WGN Channels

Now we turn our attention to analysing the performance of an optimal detector of binary antipodal signals in a WGN channel. For this analysis, we consider the scenario where we transmit many signals sequentially and independently. The signals are chosen randomly; in a practical system, of course, they may appear to be random, but they would represent a source encoded bit sequence. The matched filter in the demodulator outputs many independent samples, each one affected by independent noise. This set-up leads us to consider the signal and the noise as rvs. Let X denote the signal rv and let $Z \sim \mathcal{N}(0, N_0/2)$ denote the noise rv. The output of the demodulator is given by

$$Y = X + Z$$

Performance is quantified in terms of the probability of detection error. Since the signals and noise in our model are i.i.d., we can study the probability of detection error by considering the simple demodulator filter output model given above. The detector implements the MAP rule on the filtered sample Y. An error occurs if one of the two mutually exclusive events is observed:

- the signal a is transmitted, but $-a$ is detected;
- the signal $-a$ is transmitted, but a is detected.

Hence, the probability of detection error can be decomposed in the following way by using the law of conditional probabilities:

$$P_{\mathrm{e}} = \mathrm{P}(\hat{X} = a \mid X = -a)\mathrm{P}(X = -a) + \mathrm{P}(\hat{X} = -a \mid X = a)\mathrm{P}(X = a)$$

The a priori probabilities are assumed to be known. Thus, our goal in this analysis is calculate the conditional error probabilities $\mathrm{P}(\hat{X} = a \mid X = -a)$ and $\mathrm{P}(\hat{X} = -a \mid X = a)$.

Let us begin with $\mathrm{P}(\hat{X} = a \mid X = -a)$. The MAP detector will output the signal estimate $\hat{x} = a$ if the following inequality is satisfied:

$$y > \frac{bN_0}{4a}$$

Consequently, the conditional probability that the detector incorrectly selects a as the signal estimate given that $-a$ was transmitted can equivalently be written as

$$\mathrm{P}(\hat{X} = a \mid X = -a) = \mathrm{P}\left(Y > \frac{bN_0}{4a} \middle| X = -a \right)$$

To evaluate this probability, we recognise that, conditioned on $X = -a, Y = -a + Z$ is a Gaussian rv with mean $-a$ and variance $N_0/2$. By substituting for Y in the probability expression on the right-hand side of the equation above
and rearranging, we have that

$$\mathrm{P}(\hat{X} = a \mid X = -a) = \mathrm{P}\left(Z > a + \frac{bN_0}{4a}\right)$$

Now we can standardise the Gaussian rv Z by dividing by $\sqrt{N_0/2}$ to obtain

$$\mathrm{P}(\hat{X} = a \mid X = -a) = \mathrm{P}\left(N > \sqrt{\frac{2a^2}{N_0}} + \frac{b\sqrt{N_0}}{2\sqrt{2}a}\right)$$

where $N \sim \mathcal{N}(0, 1)$ is a standard Gaussian rv. Thus, we have shown that the conditional error probability is the area under the upper tail of a standard Gaussian density function.

In the communication literature, the upper tail distribution is typically written in terms of the so-called Q function:

$$Q(x) = \frac{1}{\sqrt{2\pi}} \int_x^\infty e^{-t^2/2} \, \mathrm{d}t$$

Figure 37 provides an illustration of the Q function. Using this notation, we can write the conditional error probability as

$$\mathrm{P}(\hat{X} = a \mid X = -a) = Q\left(\sqrt{\frac{2a^2}{N_0}} + \frac{b\sqrt{N_0}}{2\sqrt{2}a}\right)$$

Additionally, we see that the average energy of each signal is just $E_\mathrm{s} = a^2$, which leads to

$$\begin{aligned}\mathrm{P}(\hat{X} = a \mid X = -a) &= Q\left(\sqrt{\frac{2E_\mathrm{s}}{N_0}} + \frac{b}{2}\sqrt{\frac{N_0}{2E_\mathrm{s}}}\right) \\ &= Q\left(\sqrt{\mathrm{SNR}} + \frac{b}{2\sqrt{\mathrm{SNR}}}\right)\end{aligned}$$

Observe that if both a and $-a$ are equally probable a priori, then this simplifies

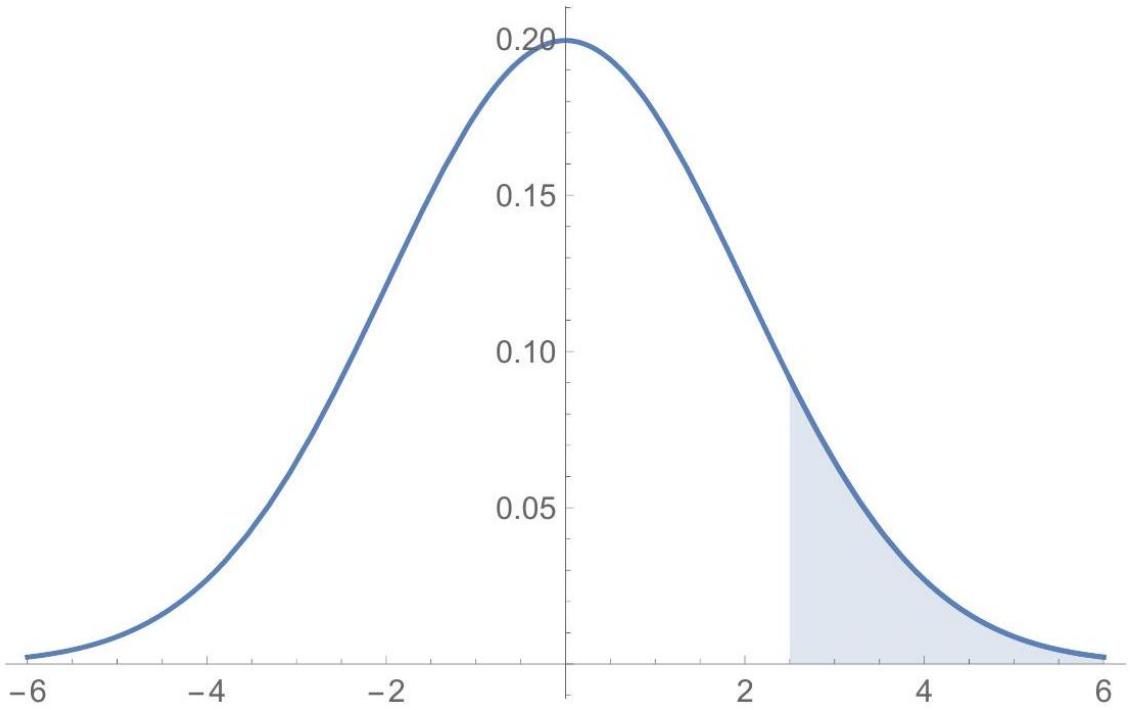


Figure 37: The shaded area is approximately the value of $Q(2.5)$.

to

$$\mathrm{P}(\hat{X} = a \mid X = -a) = Q(\sqrt{\mathrm{SNR}})$$

Equally, when the SNR is high, we have the approximation

$$\mathrm{P}(\hat{X} = a \mid X = -a) \simeq Q(\sqrt{\mathrm{SNR}})$$

It is worth noting that as the SNR increases, the conditional probability of error decreases exponentially quickly.

Now, we turn our attention to the remaining conditional error probability $\mathrm{P}(\hat{X} = -a \mid X = a)$. Using the MAP detection rule, we an write

$$\mathrm{P}(\hat{X} = -a \mid X = a) = \mathrm{P}\left(Y < \frac{bN_0}{4a} \middle| X = a\right) = \mathrm{P}\left(Z < -a + \frac{bN_0}{4a}\right)$$

Standardising the Gaussian rv yields

$$\mathrm{P}(\hat{X} = -a \mid X = a) = \mathrm{P}\left(N < -\sqrt{\frac{2a^2}{N_0}} + \frac{b\sqrt{N_0}}{2\sqrt{2}a}\right)$$

Since $1 - Q(-x) = Q(x)$, we can write

$$\begin{aligned}\mathrm{P}(\hat{X} = -a \mid X = a) &= Q\left(\sqrt{\frac{2a^2}{N_0}} - \frac{b\sqrt{N_0}}{2\sqrt{2}a}\right) \\ &= Q\left(\sqrt{\mathrm{SNR}} - \frac{b}{2\sqrt{\mathrm{SNR}}}\right)\end{aligned}$$

Again, if both a and $-a$ are equally probable a priori, then this conditional error probability simplifies to

$$\mathrm{P}(\hat{X} = -a \mid X = a) = Q(\sqrt{\mathrm{SNR}})$$

Let $p_a = \mathrm{P}(X = a)$, such that $\mathrm{P}(X = -a) = 1 - p_a$. The conditional error probability expressions given above can now be substituted into the equation for the probability of detection error to obtain

$$P_\mathrm{e} = p_a Q\left(\sqrt{\mathrm{SNR}} - \frac{b}{2\sqrt{\mathrm{SNR}}}\right) + (1 - p_a)\, Q\left(\sqrt{\mathrm{SNR}} + \frac{b}{2\sqrt{\mathrm{SNR}}}\right)$$

If a and $-a$ are equally probable, then we have the simple expression

$$P_\mathrm{e} = Q(\sqrt{\mathrm{SNR}})$$

7.2.3 MAP Detection for Binary Nonantipodal Signals

We now generalise the previous discussion to nonantipodal signal constellations. An example of a nonantipodal constellation is shown in Figure 38. Let $\mathcal{A} = \{a_1, a_2\}$, where $a_1, a_2 \in \mathbb{R}$ and $a_1 > a_2$. The MAP detection rule, specialised to this constellation, is

$$f_{Y|X}(y \mid a_1)\, \mathrm{P}(a_1) \underset{\hat{x}=a_1}{\overset{\hat{x}=a_2}{\lessgtr}} f_{Y|X}(y \mid a_2)\, \mathrm{P}(a_2)$$

Figure 38: A binary nonantipodal constellation corresponding to the common modulation technique known as on-off keying (OOK).

or, alternatively,

$$\ln\left(\frac{f_{Y|X}(y \mid a_1)}{f_{Y|X}(y \mid a_2)}\right) \underset{\hat{x}=a_1}{\overset{\hat{x}=a_2}{\lessgtr}} b$$

where $b = \ln(\mathrm{P}(a_2)) - \ln(\mathrm{P}(a_1))$.
For a zero-mean WGN channel with variance $N_0/2$, we can carry out the same calculations that we did for the antipodal case to simplify the MAP detection rule and to analyse the probability of detection error. However, we will use a more insightful method to calculate the error probability here.

First, note that the midpoint of the constellation signals is

$$\mu = \frac{a_1 + a_2}{2}$$

Also, we can define an antipodal constellation, say $\mathcal{A}' = \{a', -a'\}$ where

$$a' = a_1 - \mu \quad \text{and} \quad -a' = a_2 - \mu$$

Similarly, we can define an rvX' that takes values from this antipodal constellation, and we can relate this rv to the nonantipodal rv X through the equation

$$X' = X - \mu$$

Now, we can rewrite the AWGN model for the nonantipodal system as

$$Y = X + Z = X' + \mu + Z$$

This suggests that, after the filter, we can simply remove the deterministic bias μ and still have sufficient information to estimate the nonantipodal signal that
was transmitted. The system model with a removed bias is just

$$Y' = Y - \mu = X' + Z$$

where X' is an rv that takes values from the antipodal constellation $\mathcal{A}'$ and $Z \sim \mathcal{N}(0, N_0/2)$. By estimating the most likely antipodal signal from $\mathcal{A}'$, we can easily map back to the nonantipodal signal by adding μ.

In this simple analysis, we have effectively just shifted the origin of the constellation to be the midpoint μ. Since the shift is constant, and both the transmitter and the receiver know this shift, we have lost nothing in performing the translation. However, we will have gained a substantial advantage in terms of performance analysis, as we will now show.

Earlier, we analysed the probability of detection error for a binary antipodal signal. We wrote this probability as

$$P_{\mathrm{e}} = p_{a'} Q\left(\sqrt{\mathrm{SNR}'} - \frac{b}{2\sqrt{\mathrm{SNR}'}}\right) + (1 - p_{a'}) Q\left(\sqrt{\mathrm{SNR}'} + \frac{b}{2\sqrt{\mathrm{SNR}'}}\right)$$

where $p_{a'}$ is the a priori probability that the positive signal is transmitted, b is the logarithm of the ratio of probabilities of the two antipodal signals, and $\mathrm{SNR}' = 2E'_{\mathrm{s}}/N_0$, with $E'_{\mathrm{s}} = (a')^2$ denoting the energy of an antipodal signal. This expression for the error probability holds for the translated model shown in eq. (18). We can express the parameters $p_{a'}, b$, and E'_{s} in terms of the nonantipodal signals easily:

$$\begin{aligned} p_{a'} &= \mathrm{P}(X' = a') = \mathrm{P}(X = a_1) \\ b &= \ln(\mathrm{P}(-a')) - \ln(\mathrm{P}(a')) = \ln(\mathrm{P}(a_2)) - \ln(\mathrm{P}(a_1)) \\ E'_{\mathrm{s}} &= (a')^2 = \frac{(a_1 - a_2)^2}{4}. \end{aligned}$$

It is important to write the error probability in terms of the actual SNR of the nonantipodal transmission. To do this, recognise that the average energy
per nonantipodal signal can be written as

$$E_{\mathbf{s}} = p_{a'}(\mu + a')^2 + (1 - p_{a'})(\mu - a')^2 = (a')^2 + \mu^2 - 2a'\mu(1 - 2p_{a'})$$

For uniformly distributed signals, the average energy is the sum of the energy expended to transmit a binary signal and the energy expended to transmit the bias. Since the bias conveys

no information, this may seem like a waste of resource. In fact, there are good reasons to use a nonantipodal constellation in some cases. For example, the constant bias creates a spike (delta function) in the spectrum of the waveform that carries the signal, which can be used to lock onto the carrier frequency at the receiver in a passband transmission [52]. The proportion of energy used to convey information is

$$\gamma = \frac{(a')^2}{(a')^2 + \mu^2 - 2a'\mu\,(1 - 2p_{a'})} = \frac{E'_{\mathrm{s}}}{E_{\mathrm{s}}}$$

Consequently, we can write the probability of detection error for nonantipodal signals as

$$P_{\mathrm{e}} = p_{a_1} Q\left(\sqrt{\gamma \mathrm{SNR}} - \frac{b}{2\sqrt{\gamma \mathrm{SNR}}}\right) + (1 - p_{a_1})\, Q\left(\sqrt{\gamma \mathrm{SNR}} + \frac{b}{2\sqrt{\gamma \mathrm{SNR}}}\right)$$

where $p_{a_1} = \mathrm{P}(X = a_1)$ and $\mathrm{SNR} = 2E_{\mathrm{s}}/N_0$. Typically, $\gamma < 1$, and we can immediately see the degradation in performance caused by the bias.

7.2.4 MAP Detection for M-ary Signals

In this section, we briefly cover the extension of the MAP detection framework to nonbinary signals. Suppose we have a constellation $\mathcal{A} = \{a_1, \ldots, a_M\}$, which clearly contains M signals. It will almost always be the case that M is a power of two, such that $m = \log M$ bits will be encoded in each signal. The signals in $\mathcal{A}$ can be real or complex.

Recall that the general MAP detection rule is

$$\hat{x} = \arg\max_{x \in \mathcal{A}} f_{Y|X}(y \mid x)\mathrm{P}(x)$$

For the real-valued WGN channel model

$$Y = X + Z$$

with $Z \in \mathcal{N}(0, N_0/2)$ and $a_k \in \mathbb{R}$ for $k = 1, \ldots, M$, the detection rule can be manipulated by taking the natural logarithm to give the equivalent rule

$$\hat{x} = \arg\min_{x \in \mathcal{A}} \left\{(y - x)^2 - N_0 \ln(\mathrm{P}(x))\right\}$$

Here, the maximisation over $\mathcal{A}$ has become a minimisation since we have taken the negative of the logarithm of $f_{Y|X}(y \mid x)\mathrm{P}(x)$. Also, we have ignored any terms that are independent of x and have multiplied everything by N_0.

If all signals are equally probable (i.e., $\mathrm{P}(x) = 1/M$ for all $x \in \mathcal{A}$), then the MAP rule reduces to

$$\hat{x} = \arg\min_{x \in \mathcal{A}} (y - x)^2$$

In other words, the best estimate of the transmitted signal is the one that is closest to the received, filtered sample y. Without going into details [53], it turns out this same rule can be applied for complex constellations when all signals are equally probable:

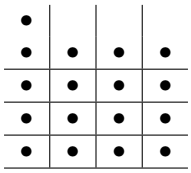
$$\hat{x} = \arg\min_{x \in \mathcal{A}} |y - x|^2$$

This minimum distance rule is only applicable when the transmitted signals are equally probable, i.e., when MAP detection is equivalent to ML detection.

Earlier, we briefly discussed decision boundaries and decision regions in the context of binary signalling. For complex signals, we can easily visu-

[52] Recall that the Fourier transform of a constant is a Dirac delta function.

[53] The details are straightforward, but would require us to develop the notion of complex Gaussian rvs. Although this is not particularly taxing, it is best not to worry too much about the rigour here, lest we lose sight of the main ideas.



(a) Standard 16-QAM Voronoi diagram. ses¿)

(b) Hexagonal 16-QAM Voronoi diagram.

Figure 39: Voronoi diagrams for QAM with $M = 16$.

alise these boundaries and regions under the minimum distance detection rule through the use of a Voronoi diagram. In this context, a Voronoi diagram is simply a partition of the plane into M regions, one for each constellation signal. Each region, called a Voronoi cell, is constructed such that it consists of points that are closer to the corresponding constellation signal than any other constellation signal in $\mathcal{A}$.

Illustrations of Voronoi diagrams for standard and hexagonal QAM constellations with $M = 16$ signals are shown in Figure 39. In each of these diagrams, in-phase and quadrature samples observed at the output of the matched filter would lie somewhere in the plane, and each sample would be mapped to the orange constellation signal that forms the nucleus of the region in which the sample resides. Each region is therefore a decision region, and each boundary is a decision boundary.

7.2.5 Sequence Detection

A typical communication system would transmit a sequence of signals by modulating a waveform, then demodulate the message at the receiver by using a matched filter and a detector. So far, we have ignored the idea of transmitting a sequence of signals and focused only on a single signal being communicated. This has allowed us to study the probability of detection error for binary and M-ary schemes. Here, we generalise the discussion to treat the detection of sequences of signals. We will cover non-dispersive and dispersive WGN channels separately.

Non-Dispersive Channels

Suppose ideal Nyquist pulses are used, such that the channel viewed from the output of the matched filter is non-dispersive. The lth matched filter output sample can be modelled as

$$Y_l = X_l + Z_l$$

Assume that a finite collection of n signals is transmitted. We can collect the n random transmitted signals in the random vector $\boldsymbol{X} = (X_1, \ldots, X_n)$. A particular choice of these signals is represented by the lower case notation $\boldsymbol{x} = (x_1, \ldots, x_n)$. Similarly, we can collect the n filter output samples in the random vector $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$, where a particular observed collection would be represented as $\boldsymbol{y} = (y_1, \ldots, y_n)$.

The optimal detector would wait for all n filter samples $\boldsymbol{y}$ to be received, then generate the most likely estimate $\hat{\boldsymbol{x}} = (\hat{x}_1, \ldots, \hat{x}_n)$ for the transmitted signal vector $\boldsymbol{x} = (x_1, \ldots, x_n)$ using the rule

$$\hat{\boldsymbol{x}} = \arg \max_{\boldsymbol{x} \in \mathcal{A}^n} f_{\boldsymbol{Y}|\boldsymbol{X}}(\boldsymbol{y} \mid \boldsymbol{x}) \mathrm{P}(\boldsymbol{x})$$

Here, the maximisation is over all possible length- n vectors of constellation signals [54]. Also, observe that the MAP detection rule requires us to evaluate the

joint conditional density $f_{\boldsymbol{Y}|\boldsymbol{X}}(\boldsymbol{y} \mid \boldsymbol{x})$ and the joint a priori distribution $\mathrm{P}(\boldsymbol{x})$, which may seem to be difficult. However, if the signal vector $\boldsymbol{x}$ is encoded in a waveform constructed as an orthonormal expansion, then each filter output is (statistically) independent of all others. Consequently, the joint conditional density factors into the product

$$f_{\boldsymbol{Y}|\boldsymbol{X}}(\boldsymbol{y} \mid \boldsymbol{x}) = f_{Y_1|X_1}(y_1 \mid x_1) \cdots f_{Y_n|X_n}(y_n \mid x_n)$$

[54] This is indicated by the notation $x \in \mathcal{A}^n$.

and the joint a priori distribution factors into

$$\mathrm{P}(\boldsymbol{x}) = \mathrm{P}\left(x_1\right) \cdots \mathrm{P}\left(x_n\right)$$

Consequently, the maximisation can be achieved by treating each signal interval separately. In this way, the large MAP sequence detection task is broken into n separate single signal detection tasks.

Dispersive Channels

If ideal Nyquist pulses are not used, the situation is considerably more complicated. Here, we will only mention the basic idea of how to cope with dispersive channels, without going into the mathematics. A full treatment would require considerably more explanation that is appropriate for this course. Interested students can find further information in Sections 3.4 and 14.7.1 of Sklar and Harris (2021) and Section 8.8.2 of Gallager (2008).

If the channel is dispersive, it can be shown that each sample at the output of the matched filter is a (discrete) convolution of the sampled channel impulse response (CIR) and the most recent transmitted signals. Thus, each output sample is affected by intersymbol interference (ISI). The goal is to use a group of output samples to estimate the group of signals that they are functionally related to. The MAP detection rule can be formulated in a similar way as discussed above (as an optimisation over vectors), but it is dependent upon the CIR as a result of the convolution. The CIR can be estimated in practice, so this, in itself, is not a problem.

The main issue encountered when executing the MAP rule in a dispersive channel is that the complexity of cycling through all possible transmitted signal vectors becomes huge as the length of the CIR grows. Specifically, for an M-ary signal constellation and a (discrete) CIR of length L, there are M^L possible signal vectors that must be considered for each MAP estimation. For binary signalling, this is not a problem, since L usually is not too large (e.g., $L \leq 10$ in some practical systems), and modern processors can run through the calculations quickly. However, for larger QAM constellations, the complexity becomes prohibitive. For example, for $M = 16$ and $L = 8$, over four million possible signal vectors would need to be considered for each signal estimation task.

To overcome the complexity issue, a dynamic programming method invented by Andrew Viterbi is typically used. The method effectively views the dispersive channel in a state-space format. At each time instant, the state of the system is defined as the collection of signals that comprise the ISI. In this sense, it may help to think of these signals as being "present" in the system at a given time, and during the next signal interval, one of the signals exits the system (because the discrete convolution progresses) and the state evolves through the inclusion of a new signal that contributes to the ISI. In this way, the state at a given time can evolve in M different ways only, since there are M constellation signals. The detector can therefore consider only these M possible evolutions and use a maximum likelihood (ML) rule to determine which state transition is the most probable. Over time, this method traces out the most likely path in a state diagram. At a given time, parts of the path that relate to the distant past can be fixed and alternatives can be discarded. But for recent state transition estimates, several possibilities can be retained in computer memory and used to update the signal sequence estimate. This algorithm is known as maximum likelihood sequence estimation (MLSE) or the Viterbi algorithm.

It should be noted that the Viterbi algorithm does not technically yield the MAP estimate for each signal, since it operates on a sequence. A separate algorithm for accomplishing this goal was developed in the 1970s by Bahl,

Cocke, Jelinek and Raviv [55]. The algorithm, which is called the BCJR algorithm, uses a rather intricate factorisation of the a posteriori probability of having transmitted a particular signal vector given a set of filtered samples. The important thing to note is that the BCJR algorithm and the Viterbi algorithm perform very similarly for most practical cases of interest, i.e., they yield a very similar probability of sequence detection error.

[55] Bahl, L., Cocke, J., Jelinek, F., & Raviv, J. (1974). Optimal decoding of linear codes for minimizing symbol error rate (corresp.). IEEE Transactions on Information Theory, 20(2), 284-287.

Chapter 8

Other Topics in Digital Communications

Chapter 9

Labs: MATLAB/ Python

9.1 Problems: OFDM Equalization

1. *OFDM Wirless LAN.* Use the Internet or any other source to find the following parameters for a 40 GHz 802.11n system.

   (a) Find the sample rate, the subcarrier spacing, FFT size, number of samples in the CP, number of occupied subcarriers and number of pilot sub-carriers.

   (b) Why is the DC carrier a null carrier?

   (c) Suppose you use 16-QAM on all data subcarriers (sub-carriers that are not used for pilots). What is the data rate?

2. *OFDM Numerology.* Find the parameters of an OFDM system with the following properties. The parameter should include the sample rate, the subcarrier spacing, FFT size, number of samples in the CP, and number of occupied subcarriers.

   (a) The maximum delay spread is 500 ns

   (b) The CP overhead is approximately 25%

   (c) The occupied bandwidth is approximately 40 MHz.

   (d) The FFT should be a power of 2.

3. *Effective SNR.* Suppose that

$$y[n] = hx[n] + w[n], \quad w[n] \sim \mathcal{C}N(0, N_0), \quad |x[n]|^2 = E_x,$$

   where h is a constant channel gain. Suppose the first N symbols are reference symbols, and we estimate the channel gain h via a simple average:

$$\widehat{h} = \frac{1}{N} \sum_{n=0}^{N-1} \frac{y[n]}{x[n]}.$$

   (a) Find the bias and variance of $\widehat{h}$ as a function of N.

   (b) For any subsequent symbol we can write,

$$y[n] = \widehat{h}x[n] + v[n], \quad v[n] = (h - \widehat{h})x[n] + w[n],$$

   where $v[n]$ is the *effective noise* including channel estimation error. Assuming $x[n]$ is independent of $w[n]$ what is the $E|v[n]|^2$, the effective noise variance.

   (c) What is the degradation in SNR in terms of the number of symbols N you use for reference signals?

4. *MSE error with a correlation.* Suppose that

$$y[n] = hx[n] + w[n], \quad w[n] \sim \mathcal{C}N(0, N_0), \quad |x[n]|^2 = 1.$$

To estimate $h[n]$, reference signals are placed once every N sub-carriers at locations $n = 0, N, 2N, \ldots$. On the reference signals, we compute the raw estimate: $\hat{h}_0[n] = y[n]/x[n]$. Suppose that we can model $h[n]$ as a stationary random process with $E(h[n]) = 0$ and $E(h[n]h^*[n-m]) = E_0 e^{-\alpha|m|}$ for some constants E_0 and m.

(a) Find the MSE of the raw channel estimate $\hat{h}_0[n] = y[n]/x[n]$ on the sub-carriers $n = 0, N, 2N, \cdots$. The MSE is $\mathbb{E}|\hat{h}_0[n] - h[n]|^2$.

(b) Suppose that for other sub-carriers, we estimate $\hat{h}[n] = \hat{h}_0[kN]$ where kN is the closest reference sub-carrier to n. Find the MSE of this nearest neighbor estimate $\hat{h}[n]$ as a function of the sub-carrier index n. This will be a periodic function in n.

(c) Suppose that the SNR E_0/N_0 is 20 dB and we want that the worst case MSE is bounded by:

$$E|\hat{h}[n] - h[n]|^2 \leq 2N_0,$$

for all sub-carriers n. That is, we want that the estimation error to be no more than 3dB the MSE on the reference sub-carriers. What is N in terms of α. That is, how closely do we need to place the reference sub-carriers as a function of the rate of variation of the random process.

5. *Bias and Variance with a frequency rotation.* Suppose that true OFDM channel is $h[n] = e^{in\theta}$ where θ is the frequency rotation per sub-carrier. We get measurements of the form,

$$y[n] = h[n]x[n] + w[n], \quad w[n] \sim \mathcal{C}N(0, N_0), \quad |x[n]|^2 = E_x.$$

(a) Find the bias and variance of the raw channel estimate $\hat{h}_0[n] = y[n]/x[n]$.

(b) Now consider a smoothed channel estimate,

$$\hat{h}[n] = \sum_{\ell=-L}^{L} \hat{h}_0[n].$$

Find the bias and variance as a function of L.

(c) Suppose that a channel delay $\tau = 0.5\mu$s relative to the beginning of the FFT window, and the sub-carrier spacing is Δf=15 kHz and the SNR $E_x/N_0 = 10$ dB. Use MATLAB to plot the bias squared, variance and MSE as a function of the window length L.