uc3m | Universidad Carlos III de Madrid
OpenCourseWare (OCW)

https://ocw.uc3m.es

Communication Theory

Bachelor in Telecommunication Technologies Engineering

Marcelino Lázaro

Teaching materials for *Communications Theory*, Bachelor in Telecommunication Technologies Engineering (and also, Bachelor in Sound and Image Engineering).

A.K.: So, we got to figuring, maybe they're looking for the exact thing that we are, so then the Roci and I, we...

A.B.: We?

Casa bajo la nieve

Managing Editor: *Raimundo Senda*

Production Editor: *M. Wallace*

Editorial Assistant: *Eduardo Warner*

Art Editor: "*Rocío* ♡"

Music by: "*BlancaNieves*"

Printed in Spain

Contents

Introduction

This chapter provides an introduction to the subject, covering the basic theoretical principles used in the design and analysis of communication systems. The chapter begins by defining what a communication system is, listing the basic functional elements, and briefly describing the main function of each of them. Then, some general classifications of communication systems into different types are made, and various aspects related to the design and analysis of a communication system are discussed, to conclude by presenting the objectives of the subject and its contents.

Definition of a communication system

The purpose of a communications system is the *transmission of information* between two points separated by a distance but physically linked by some physical structure (natural or artificial) that can be used for it.

Transmission can therefore be defined as the process of sending or transporting information from one point (source) to another point (destination) through a channel or transmission medium, as illustrated in Figure 1 .

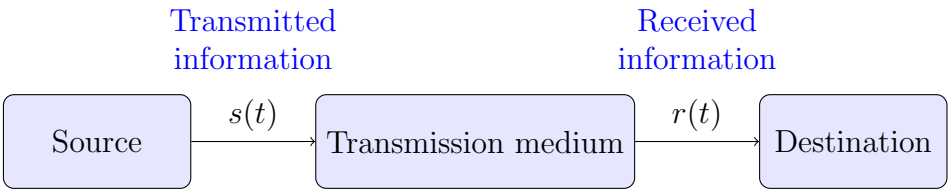


Figure 1: General representation of a commuunication system.

The transmission medium can be any element that allows information to be sent between a source and a destination: a pair cable, a waveguide, a coaxial cable, an optical fiber, or the atmosphere itself (using the entire radio spectrum) are some common examples. However, some storage devices, such as CD-ROMs, DVDs, etc., that allow information to be transported between two points, can also be considered transmission media.

In general, the physical manifestation of the information is not adequate to transmit it directly through the transmission media. Typically, it is necessary to process the physical manifestation of the source in order to efficiently introduce the information that it contains into the medium or channel. Most transmission media are designed to transmit electrical or electromagnetic signals, while the physical manifestation of the sources is in many cases not electrical in nature. To take a simple example, in an audio source (voice, music, etc.), the physical manifestation of the

information consists of acoustic pressure waves traveling through the air. In order to be able to transmit this type of information, it is first necessary to convert the physical manifestation of the information into an electrical signal on the source side. This is done using a *transducer*. The classic example of a transducer for audio signals is a microphone, which converts acoustic pressure waves into an electrical signal representing the variation of sound pressure over time. Figure 2 shows an example of an electrical signal associated with a voice signal.

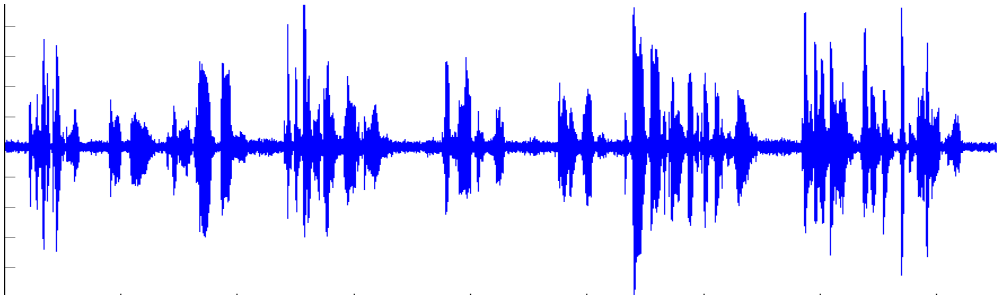


Figure 2: Example of an electrical signal associated to a voice signal.

Once the physical manifestation of the information has been converted into an electrical signal suitable to represent it, this electrical signal will be processed by the communications system, and sent over a particular transmission medium. At the receiving end it is further processed to convert the received electrical signal into the corresponding physical manifestation of the information.

Basic functional elements in a communication system

The previous section defined a communication system and briefly described the basic steps that must be followed to transmit information between two points. In this section, these elementary steps will be specified in the definition of the basic functional elements that are part of a communication system.

A communications system is made up of many elements, but from a functional point of view, the basic elements are the five shown in Figure 3

- Source of information
- Transmitter
- Channel
- Receiver
- Destination of the information

Each of these functional elements is briefly described below.

Information Source

As its name implies, the information source generates the information to be transmitted to the other end of the communication system. The message to be transmitted is the physical manifes-

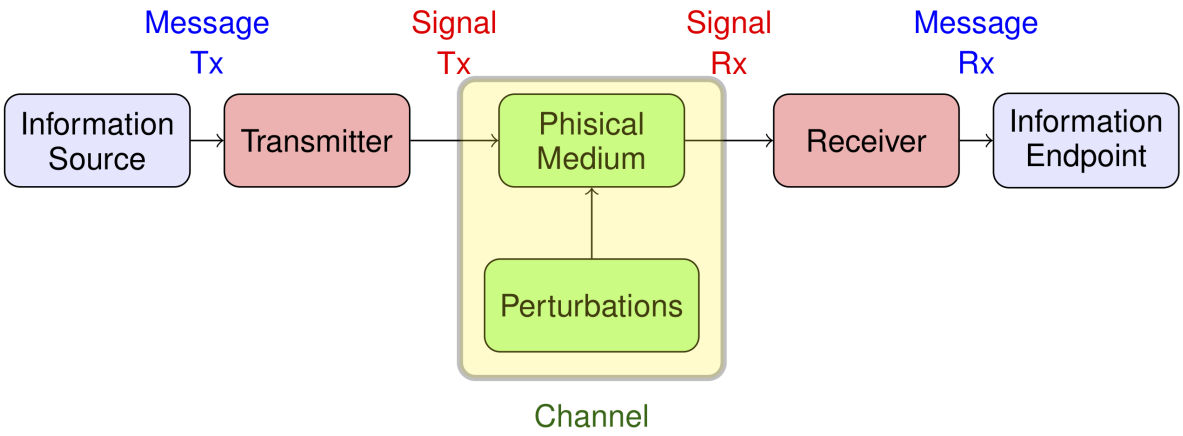


Figure 3: Functional block diagram of a communication system.

tation of the information produced by the source. As mentioned earlier, regardless of the type of source, it is common for a transducer to convert the information into an electrical signal that represents it (e.g., a microphone converts an acoustic signal into an electrical signal, or a video camera converts a sequence of images into another electrical signal).

There are many kinds of information sources. If they are classified according to the format in which the information is represented, they can be classified as analog or digital sources. An analog source produces information that manifests itself in continuous variations of some kind of physical quantity. Some examples would be the pressure in the air when you speak or the continuous change in temperature of a thermometer over time. With this type of source, the message to be transmitted is represented by a continuous waveform that represents the variation of the corresponding physical quantity, as in the example of the voice signal in Figure 2. In a digital source, on the other hand, the information is represented by a set of symbols belonging to a finite alphabet, sent sequentially at discrete time intervals (a symbol is sent every T seconds). Examples of this type of source would be text files, where the alphabet consists of the possible characters that can appear in the text, or binary data systems, where the information is encoded as a sequence of ones and zeros (binary alphabet, $\{0, 1\}$ or bits).

Transmitter

The function of the transmitter is to convert the information signal or message, regardless of its format, into an electrical or electromagnetic signal suitable for transmission over the physical medium used by the system to carry out the communication. This medium is usually referred to generically as the communication channel. In order to perform this task efficiently, relative knowledge of the channel is required: for example, it is necessary to know how much the medium will attenuate the transmitted signal in order to amplify it sufficiently, or in which frequency range the medium will allow the transmission to be performed with the least distortion. The transmitter uses this relative knowledge of the channel to generate a signal that matches the characteristics of the channel so that the signal suffers as little distortion as possible during transmission.

The process by which the transmitter adapts the signal to the characteristics of the channel is generically called *modulation*. The process to be carried out depends on several factors, such as the nature of the information source, the specific characteristics of the channel, the required features

or the available resources, in particular the available energy at the transmitter and the available bandwidth on the channel, but there are some general aspects that can be discussed at this time. Regarding the frequency range of the signal to be transmitted, two different operating strategies can be adopted: baseband transmission or bandpass transmission. In a baseband transmission, the signal is transmitted in the same frequency band that it naturally occupies, which is usually low frequencies. In a bandpass transmission, the frequency range occupied by the signal spectrum is modified. The signal is processed so that its frequency response is transferred to another frequency band centered around a certain center frequency (ω_c rad/s or f_c Hz), which may be more appropriate for transmission for various reasons.

Channel

The physical medium used to transmit the signal from the transmitter to the receiver is generally referred to as a channel. There are several types of media that can be used to transport signals in a communications system. The most common are cables of various types, such as pair cables used in traditional telephone systems or coaxial cables used in television systems, waveguides, fiber optics, and the radio spectrum, where the medium is the atmosphere itself and the electromagnetic energy carrying the information is introduced or extracted from it through the use of antennas. An important aspect to take into account in this last transmission medium is that the medium is unique and therefore shared among all its potential users. This means that, in practice, it is a scarce resource and that access to it is regulated by public administrations; otherwise, the interferences between the different users trying to use it in an unregulated way would make its efficient use impossible.

Each physical medium has its own characteristics, but regardless of the physical medium, a communication channel will always introduce a series of *perturbations* or *distortions* on the signals during their transmission. In the best case, in what could be considered an ideal scenario, the channel will introduce two effects into the transmitted signals: a delay and an attenuation; both effects are inherent to the transmission of electrical or electromagnetic signals over a medium:

- The signals suffer an attenuation when they propagate through any medium.
- It takes a certain time for signals to travel a certain distance through any medium.

If only these two effects appear during transmission, the received signal $r(t)$ can be written in terms of the transmitted signal $s(t)$ as

$$r(t) = C\, s(t - t_0),$$

where the constant $C < 1$ defines the attenuation and t_0 the delay. These two effects are inherent to the transmission of electromagnetic signals and therefore unavoidable, but in most cases they are not problematic. The attenuation can be compensated by amplification with a gain factor of $G = 1/C$. As for delay, if it is within reasonable limits, it usually does not have a significant effect, and given the speed of transmission of electromagnetic signals in most applications, it is not a problem in practice. Therefore, a medium that includes only delay and attenuation could be considered a system with ideal response.

However, in addition to these two effects, other types of undesirable effects will appear in practice, mainly linear distortion, non-linear distortion and noise, especially thermal noise, which

is always present in the transmission of electrical or electromagnetic signals due to the inherent thermal motion of charge carriers (electrons, photons,...).

In this subject, due to its introductory nature, only linear distortions and thermal noise will be considered, and nonlinear distortions will not be taken into account. In this case, the model to be used for the linear distortion is a time-invariant model characterized by an impulse response in the time domain, $h(t)$, and the corresponding representation of this response in the frequency domain, $H(j\omega)$. The relationship between the response of the system in the time and frequency domains is given by the Fourier transform

$$H(j\omega) = \mathcal{FT}\{h(t)\} = \int_{-\infty}^{\infty} h(t)\, e^{-j\omega t}\, dt$$

$$h(t) = \mathcal{FT}^{-1}\{H(j\omega)\} = \frac{1}{2\pi}\int_{-\infty}^{\infty} H(j\omega)\, e^{+j\omega t}\, d\omega.$$

Considering the linear distortion and noise, the communication channel model will be the so-called linear channel model, which is given by the relationship

$$r(t) = \int_{-\infty}^{\infty} s(\tau)\, h(t-\tau)\, d\tau + n(t) = s(t) * h(t) + n(t),$$

and that is shown conceptually in the block diagram of Figure 4: the channel output is the result of a linear distortion given by $h(t)/H(j\omega)$ and the sum of the noise $n(t)$.

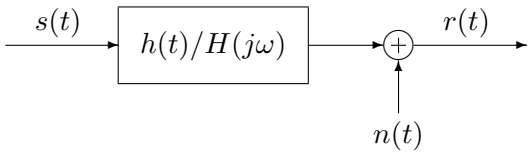


Figure 4: Usual channel model.

Note that the effect of delay and attenuation can be included in the impulse response of the system. In particular, an ideal system without linear distortion of the signal and without noise, which produces on the transmitted signal only an attenuation given by a factor $C < 1$ and a delay t_0, so that the received signal

$$r(t) = C\, s(t - t_0),$$

can be modeled by using a linear system with response

$$h(t) = C\, \delta(t - t_0) \leftrightarrow H(j\omega) = C\, e^{-j\omega t_0}.$$

Receiver

The receiver must recover the original information signal (message) from the received signal. Since this signal has undergone some distortion during transmission (in the channel model above, linear distortion plus the effect of noise), it may not always be possible to accurately recover the transmitted information signal. The receiver must be designed to recover the information signal with the highest possible fidelity, given the distortions that the signal has undergone during transmission. The way this fidelity is measured depends on the type of information signal being transmitted. If the information being transmitted is in analog format, the goal is for the waveform

of the received signal to resemble that of the transmitted signal as closely as possible, and the usual way to quantify this resemblance is by what is called the signal-to-noise ratio (SNR or S/N). This ratio measures the ratio among the energy or power of the transmitted signal and the energy or power of the difference between the transmitted and received signals (which is considered noise). If the format of the information transmitted is digital, the objective is to receive the least number of erroneous symbols that represent the information, so that the fidelity is quantified with an error probability, of symbols or of bits in binary systems, BER of *Bit Error Rate*. In any case, the receiver must generally perform the following tasks:

- Demodulate the signal, which means undoing all the transformations that were made in the transmitter to condition the signal for its transmission through the communications channel, such as returning the signal to its original frequency band if a bandpass transmission.
- Minimize the effect of noise on the information signal.
- Compensate, if possible, the linear distortions introduced by the channel.

As with the transmitter, the way these three tasks are performed depends on many factors, such as the format of the information being transmitted or the bandwidth being used. Throughout the course, we will see how these functions materialize for the different types of modulation.

Analog and digital communications systems

Similar to how information sources have been classified according to the format of the information they produce, leading to the distinction between analog and digital sources, communication systems can also be classified according to the format in which they transport information into two broad types: analog systems and digital systems. An analog communication system sends information encoded in a particular continuous waveform, while a digital system sends information encoded in a sequence of symbols sent sequentially at a particular rate (R_s symbols per second). The most common case is binary systems, where the information is contained in a sequence of bits (ones and zeros) sent at a certain bit rate (R_b bits per second).

It is important to note that an analog/digital communication system is not limited to transmission from a source of the same type (analog/digital). In particular, an analog signal can be digitized, converted to a sequence of bits at a given rate, and then converted back to analog (analog-to-digital, or A/D, and digital-to-analog, or D/A, conversion processes). The analog-to-digital conversion of a signal consists of two steps. An analog signal is continuous both in time and in the range of possible amplitudes (there is continuity in the two axes of the time representation of the signal, as shown in Figure 5). To convert it to digital format, the signal is discretized both in time and in amplitude, as also shown in the figure:

- From continuous time to samples of the signal in discrete time through periodic sampling

$$s[n] = s(t)\big|_{t=nT_m} = s(nT_m),$$

  where T_m is the sampling interval. If the signal $s(t)$ is band-limited, with a bandwidth of B Hz, the well-known Nyquist theorem for sampling shows that there is no loss of information in the sampling process; this means that it is possible to recover the original signal from

its samples by interpolation with sinc-functions (which is equivalent to interpolation with a low-pass filter of the bandwidth of the signal). To do this, it is only necessary to ensure that the sampling interval T_m is small enough to guarantee that its inverse, the sampling frequency, is at least twice the bandwidth of the signal expressed in Hz.

$$\frac{1}{T_m} = f_m \text{ samples/s} \geq B \text{ Hz}.$$

- After sampling, digitization is completed by quantizing each of the resulting samples with n bits. The figure shows an example of a 3-bit quantization, in which the dynamic range of the signal is divided into 8 possible values (since 3 bits allow the representation of $2^3 = 8$ values). In this step, a distortion of the information signal occurs because the original values of the signal in each sample (yellow circles in the figure) are passed to the closest quantized value (blue circles in the figure). This effect is known as quantization noise.

The process of converting an analog signal to digital therefore produces a sequence of bits. The rate of said sequence will be given by the product between the sampling rate and the number of bits per sample

$$R_b \text{ bits/s} = f_m \text{ (samples/s)} \times n \text{ (bits/sample)}.$$

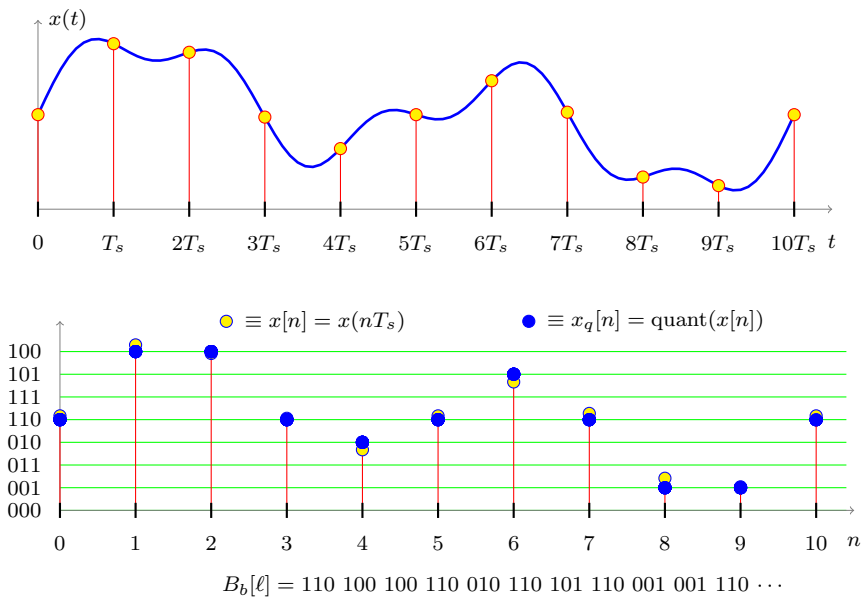


Figure 5: Analog-to-digital (A/D) conversion: from a signal to a bit sequence.

Although there is always some distortion of the information signal due to the quantization noise, this distortion can be negligible in many cases. All you have to do is use a sufficient number of bits per sample. If the number of bits is increased, the number of quantization levels will be 2^n, where the difference between levels within the range of the signal will progressively decrease as n increases, arriving at a moment when the distortion due to quantization will be practically negligible when reconstructing the signal in continuous time by interpolating with sincs (low-pass filtering). The cost of reducing distortion by increasing the number of bits per sample is an increase in the bit rate of the bit sequence resulting from the A/D conversion.

The current trend is for analog sources not to be transmitted using an analog communication system, but to be converted to digital format, transmitted using a digital communication system,

and finally converted back to analog format. This is because, although each type of system has its advantages and disadvantages over the other, the advantages of digital systems generally outweigh their disadvantages for most applications. This means that digital communication systems currently have a clear advantage over analog communication systems. The advantages of digital systems over analog systems that have made them clearly dominant will be discussed in detail in Chapter 3, which is devoted to the study of digital communication systems.

Although it is also possible to transmit digital information using analog communication systems, this option is much less common, and in practice is limited to a few systems with very specific requirements.

Design of a communications system

There are many factors to consider when designing a communications system. Some of the most important are:

- Quality or services that are required.
- Use of resources.
- Economic cost.
- Technology that is available for the implementation.

Each of these factors will be briefly discussed below.

Required quality

Among the specifications of a communication system, one of the most important is the quality required. The design of the system must be aimed at achieving the specified quality within the constraints imposed by the available resources and at the lowest possible cost, without in any case exceeding the maximum cost that can be assumed for the system. Of course, there is always a compromise between these factors; the greater the resources available, the higher the quality that can be achieved.

The way to specify the quality of a communication system is different depending on the type of system. For analog communications systems, quality is tied to the fidelity of the received signal: the received signal should be as close as possible to the originally transmitted signal. As discussed earlier, the transmission process introduces distortion and interference, primarily thermal noise. This assumes that the received signal will be different from the original signal. To measure how faithful a signal is, the signal-to-noise (S/N) ratio is usually used as a figure of merit, which gives us the relationship between the power of the signal and the noise.

In some cases, when using this measure, the combined effect of all the distortions that occur on the transmitted signal is considered to be noise, i.e., the difference between the transmitted signal and the received signal. In fact, it includes the effect of noise and other linear or nonlinear distortions that occur. For this reason, it is also referred to as the signal to noise ratio. Obviously,

the higher the value of this signal-to-noise (or distortion) ratio, the higher the quality of the system.

On the other hand, in digital communication systems, since the information is contained in a sequence of symbols of a finite alphabet (usually binary, ones and zeros), the figure of merit that is usually used to quantify the quality of the system will be the *error rate (or error probability)* of the symbols or bits (in this case it is commonly known by the acronym BER, from *Bit Error Rate* or *Bit Error Ratio*). Obviously, the lower this error rate or probability, the higher the quality of the digital system.

Available Resources

The resources available to the communications system are critical in limiting the maximum achievable performance. Of the resources to consider, bandwidth and power are usually the most important.

Bandwidth will, in practice, be limited in most applications. Either due to physical limitations, given the limited bandwidth of some of the common transmission media, such as cables, or due to administrative limitations, basically when transmitting over the radioelectric spectrum, since the use of this shared medium is completely regulated.

The bandwidth is related to the quality of the signal. In analog systems, more precise applications require more bandwidth. There are two main reasons for this. In some cases, the actual bandwidth of analog signals may be greater than the bandwidth available for transmission, so signals must be filtered to reduce their bandwidth before transmission. In this situation, the more frequency components that are eliminated (as the bandwidth is reduced), the greater the distortion of the information signal. In other cases, one of the ways to reduce the effect of noise is to spread the spectrum of the signal so that there is some redundancy in the transmitted signal. The more the bandwidth of the signal is increased, the greater the protection against noise that is obtained during transmission. This will be seen in more detail when we analyze the benefits of what are known as angle modulations, or phase and frequency modulations. In digital systems, bandwidth refers to the maximum transmission rate that can be obtained with minimum characteristics, so the greater the bandwidth, the greater the transmission capacity (transmission at a higher speed).

Cost of existing system and technologies

The cost of the system is another factor to consider when designing it, since there is usually a target cost or a maximum cost that cannot be exceeded. This cost is related to the technologies used in the implementation of the transmitters and receivers and in the choice of the transmission medium to be used. Therefore, when deciding on some design options, knowledge of the possible technologies to be used and their corresponding costs will be an important aspect.

Despite this importance, this subject will not deal with this aspect, as it is outside the objectives of the subject, which, as we will see in the next section, is more theoretical than practical and will focus on the basic theoretical aspects that govern the operation of a communication system.

Objectives and organization of the course

After a brief introduction to communication systems, this final section of the chapter presents the basic objectives of the subject and the organization of its content to achieve those objectives.

Course objectives

This subject attempts to establish the fundamental theoretical principles that are applied in the design and analysis of communication systems, both analog and digital. For this purpose, the mathematical characterization of a communication system and of the signals present in it, both transmitted signals and interfering signals such as noise, will be essential. This characterization will allow the analysis of a communication system and the derivation of the theoretical principles that define the optimal design of each of the various functional elements that make it up. In particular, the following fundamental objectives can be established

- To introduce the statistical characterization of signals related to communication systems, information signals and, in particular, thermal noise, which is always present in the transmission of an electromagnetic signal.
- To introduce the concept of modulation in analog communication systems and to study the most common types of modulation: amplitude modulation and angle (phase and frequency) modulation.
- To form the core knowledge base for digital communications, presenting in a simplified way the concept of digital modulation, transmission over Gaussian channels, where the main element of distortion is thermal noise, and studying the basic theoretical principles to design a digital demodulator, applying the statistical principles of decision theory and the vectorial representation of signals. Finally, information theory principles will be used to obtain some of the fundamental limits that can be reached in a digital communication system.

Course organization

In order to meet the objectives presented in the previous section, after this brief introduction the subject has been organized into the following chapters:

1. Noise in communication systems
   This chapter presents the statistical characterization of the signals in a communications system, and in particular the thermal noise signal, which is always present as a distortion element in the transmission of electrical or electromagnetic signals. To carry out this characterization, some basic concepts of random variables and random processes are reviewed in order to present the statistical model commonly used to model thermal noise. This model is used to calculate the signal-to-noise ratio in the transmission of an information signal.
2. Analog modulations
   This chapter introduces various modulation techniques used in analog communication systems. In particular, the most common types of amplitude modulation and phase and frequency angle modulation are introduced. For each type of modulation, its description in the

time domain, its spectral characteristics, and its power requirements are presented. Finally, the behavior of each modulation against noise is analyzed, evaluating the signal-to-noise ratio obtained with each of them.

3. Modulation and detection in Gaussian channels
   This chapter presents the basic principles that govern the design and analysis of a digital communications system. It introduces the concept of digital modulation as a mechanism for transmitting digital information over analog channels, and the concept of detection as a mechanism for recovering the transmitted digital information from the analog signal received over the communication channel. The simplest model of this channel is used: a Gaussian additive channel.

4. Fundamental limits in digital communications systems
   Finally, the last chapter presents some of the fundamental limits that can be reached with a digital communication system. In particular, it examines how to calculate the maximum amount of information that can be reliably transmitted by a digital communication system, known as the channel capacity. Obtaining this limit is based on the application of information theory and the use of quantitative information measures, which are presented in this chapter as tools for the analysis of digital communication systems.

Recommended bibliography

There are excellent books on communication systems. For this course, only a small number of them are recommended, which adequately cover all the contents of the course, in order to avoid an excessive number of references for the student. Also with this objective, a distinction has been made between what is called the basic bibliography, where there are two references that cover all the contents of the subject, and what is called the supplementary bibliography, where texts are cited that allow going deeper into some of the contents of the subject beyond its own objectives. These bibliographical references are presented below, with a brief commentary on each one.

Basic bibliography

1. A. Artés et al. “Comunicaciones Digitales”, Pearson Education, 2007
   Althought it is written in Spanish, the first basic reference is this excellent book, written by several university professors, oriented towards its use as a learning manual, which is why it is very appropriate as a reference for the subject. This book fundamentally deals with digital communications systems with a clear focus on the contents usually covered in degrees related to Telecommunications Engineering. Chapter 3 covers most of the contents of chapter 1 of the course. Chapter 4 and Chapter 9 cover, respectively, all the contents of chapters 3 and 4 of the subject.

   This book is available online through the website of its first author, Professor Antonio Artés Rodríguez, from the Carlos III University of Madrid.

   - Available online: www.tsc.uc3m.es/∼antonio/libro_comunicaciones

2. J.G. Proakis and M. Salehi. “Communication Systems Engineering” (2nd Ed.), Prentice-Hall, 1994
   This is another excellent text on communications systems, both digital and analog. Its compact notation and its modularity facilitate the task of monitoring the contents of the

subject despite its different sequencing with respect to the one followed in the subject. Chapters 4 covers the contents of chapter 1 of the subject; Chapters 3 and Chapter 5 cover the contents of chapter 2 of the subject; and finally Chapter 6 and Chapter 9 do it with those of chapter 4. Although Chapter 7 covers many of the contents of chapter 3 of the subject, in this case the approach is slightly different from that followed in the subject.

Suplementary bibliography

1. A. Papoulis. "Probability, random variables, and stochastic processes", (3rd Ed.), McGraw-Hill, 1991
   One of the reference books on the foundations of probability theory and stochastic processes. An excellent reference for all the statistical concepts that are handled in the subject.

2. A.B. Carlson. "Communication Systems" (2nd Ed.), McGraw-Hill, 1986
   Classic introductory text to analog and digital transmission. Basically, it consists of three parts: the first introduces the basics of random signals and processes; the second covers analog communications, while the third part is devoted to digital communications. The development of analog communications is simple and intuitive and is illustrated with block diagrams of systems and basic electrical circuits.

3. S. Haykin. "An Introduction to Analog and Digital Communications", Willey, 1989
   Another classic text that deals with analog and digital communication systems, although in this case with a clear emphasis on the latter. It is an interesting book for the introductory treatment of communication theory and the mathematical nature of its formulation.

4. B. Sklar. "Digital communications : fundamentals and applications", Prentice Hall, 2001
   Advanced book on digital communications, interesting for those students with a basic training in probability theory. It presents an excellent introduction to signal fundamentals, signal spectrum, and baseband transmission. From this introduction, the book presents multiple variants of modulations and modulation techniques that, although they go beyond the objectives of this subject, may be of interest to a student who has taken it.

5. T.M. Cover and J.A. Thomas. "Elements of Information Theory". Wiley. 2006
   Classic book on Information Theory, surely the most frequent reference in this field. The most common quantitative measures of information are introduced, such as entropies or mutual information, and some applications of them are presented.

Chapter 1

Noise in the communication systems

In any communication system, there are several types of signals. Some signals are deterministic. Others, however, are random in nature, since their specific value at a given time is not known a priori, but only their statistical parameters. Examples of this type of signal are:

- The signals that carry the information to be transmitted. Every information signal has a certain degree of uncertainty. If this were not the case, it would not really contain information (e.g., if the receiver knew precisely the signal to be transmitted, it would not be necessary to transmit it).
- The thermal noise that always appears in the transmission of electromagnetic signals.

Random processes are used to characterize signals of this type, where some of their statistical properties are known, but not their specific values at a given time.

On the other hand, the information to be transmitted is, by its nature, also modeled by random processes. This is because any information signal must have some degree of uncertainty. If it does not, it contains no information.

For this reason, this chapter will use the theory of random processes to characterize communication signals and, in particular, thermal noise. Modeling this noise will be fundamental to the design and analysis of communications systems, since thermal noise is one of the main sources of distortion that occurs in any communications system. We will also analyze how linear systems affect the statistical parameters that define a process, and finally we will study the signal-to-noise ratio under different circumstances in a communications system.

Before turning to the use of random processes as a tool for characterizing signals in a communications system, we will briefly review some concepts related to probability, random variables, and random processes. Although these concepts would be part of the previous topic of *Statistics*, they are included here for completeness.

1.1 Probability

In this section, we will briefly review some of the basic concepts of probability theory. We will focus on those aspects that are necessary for the treatment of random processes in the field of

communications signal modeling.

Probability theory deals with mass phenomena. There are countless examples: games of chance, the motion of electrons, birth and death rates, etc. Probability theory tries to establish averages for such phenomena. In particular, its purpose is to describe and predict these averages in terms of probabilities of events.

1.1.1 Probability space

Before we can define what a probability space is, it is necessary to make several definitions.

Random experiment

The fundamental concept upon which probability theory is based is the *random experiment*. A random experiment is one whose outcome cannot be accurately predicted. Flipping a coin, rolling a die, drawing a card from a deck, or measuring the voltage across a pair of copper wires are some examples of random experiments.

Sample spaces

Every random experiment has certain output values, or possible outcomes of the experiment. In the case of a coin toss, that the figure facing up is heads or tails, in the case of a die, that the number of points on the face facing up is 1, 2, 3, 4, 5 or 6. The sample space is defined as the set of all possible outputs of an experiment. It is usually denoted by the Greek letter omega, Ω.

In nature, there are two types of sample spaces:

- Discrete, when the experiment has as possible outcomes a finite number of possible values, or a contable infinite number of values.
- Non-discrete (or continuous), when the sample space corresponds to continuous sets of possible values (or in other words, the number of possible output values is uncountable infinite).

Examples of the former are the die or coin mentioned above. In that case the sample space is for the coin heads and tails, in the case of the die, 1, 2, 3, 4, 5 and 6. An example of a random variable with a continuous sample space is the value of the voltage across a resistor, which can take any value within a range of voltage values. In this case the sample space is the entire continuous set of possible values.

There are also mixed spaces, with part discrete sample space and part continuous, although they will not be discussed in this chapter.

Events

An event is a subset of the sample space over which it is possible to define a probability. For this probability measure to make sense, it must satisfy a number of conditions, which will be discussed a little later. First, let us see what is meant by a probability.

The probability of an event E is a number, $P(E)$, non-negative, defined between 0 and 1 ($0 \leq P(E) \leq 1$), assigned to this event and describing how probable or improbable this event is. This number can be interpreted as

If a given experiment is performed a number N of times (assuming N is sufficiently large), and event A occurs N_A times, then we can say that the probability will be fairly close to the ratio N_A/N:

$$P(A) \approx \frac{N_A}{N}$$

This may be an intuitive definition of probability, i.e., a measure that tells us how often an event occurs when a given experiment is performed.

In the case of discrete spaces, the idea is simple. What is the probability that a die will roll a 5? If the die is not tricked, this probability is 1/6. But in the case of continuous spaces, there is an important nuance to keep in mind. For example, what is the probability that the voltage across a resistor is 1 volt? The answer is 0. Although this may seem counterintuitive, the explanation is that the set of values it can take is infinite, so the probability of having any of them is zero. In short, it is not possible to define a probability for a particular value. What is possible is to define the probability that the voltage value is in a certain interval, say between 0.99 and 1.01 volts. This event has a probability associated with it.

Thus, in experiments with discrete sample spaces, the events must consist of a subset of the sample space, including single-element events. In the case of continuous sample spaces, each event must have a probability, so an event must be a "region" of the sample space (not a single value). The sigma field, denoted by $\mathcal{B}$, is usually defined as the collection of subsets of Ω, that is, the set of all possible events.

Some definitions of events that may be useful are the following:

- Trivial event: it is the event that occurs in every experiment, i.e., its probability is 1.
- Null set ($\varnothing$): The one that does not have any element.
- Event union ($E_1 \cup E_2$): it is the event that occurs when E_1, E_2 or both occur.
- Event intersection ($E_1 \cap E_2$): the event that occurs when events E_1 and E_2 occur at the same time.
- Exclusive or disjunct events: those for which $E_1 \cap E_2 = \varnothing$. For them it is satisfied that $P(E_1 \cup E_2) = P(E_1) + P(E_2)$.
- Complement of an event (E^c): the sample space minus the event itself, i.e., the one that satisfies that

$$E \cup E^c = \Omega, \quad E \cap E^c = \varnothing.$$

Probability space

The probability space is defined as the triplet (Ω,$\mathcal{B}$,P); that is, the sample space, the space with the different events and the probability measure that indicates the probability of each event. Some of the properties that the probability measure on events must satisfy are the following:

1. $P(E^c) = 1 - P(E)$.
2. $P(\varnothing) = 0$.
3. $P(E_1 \cup E_2) = P(E_1) + P(E_2) - P(E_1 \cap E_2)$.
4. If $E_1 \subset E_2$ then $P(E_1) \leq P(E_2)$.

1.1.2 Conditional probability

Suppose there are two events, E_1 and E_2, defined on the same probability space with corresponding probabilities $P(E_1)$ and $P(E_2)$. These probabilities are sometimes referred to as the a priori probabilities of each event. If one of the events is known to have occurred, say E_2, this can give us some information about the other event, which changes its a priori probability (without knowing that either event has occurred). This new probability is called the conditional probability. The conditional probability of the event E_1 given the event E_2, denoted as $P(E_1|E_2)$ is defined as:

$$P(E_1|E_2) = \begin{cases} \dfrac{P(E_1 \cap E_2)}{P(E_2)}, & P(E_2) \neq 0 \\ 0, & P(E_2) = 0 \end{cases}.$$

Example

A fair (unloaded) die is rolled, and the following events are defined

- E_1: the outcome is greater than 3
- E_2: the outcome is an even number

The probabilities of each of these events are easily calculated by adding the probabilities of each of the possible initial values of the sample space that are part of the event.

$$P(E_1) = P(4) + P(5) + P(6) = \frac{1}{2}$$

$$P(E_2) = P(2) + P(4) + P(6) = \frac{1}{2}$$

$$P(E_1 \cap E_2) = P(4) + P(6) = \frac{1}{3}$$

The conditional probability $E_1|E_2$ is

$$P(E_1|E_2) = \frac{1/3}{1/2} = \frac{2}{3}$$

It is checked whether the result obtained is consistent with the probability of having a 4 or a 6 if the sample space is the event E_2. Knowing that the result is even changes the probabilities of having a result greater than 3, compared with the situation where there is no prior information.

Statistically independent events

An important statistical definition follows from the conditional probability. If it happens that $P(E_1|E_2) = P(E_1)$ this means that knowledge of E_2 does not provide information about E_1 and therefore does not change its probability with respect to the a priori probability (without the knowledge that E_2 has occurred). In this case, the two events are said to be *statistically independent*.

Formally, two events are said to be *statistically independent* when the conditional probabilities coincide with the a priori probabilities.

$$P(E_1|E_2) = P(E_1) \text{ and } P(E_2|E_1) = P(E_1).$$

Given the relationship between a priori probabilities and conditional probabilities, through the probability of intersection, the probability of the intersection of two statistically independent events is equal to the product of the probabilities of each event.

$$P(E_1 \cap E_2) = P(E_1) \times P(E_2).$$

Law of total probability

If the events E_i, with $i = 1, \ldots, N$ form a partition of the sample space Ω, which means that the following conditions are satisfied:

- $\cup_{i=1}^{N} E_i = \Omega$,
- $E_i \cap E_j = \varnothing$ for all $i \neq j$,

that is, that the union of the events forms the whole sample space, being the events disjoint among themselves, then, if for an event A the conditional probabilities $P(A|E_i)$ are available for all the events that form the partition, $i = 1, \ldots, N$, the probability of event A, $P(A)$, is obtained by means of the *Law of total probability*.

$$P(A) = \sum_{i=1}^{N} P(A|E_i)P(E_i).$$

Bayes rule

On the other hand, *Bayes' Rule* (although its idea is due to Bayes, it was finally formulated by Laplace) tells us that the conditional probabilities of the events of the partition given A, $P(E_i|A)$, are obtained by the following expression:

$$P(E_i|A) = \frac{P(A|E_i)P(E_i)}{P(A)} = \frac{P(A|E_i)P(E_i)}{\sum_{j=1}^{N} P(A|E_j)P(E_j)}.$$

1.2 Random variable

A *random variable* (r.v.) is a function that assigns a number to each of the possible outcomes of a random experiment, that is, to each of the elements of the sample space. In this section, we will focus on real random variables, for which the number assigned to each possible outcome of the random experiment belongs to the set of real numbers.

$$\Omega \to I\!R$$
$$\lambda \in \Omega \to X(\lambda) \in I\!R$$

Therefore, a (real) r.v. maps the results of a random experiment on the real line.

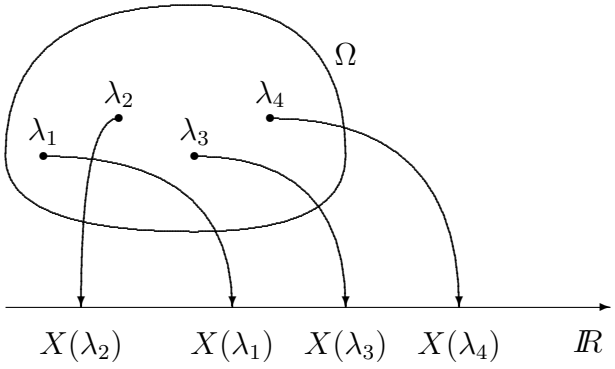


Figure 1.1: Random variable seen as a map from Ω to $I\!R$.

For example, in the die toss experiment, an assignment is already implicit (the number of dots on the up side). In other cases, such as flipping a coin, it is possible to assign one number to heads and one to tails (e.g. heads $\equiv 0$, tails $\equiv 1$). Random variables are usually denoted by uppercase X, Y, and the implicit dependence on the elements of the sample space of the random experiment, λ_i, is usually not expressed. Again, when classifying in terms of the type of values it can take, we will have mainly two categories of random variables:

- Discrete: finite set of values.
- Continuous: continuous range of values (in one or several intervals).

As for the values into which the output of the random experiment is translated, the set of real numbers that have an associated result of the sample space is called the *range* (or *domain*) of a r.v:

$$\mathcal{A}_X = \{x \in I\!R \ : \exists\, \lambda \in \Omega \text{ such that } \ X(\lambda) = x\}.$$

In the case of discrete random variables, it is also sometimes referred to as the "alphabet" of the random variable.

Probabilistically, a random variable is usually characterized by two functions (which are linked to each other):

- Distribution function, $F_X(x)$ (also known as cumulative distribution function).

- Probability density function, $f_X(x)$.

Each of these functions is described below.

1.2.1 Cumulative distribution function (CDF)

The cumulative distribution function (CDF) of a random variable is defined as

$$F_X(x) = P(X \leq x),$$

i.e., as the probability that the random variable X takes a value less than or equal to the argument x. The main properties of the distribution function are the following:

1. $0 \leq F_X(x) \leq 1$.
2. $x_1 < x_2 \rightarrow F_X(x_1) \leq F_X(x_2)$ ($F_X(x)$ is not decreasing).
3. $F_X(-\infty) = 0$ y $F_X(\infty) = 1$ ($\lim_{x\to-\infty} F_X(x) = 0$ and $\lim_{x\to\infty} F_X(x) = 1$).
4. $F_X(x^+) = F_X(x)$ ($F_X(x)$ is continuous on the left side).
5. $F_X(b) - F_X(a) = P(a < X \leq b)$.
   To calculate other probabilities including or not the extreme limits of the interval

$$P(a \leq X \leq b) = F_X(b) - F_X(a^-).$$
$$P(a < X < b) = F_X(b^-) - F_X(a).$$
$$P(a \leq X < b) = F_X(b^-) - F_X(a^-).$$

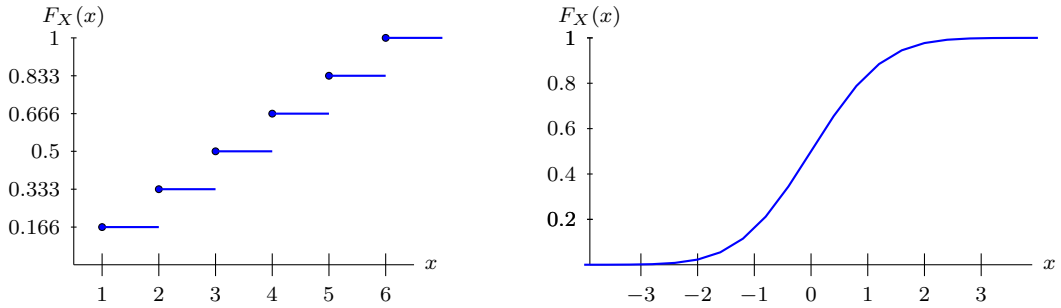
6. $P(X = a) = F_X(a) - F_X(a^-)$.
7. $P(X > x) = 1 - F_X(x)$.

In the above expressions, the following notation has been used as notation

$$F_X(x^\pm) = \lim_{\varepsilon\to 0} F_X(x \pm \varepsilon).$$

This distinction $F_X(x^\pm)$ is made to take into account the particular case of distribution functions for a discrete r.v., for which $F_X(x_i^-) \neq F_X(x_i)$, where $\{x_i\}_{i=1}^N$ is the discrete set of values that form the range of X. In general, for continuous random variables $F_X(x) = F_X(x^-)$, which implies that the probability of taking a particular value is zero, $P(X = a) = 0$. In any case, for both discrete and continuous variables, $F_X(x) = F_X(x^+)$ is satisfied (see property 4).

For discrete random variables $F_X(x)$ is a step function, with discontinuities at the discrete values that form the range of the random variable. For a continuous variable it has continuous variation. Figure 1.2 shows examples of discrete distribution function, in this case the experiment throwing a die, and continuous.



(a) Discrete (Die) (b) Continuous (Gaussian: $\mathcal{N}(0,1)$)

Figure 1.2: Examples of the cumulative distribution function for discrete and continuous random variables.

Frequency or probabilistic interpretation

To present an empirical, constructive interpretation of the distribution function, we can write:

$$F_X(x) = P(X \leq x) = \lim_{n\to\infty} \frac{n_x}{n},$$

where n is the number of realizations of the random experiment, and n_x is the number of outcomes for which $X \leq x$. Obviously, we can never do an infinite number of experiments, but we can make an estimate from a limited number of experiments. Figure 1.3 shows 100 realizations of a Gaussian random variable and the resulting estimate compared to the theoretical distribution function.

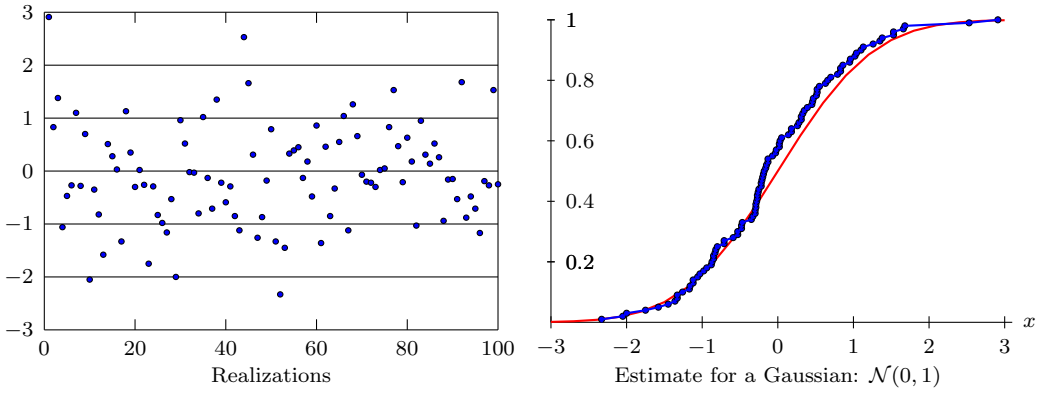


Figure 1.3: Estimation of the distribution function by means of its frequency interpretation.

1.2.2 Probability density function

The other function used to characterize a random variable is the probability density function (PDF), which is denoted as $f_X(x)$. The probability density function is defined as the derivative of the distribution function

$$f_X(x) = \frac{d}{dx} F_X(x).$$

This function indicates how the probability of the random variable is distributed. Its main properties are the following:

1. $f_X(x) \geq 0$.
2. $\int_{-\infty}^{\infty} f_X(x)dx = 1$.
3. $\int_{a^+}^{b^+} f_X(x)dx = P(a < X \leq b)$.
4. In general, $P(X \in A) = \int_A f_X(x)dx$.
5. $F_X(x) = \int_{-\infty}^{x^+} f_X(u)du$.

In the case of continuous variables it has a continuous variation, and in the case of discrete variables, the PDF includes pulses located at the discrete values that the variable can take (the derivative of a function with steps). The value at each of these discrete values corresponds to the probability that the random variable takes that value.

The nuance a^+ is used to treat discrete signals. In this case, the impulse is located at a, and integrating from a^+ does not include it. For continuous variables we can use a directly.

In the case of a discrete variable, its alphabet reduces to a set of finite values $\{x_i\}_{i=1}^N$. In this case, sometimes instead of working with the PDF, one works with the *probability mass function*, or sometimes the so-called *mass points*. In this case, the probability mass function or mass points is defined as the set of values $\{p_i\}_{i=1}^N$ such that

$$p_i = P(X = x_i),$$

which of course meet the following conditions

1. $p_i \geq 0$.
2. $\sum_{i=1}^{N} p_i = 1$.

The difference with the PDF is that it is usually represented as a function of i instead of with respect to x_i, but conceptually there is no difference between both representations.

On other occasions, for discrete random variables, once the sample space $\{x_i\}_{i=1}^N$ is known, the probabilities of each of the values in that space are denoted as $p_X(x_i)$.

In this course, we will generally work with the PDF, but when working with discrete random variables, we will frequently use the $p_X(x_i)$ notation instead of the $f_X(x)$ notation.

Frequency or probabilistic interpretation

To give an empirical interpretation of the PDF, we can define the probability density function as

$$f_X(x) = \lim_{\Delta x \to 0} \frac{P(x \leq X \leq x + \Delta_x)}{\Delta_x},$$

i.e.

$$f_X(x) = \frac{\text{Probability in an interval}}{\text{Length of the interval}} = \text{Probability Density},$$

when the length of the interval is taken to the infinitesimal limit. Using the frequency definition of probability,

$$f_X(x) = \lim_{\Delta_x \to 0} \left\{ \frac{1}{\Delta_x} \lim_{n \to \infty} \frac{n_x}{n} \right\},$$

where n is the number of realizations of the randomized experiment, and n_x is the number of outcomes for which $x \leq X < x + \Delta_x$.

This is equivalent to making a histogram, which consists of dividing the real line into intervals of width Δ_x and raising a vertical bar with the relative frequency of each interval. In this case, it can be seen that a histogram tends to the probability density function when the number of realizations increases and the interval length decreases. Figure 1.4 shows a histogram with a value $\Delta_x = 1.0$ made from 100 realizations.

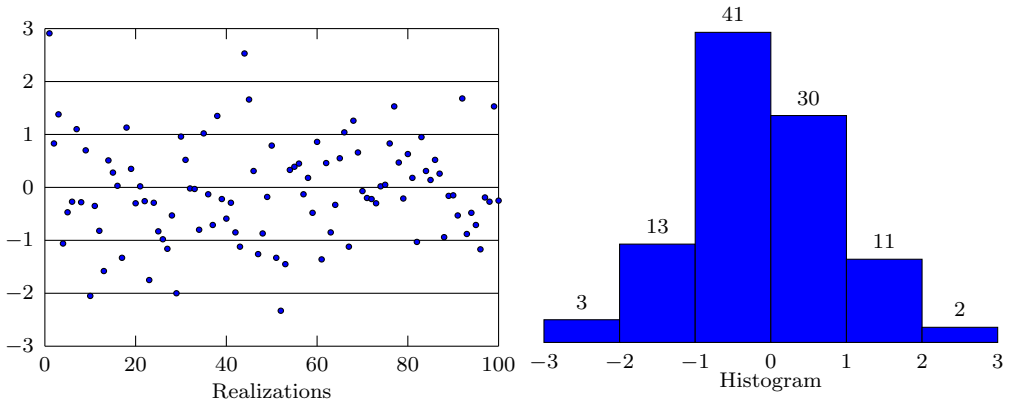


Figure 1.4: Approximation of the PDF by a histogram.

Figure 1.5 shows a histogram with a value $\Delta_x = 0.2$ made from 10000 realizations, and compares the normalized histogram with the theoretical probability density function.

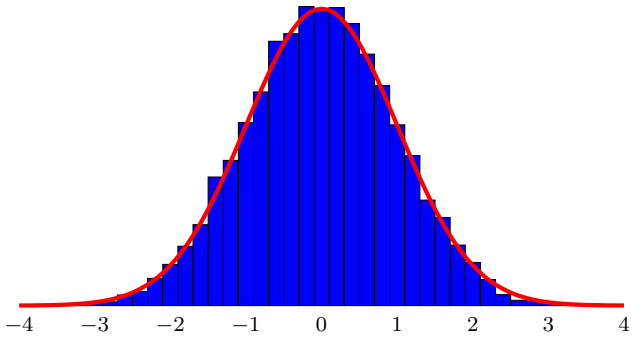


Figure 1.5: Approximation of the PDF by a histogram using 10000 realizations.

1.2.3 Random variables of interest

The following are the most common random variables used in communications.

Bernoulli

The Bernouilli random variable is a discrete random variable that takes two values, 1 and 0, with probabilities

- $P(1) = p$,
- $P(0) = 1 - p$,

respectively.

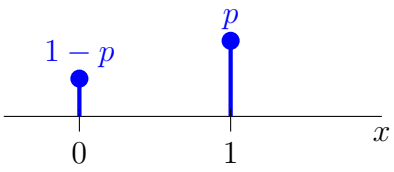


Figure 1.6: $f_X(x)$ for a Bernoulli random variable.

This is a distribution with one parameter, in this case p. Its probability density function is, obviously:

$$f_X(x) = \begin{cases} 1-p, & x = 0 \\ p, & x = 1 \\ 0, & \text{in other case} \end{cases} .$$

A Bernoulli random variable is a good model for, e.g.

- *Binary data generator.* In this case, it is normal that the parameter p is $1/2$, that is, that the 1's and 0's are equiprobable.
- *Error model.* Errors will occur in any transmission over a communications channel. An error can be modeled as the sum modulo-2 (XOR) of the input bit with a 1 (in the sequence modeling errors, a 1 is indicating an error, and a 0 is indicating a correct bit). Therefore, this type of variables can also be used to model errors. In this case, the parameter p is precisely the bit error rate.

Binomial

It is also a discrete random variable. This variable models the number of 1's in a sequence of n independent Bernoulli experiments, so it has two parameters, n and p. Its probability density function is as follows:

$$f_X(x) = \begin{cases} \binom{n}{x} p^x (1-p)^{n-x}, & 0 \leq x \leq n \text{ and } x \in \mathbb{Z} \\ 0, & \text{in other case} \end{cases} .$$

This variable can be used, for example, to model the *total number of bits received with error* when a sequence of n bits is transmitted over a channel with bit error rate p.

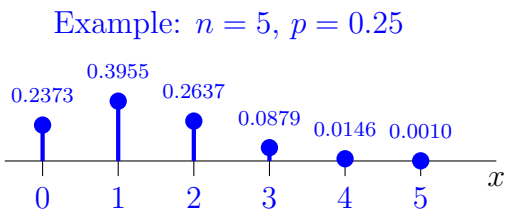


Figure 1.7: $f_X(x)$ for a binomial random variable.

Uniform

This is a continuous random variable of two parameters, a and b, which takes values in the interval (a,b) with the same probability for intervals of equal length. Its probability density function is

$$f_X(x) = \begin{cases} \frac{1}{b-a}, & a < x < b \\ 0, & \text{in other case} \end{cases} .$$

$$f_X(x) = \begin{cases} \frac{1}{b-a}, & a < x < b \\ 0, & \text{else} \end{cases}$$

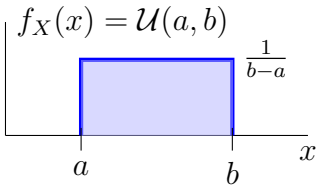


Figure 1.8: $f_X(x)$ for a uniform random variable.

Sometimes the notation $\mathcal{U}(a,b)$ is used to denote a uniform distribution between a and b. This model is used for continuous variables with known range for which nothing else is known. For example, to model a *random phase in a sinusoid*, a uniform r.v. between 0 and 2π is usually used.

Gaussian (Normal)

It is a continuous random variable with two parameters, μ and σ. Its probability density function is a Gaussian with mean μ and variance σ^2 (or what is the same, standard deviation σ),

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}.$$

It is sometimes denoted as $\mathcal{N}(\mu, \sigma^2)$. The Gaussian is the most important and undoubtedly the most widely used r.v. in communications. The main reason is that thermal noise, which is the major source of noise in communications systems, has a Gaussian distribution.

The distribution function, $F_X(x)$, for a Gaussian r.v. of zero mean and unit variance is commonly denoted as $\Phi(x)$.

$$\Phi(x) = P(X \leq x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}\, dz.$$

A function related to this distribution function, which is very often used, is the $Q(x)$ function, which is defined from the distribution function as

$$Q(x) = 1 - \Phi(x) = P(X > x)$$

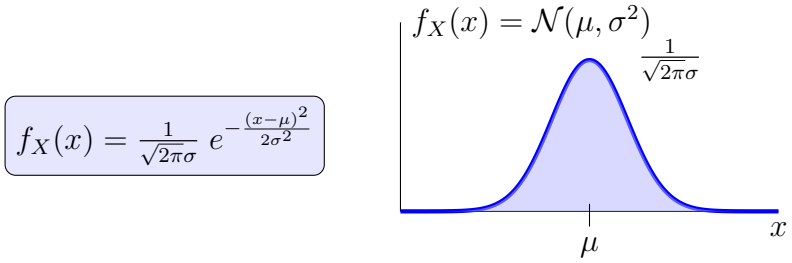


Figure 1.9: Probability density function of a Gaussian (Normal) random variable.

which gives the probability that the Gaussian random variable of zero mean and unit variance takes values greater than the argument of the function. This function, formally defined as

$$Q(x) = \int_x^{+\infty} f_X(z)\, dz = \int_x^{+\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}\, dz$$

has no analytical solution ($\Phi(x)$ does not either). However, it can be calculated numerically and is usually tabulated for its positive values, as shown in Table A.3, in Appendix A. Figure 1.10 shows the graphical interpretation of the value of this function as the area under the curve of the Gaussian distribution $\mathcal{N}(0,1)$ for positive and negative values of the argument x.

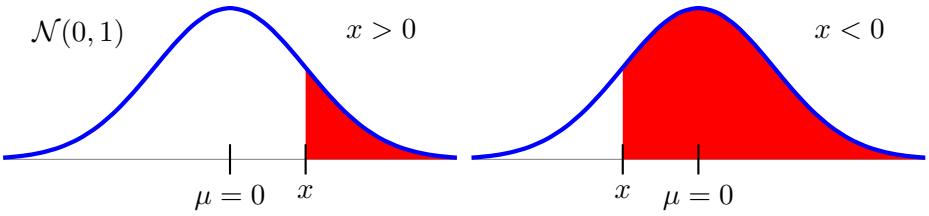


Figure 1.10: Graphical interpretation of $Q(x)$ for positive and negative arguments x.

From this figure it is easy to extract some of the properties properties of this function

1. $Q(0) = \frac{1}{2}$.
2. $Q(+\infty) = 0$.
3. $Q(-x) = 1 - Q(x)$.

Due to the symmetry $Q(-x) = 1 - Q(x)$, tables of this function are usually presented only for positive values of the argument of the function, since for negative values it can be obtained from that relation.

For a distribution with mean μ and variance σ^2, i.e. $\mathcal{N}(\mu, \sigma^2)$, a simple change of variable serves to estimate $P(X > x)$ via the $Q(x)$ function as

$$P(X > x) = Q\left(\frac{x - \mu}{\sigma}\right).$$

The function $Q(x)$ is of great interest in this course because it will be used, as we will see, to evaluate error probabilities in digital communications systems. Given its importance in the

subject, some examples of how this function can be used to obtain probabilities that a random variable with a given mean μ and variance σ^2 can take values in certain ranges will be illustrated below. From its definition, it is evident how probabilities that a random variable will take values greater than a certain threshold x can be calculated. Figure 1.11 shows some of the symmetries of the function $Q(x)$ that also allow to obtain probabilities that the Gaussian random variable takes values smaller than a certain argument (on the right side of the figures), from an equivalent problem on the very definition of the function $Q(X)$ (on the left side). And in Figure 1.12, an example of how to calculate the probability that the random variable takes values in an interval between two thresholds is illustrated. Such a problem can be solved by reformulating it as the difference between two problems with a single threshold.

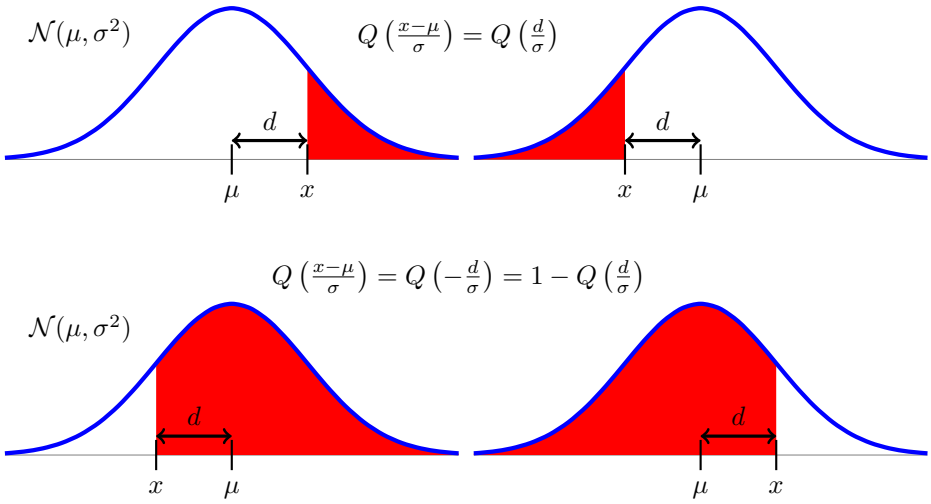


Figure 1.11: Symmetries of $Q(x)$.

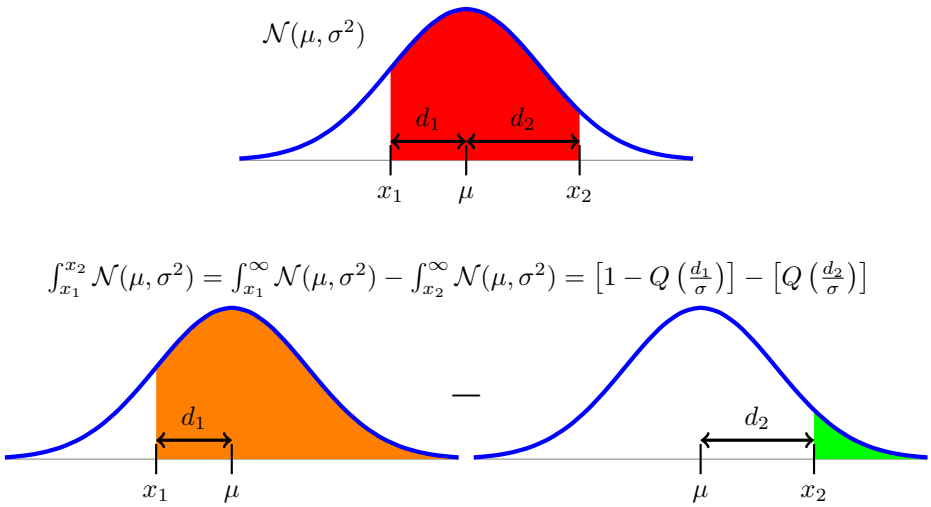


Figure 1.12: Computation of the probability of a Gaussian random variable taking values in a given interval using $Q(x)$.

1.2.4 Functions of a random variable

A function of a random variable $Y = g(X)$ is itself a random variable. To find its distribution function we can start from the definition of $F_Y(y)$:

$$F_Y(y) = P(Y \leq y) = P(g(X) \leq y)$$

This probability is

$$F_Y(y) = P(x \in B_X^g(y)),$$

where $B_X^g(y)$ is the set of values for X such that $g(x) < y$, i.e.

$$B_X^g(y) = \{x \in I\!R : g(x) \leq y\}.$$

Example

For the transformation $Y = -2X$, we want to calculate $F_Y(y)$.

In this case, it is straightforward to calculate $B_X^g(y)$,

$$B_X^g(y) = \{x \in I\!R : -2x \leq y\} = \{x \geq -y/2\},$$

and therefore

$$F_Y(y) = P(Y \leq y) = P(X \geq -y/2).$$

This probability for a certain random variable X can be calculated when $F_X(x)$ or $f_X(x)$ are known.

On the other hand, the probability density function of the random variable Y can be calculated, from $f_X(x)$ and the transformation $g(x)$, as follows

$$f_Y(y) = \sum_{i=1}^{N_r} \frac{f_X(x_i)}{|g'(x_i)|},$$

where N_r and $\{x_i\}_{i=1}^{N_r}$ are the number of solutions and the solutions themselves of the equation $y = g(x)$, respectively. The function $g'(x)$ is the derivative of $g(x)$. In order to obtain this expression it is necessary that the equation has a finite number of solutions, that for all of these solutions the derivative $g'(x_i)$ exists and that the derivative at the solutions is not zero.

Example

We have a Gaussian random variable X with zero mean and unit variance, i.e. $\mu = 0$ and $\sigma = 1$. We want to find the probability density function of the random variable

$$Y = aX + b.$$

In this case $g(x) = ax + b$, and the derivative is $g'(x) = a$. Equation $y = ax + b$ has a single solution

$$x_1 = \frac{y - b}{a}$$

Using this information

$$f_Y(y) = \frac{f_X\left(\frac{y-b}{a}\right)}{|a|} = \frac{1}{\sqrt{2\pi}|a|} e^{-\frac{(y-b)^2}{2a^2}}.$$

It can be seen that this distribution is a Gaussian distribution of mean b and variance a^2, i.e.

$$f_Y(y) = \mathcal{N}(b, a^2).$$

An important conclusion can be drawn from this example: *a linear function of a Gaussian random variable is also a Gaussian random variable.*

1.2.5 Statistic moments

We will now see how to compute some statistical moments associated with a random variable. Remember that a random variable is the result of a random experiment. If the PDF is known, it is possible to obtain some statistics about it, which is equivalent to saying the statistics of the random experiment.

Expected value (Mean)

The expected value (mathematical expectation) of a random variable is equivalent to its (arithmetic) mean, and is often denoted as m_X. The expected value measures the mean value obtained when the number of experiments is sufficiently large. This expected value is defined as

$$m_X = E[X] = \int_{-\infty}^{\infty} x\ f_X(x)\ dx.$$

Expected value of a function of X.

The expected value of a function of a random variable, $Y = g(X)$, is obtained as

$$E[g(X)] = \int_{-\infty}^{\infty} g(x)\ f_X(x)\ dx.$$

Moment of order n

In general, the moment of order n is the expected value (the mean) of X^n, and is defined as

$$m_X^{(n)} = \int_{-\infty}^{\infty} x^n\ f_X(x)\ dx.$$

It can be seen as the expected value of a function of X, in this case the function $g(x) = x^n$. Therefore, the mean is the first order moment.

Variance

The variance can be viewed as the expected value for the particular case of the function

$$g(x) = (x - m_X)^2.$$

Therefore,

$$\sigma_X^2 = E\left[(X - m_X)^2\right] = \int_{-\infty}^{\infty} (x - m_X)^2\ f_X(x)\ dx.$$

σ_X^2 is the variance of the random variable and σ_X is the standard deviation. These parameters give us an idea of the variability of the random variable. Interestingly, the mean and the variance have the following relationship (by means of the second order moment):

$$\sigma_X^2 = E\left[(X - E(X))^2\right] = E\left[X^2\right] - (E[X])^2 .$$

$$\sigma_X^2 = E\left[(X - m_X)^2\right] = m_X^{(2)} - (m_X)^2 .$$

Properties

Below, some of the properties of these statistics are shown. In this list of properties, c denotes an arbitrary constant.

1. $E[X + Y] = E[X] + E[Y] = m_X + m_Y$ (Linearity)
2. $E[c] = c$
3. $E[c\,X] = c\,E[X]$
4. $E[X + c] = E[X] + c$
5. $\mathrm{Var}(c) = 0$
6. $\mathrm{Var}(c\,X) = c^2\,\mathrm{Var}(X)$
7. $\mathrm{Var}(X + c) =$Var(X)

1.2.6 Multidimensional (multiple) random variables

When two random variables are defined on the same sample space Ω, it is possible to work with them jointly. This case can be posed as a multidimensional problem, or also as a problem of vectors of random variables. We will follow the first alternative.

Joint probability density and distribution functions

For two random variables X and Y, their *joint distribution function* is defined as

$$F_{X,Y}(x, y) = P(X \leq x, Y \leq y).$$

The *joint probability density function* is

$$f_{X,Y}(x, y) = \frac{\partial^2}{\partial x \partial y} F_{X,Y}(x, y).$$

These two functions have the following properties (most of them are extension of the properties of CDF and PDF for a single random variable):

1. $F_X(x) = F_{X,Y}(x, \infty)$.

2. $F_Y(y) = F_{X,Y}(\infty, y)$.
3. $f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\, dy$.
4. $f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\, dx$.
5. $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{X,Y}(x,y)\, dx\, dy = 1$.
6. $P((X,Y) \in A) = \int\int_{(x,y)\in A} f_{X,Y}(x,y)\, dx\, dy$.
7. $F_{X,Y}(x,y) = \int_{-\infty}^{x}\int_{-\infty}^{y} f_{X,Y}(u,v)\, du\, dv$.

Conditional probability density function

As in the case of events, knowing the outcome of a random variable can condition the knowledge about the other. The probability density function of the variable Y conditioned by $X = x$ is defined as

$$f_{Y|X}(y|x) = \begin{cases} \frac{f_{X,Y}(x,y)}{f_X(x)}, & f_X(x) \neq 0 \\ 0, & \text{in other case} \end{cases}.$$

The definition of *statistically independent random variables* rises from this definition. If knowledge of X contributes nothing to knowledge of Y and vice versa, then

$$f_{Y|X}(y|x) = f_Y(y), \text{ and also } f_{X|Y}(x|y) = f_X(x).$$

Therefore, two random variables are independent if their conditional distributions are equal to the marginal distributions. From this definition of independency, an implication naturally appears: for independent random variables, the joint distribution is equal to the product of the marginal distributions

$$f_{X,Y}(x,y) = f_X(x) \times f_Y(y).$$

Statistic moments

The expected value of a function $g(X,Y)$ of the random variables X and Y is obtained as

$$E\left[g(X,Y)\right] = \int_{\infty}^{\infty}\int_{\infty}^{\infty} g(x,y)\, f_{X,Y}(x,y)\, dx\, dy.$$

It is interesting to highlight the following particular cases:

- If $g(X,Y) = X \times Y$, the expectation of the product of the two random variables is obtained, which is called the *correlation* between X and Y.
- For the case $g(X,Y) = (X - m_X) \times (Y - m_Y)$, the so called *covariance* is obtained.

The normalized version of the covariance is what is known as the correlation coefficient, $\rho_{X,Y}$, which is defined as

$$\rho_{X,Y} = \frac{\text{Cov}(X,Y)}{\sigma_X \sigma_Y}.$$

The range for this coefficient is $0 \leq |\rho_{X,Y}| \leq 1$, or equivalently

$$-1 \leq \rho_{X,Y} \leq +1.$$

Some particular values of this coefficient give us special information about the involved random variables.

- The value $\rho_{X,Y} = 0$ means that the variables are *incorrelated*. If two random variables are independent, then they are always incorrelated. However, the converse is not true: uncorrelation does not imply independence.
- On the other hand $\rho_{X,Y} = \pm 1$ indicates a linear relationship between the random variables, i.e. $Y = aX + b$. In this case, $\rho_{X,Y} = +1$ indicates a positive value of a, while $\rho_{X,Y} = -1$ indicates that a is negative.

It is common to use the notation ρ, without referring to the random variables involved when they are implicit.

Intuitively, the correlation will indicate the degree of statistical relationship between the two random variables. In general, a high correlation indicates a high relationship, and a low correlation usually indicates a low relationship.

Functions of multidimensional random variables

For multidimensional (or multiple) random variables, as for one-dimensional ones, functions can be defined on the variables X and Y

$$\begin{cases} Z = g(X,Y) \\ W = h(X,Y) \end{cases}.$$

To obtain $F_{Z,W}(z,w)$ the procedure is similar to the unidimensional case:

$$F_{Z,W}(z,w) = P(Z \leq z, W \leq w) = P((x,y) \in B^{g,h}_{X,Y}(z,w)),$$

where now

$$B^{g,h}_{X,Y}(z,w) = \{(x,y) \in I\!R^2 : g(x,y) \leq z, h(x,y) \leq w\}.$$

As in the case of a single r.v., if the roots (solutions) $\{x_i, y_i\}$ of the equations

$$\begin{cases} z = g(x,y) \\ w = h(x,y) \end{cases},$$

are know, then the joint PDF of the new variables is obtained as follows

$$f_{Z,W}(z,w) = \sum_i \frac{f_{X,Y}(x_i,y_i)}{|\text{det}\boldsymbol{J}(x_i,y_i)|}.$$

where det$\boldsymbol{J}$ denotes the determinant of the Jacobian matrix $\boldsymbol{J}$. It is necessary that the number of solutions is finite and that the Jacobian is not zero. The Jacobian is defined as

$$\boldsymbol{J}(x,y) = \begin{bmatrix} \frac{\partial z(x,y)}{\partial x} & \frac{\partial z(x,y)}{\partial y} \\ \frac{\partial w(x,y)}{\partial x} & \frac{\partial w(x,y)}{\partial y} \end{bmatrix}.$$

Most of the previous definitions and results for two random variables can be immediately extended to a larger number of random variables.

Jointly Gaussian (normal) random variables

Jointly Gaussian random variables are also sometimes called multidimensional Gaussian random variables. Due to its importance for this subject, some of its properties are presented below. First, they are defined probabilistically. Two jointly Gaussian random variables X and Y are characterized by a joint probability density function that is a two-dimensional Gaussian

$$f_{X,Y}(x,y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} e^{-\frac{1}{(1-\rho^2)}\left(\frac{(x-\mu_X)^2}{2\sigma_X^2} + \frac{(y-\mu_Y)^2}{2\sigma_Y^2} - \frac{\rho(x-\mu_x)(y-\mu_Y)}{\sigma_X\sigma_Y}\right)}.$$

When X and Y have this type of jointly Gaussian distribution, not only are X and Y have individually a Gaussian distribution (each one is a Gaussian r.v.) but the conditional probabilities are Gaussian as well. This is the main difference between two random variables that each have a Gaussian distribution and two random variables with a jointly Gaussian distribution. With a jointly Gaussian distribution, the individual random variables are as follows: X is Gaussian with mean μ_X and variance σ_X^2, Y is Gaussian with mean μ_Y and variance σ_Y^2, and also its correlation coefficient is ρ.

This concept can be extended to an arbitrary number n of random variables, arriving at the expression for the distribution of a n-dimensional Gaussian, parameterized by a vector of means and a matrix of covariances

$$f_{\mathbf{X}}(x_1, x_2, \ldots, x_n) = \frac{1}{\sqrt{(2\pi)^n \det(\boldsymbol{C})}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})\boldsymbol{C}^{-1}(\mathbf{x}-\boldsymbol{\mu})^T}.$$

where $\mathbf{X}$ is the vector of the random variables, $\mathbf{X} = (X_1, X_2, \ldots, X_n)$, $\mathbf{x} = [x_1, x_2, \ldots, x_n]^T$, and the vector of means is $\boldsymbol{\mu} = [\mu_1, \mu_2, \ldots, \mu_n]^T$. Finally, $\boldsymbol{C}$ is the covariance matrix that contains in the i-th row and j-th column the covariance among the i-th and the j-th variables:

$$C_{i,j} = \mathrm{Cov}(X_i, X_j) = \rho_{i,j}\sigma_i\sigma_j,$$

i.e.,

$$\boldsymbol{C} = \begin{bmatrix} \sigma_1^2 & \rho_{1,2}\sigma_1\sigma_2 & \ldots & \rho_{1,n}\sigma_1\sigma_n \\ \rho_{1,2}\sigma_1\sigma_2 & \sigma_2^2 & \ldots & \rho_{2,n}\sigma_2\sigma_n \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{1,n}\sigma_1\sigma_n & \rho_{2,n}\sigma_2\sigma_n & \ldots & \sigma_n^2 \end{bmatrix}.$$

The main properties of jointly Gaussian random variables are:

1. Jointly Gaussian random variables are completely characterized by their vector of means $\boldsymbol{\mu}$ and their covariance matrix $\boldsymbol{C}$. These two parameters are called *second-order statistics*, and they fully describe these random variables.

2. If n random variables are jointly Gaussian, then any subset is also jointly Gaussian distributed. In particular, all individual variables are Gaussian.

3. Any subset of jointly Gaussian r.v.(s), conditioned on another subset of the same original jointly Gaussian variables, has a jointly Gaussian distribution (the parameters, means and covariances, can be modified in this case).

4. Any set of random variables obtained as linear combinations of $(X_1, X_2, \ldots, X_n)$

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} a_{1,1} & a_{1,2} & \ldots & a_{1,n} \\ a_{2,1} & a_{2,2} & \ldots & a_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n,1} & a_{n,2} & \ldots & a_{n,n} \end{bmatrix} \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{bmatrix} + \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{bmatrix},$$

   are jointly Gaussian. In particular, individually any linear combination Y_i is Gaussian.

5. Two uncorrelated jointly Gaussian random variables are independent. Therefore, *for jointly Gaussian random variables, independence and uncorrelation are equivalent.* This is not true in general for other types of rvariables)

6. If the variables are uncorrelated, $\rho_{i,j} = 0 \; \forall i \neq j$, that is, $\boldsymbol{C}$ is a diagonal matrix.

Sum of random variables

Given a sequence of random variables, $(X_1, X_2, \ldots, X_n)$, which basically have the same properties, it seems logical to think that the behavior of their average,

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

be, so to speak, "less random". The *Law of large numbers* and the *Central limit theorem* rigorously state this intuition.

Law of Large Numbers (weak) This law states that if the random variables $(X_1, X_2, \ldots, X_n)$ are *uncorrelated* and all have the same mean m_X and variance $\sigma_X^2 < \infty$, regardless of their distribution, for any $\varepsilon > 0$,

$$\lim_{n \to \infty} P(|Y - m_X| > \varepsilon) = 0.$$

This means that the average (Y) converges, in probability, to the mean of the variables, m_X. In other words, the more variables we add, the more their combination resembles the mean (the lower their variance).

Central Limit Theorem This theorem goes further than the Law of Large Numbers. Not only does it say that the average of random variables converges to the mean, but it also tells us what their distribution is like. Specifically, the theorem states that: if $(X_1, X_2, \ldots, X_n)$ are n *independent* random variables with means $m_1, m_2, \ldots, m_n$, and variances $\sigma_1^2, \sigma_2^2, \ldots, \sigma_n^2$, then the distribution of

$$Y = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{X_i - m_i}{\sigma_i}$$

converges to a normal distribution, with zero mean and unit variance, $\mathcal{N}(0, 1)$.

In the particular case that they are *independent and identically distributed (i.i.d)*, that is, that they all have the same distribution with the same mean m and the same variance σ^2, the average

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

converges to a normal distribution $\mathcal{N}(m, \frac{\sigma^2}{n})$. This occurs even if the original distribution is not Gaussian.

Remark: It should be noted that the Law of Large Numbers is valid for uncorrelated random variables while the Central Limit Theorem requires independence between random variables, which is a stronger constraint.

1.3 Random processes

A random process, or stochastic process, is the natural extension of the concept of random variable to work with signals. Communication systems work with signals, which are time functions. As has already been mentioned several times, sometimes it is possible to characterize these signals deterministically, and in other times it will be necessary to treat them as random signals: the clearer examples are the thermal noise in any device, or the information signals themselves. These signals will be characterized as random processes.

Perhaps the most intuitive way to see what a random process is is to think of it as a set of time signals corresponding to each of the possible outcomes of a random experiment. Each output of an experiment has a time function associated with it. A real random variable assigns a real value to each value in the sample space ($\Omega \to I\!R$), that is, $\lambda_i \in \Omega \to X(\lambda_i) \in I\!R$. A stochastic process can be interpreted as a situation in which the assignment of values from the sample space to the real line varies with time $X(t, \lambda)$. From this point of view, each output of the experiment is associated with a time function that specifies its actual value assigned at a time t. Below are some examples of random processes.

Below are several examples of random processes.

Example

Random experiment: throwing a dice, with 6 possible outcomes

$$\lambda \in \{\lambda_1, \lambda_2, \lambda_3, \lambda_4, \lambda_5, \lambda_6\}$$

The random process is defined by selecting the 6 signals associated with each possible outcome of the experiment.

$$X(t, \lambda_i) = \frac{1}{2} + \sin(\omega_0 t - \theta_i)$$

$$\text{con } \theta_i = (i-1)\frac{2\pi}{6}$$

$$\text{para } i \in \{1, 2, 3, 4, 5, 6\}$$

From now on we will call this example *Example I*, and the 6 functions that make up the process are shown in Figure 1.13.

Example

Same random experiment: throwing a dice, with 6 possible outcomes

$$\lambda \in \{\lambda_1, \lambda_2, \lambda_3, \lambda_4, \lambda_5, \lambda_6\}$$

The random process is defined by defining each of the 6 signals associated with each possible output of the experiment, which are now

$$X(t, \lambda_i) = \frac{1}{2}\sin(\omega_0 t - \theta_i) + \frac{1}{2}\cos(\omega_0 t)$$

$$\text{with } \theta_i = (i-1)\frac{2\pi}{6}$$

$$\text{for } i \in \{1, 2, 3, 4, 5, 6\}$$

From now on we will call this example *Example II*, and the 6 functions that define the process are shown in Figure 1.14.

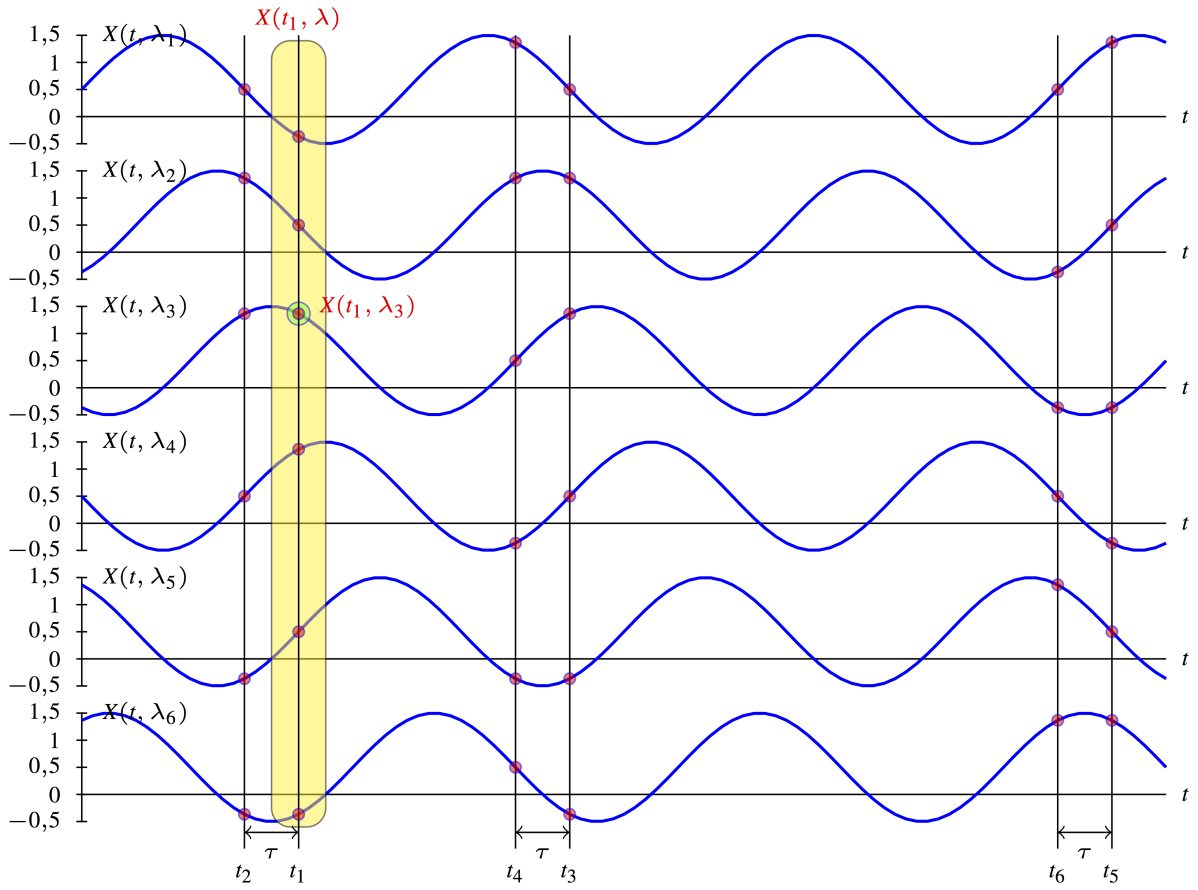


Figure 1.13: Example I: Signals associated to each outcome in Ω.

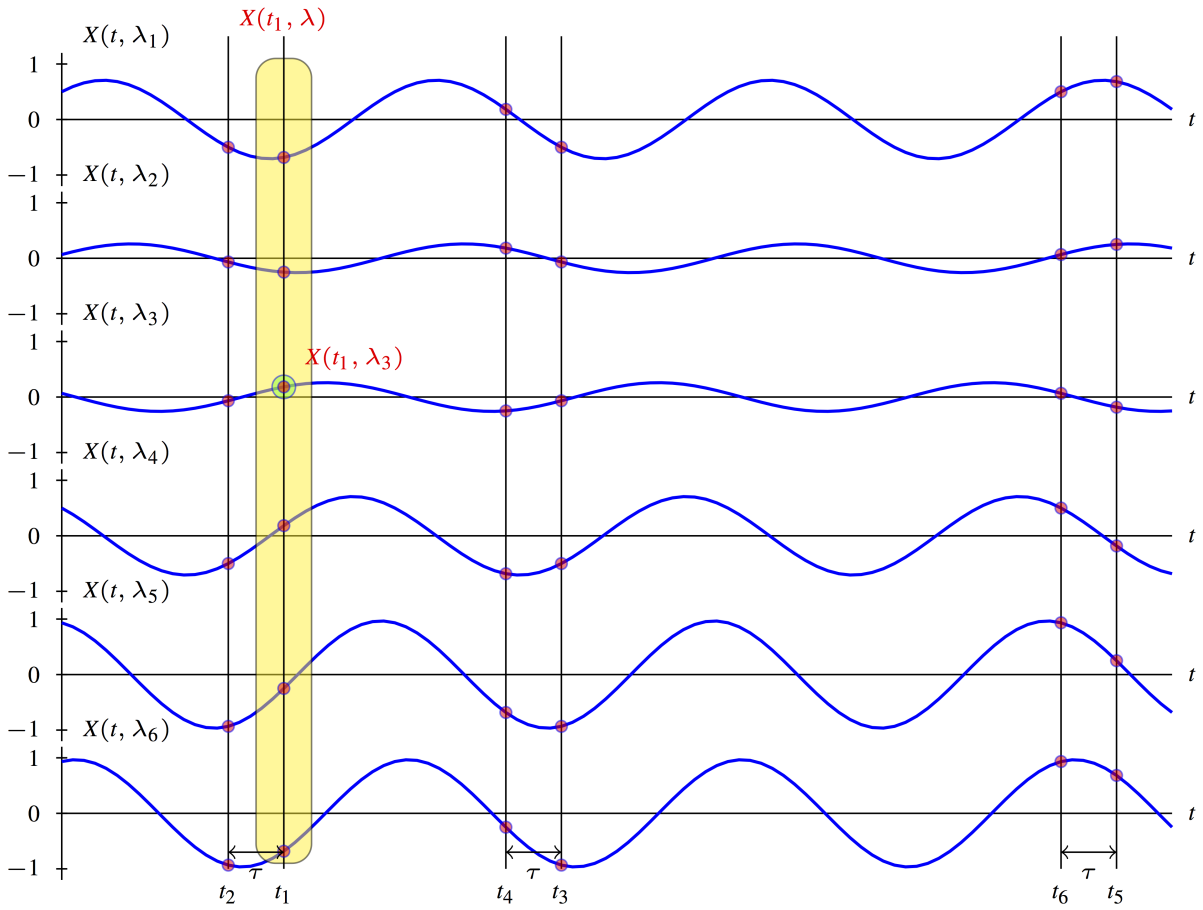


Figure 1.14: Example II: Signals for each outcome in Ω.

In view of the previous examples, and taking into account that a random process has two arguments, the time index and the result of the random experiment, the following values or functions can be identified for a process:

- $X(t_i, \lambda_j)$, the case in which the values of the two arguments are fixed, is an individual outcome of the experiment at a given moment, which gives rise to a real value.
- $X(t, \lambda_i)$, when the result of the random experiment is fixed, it is a time signal that indicates the real number assigned at each instant to a possible output, λ_i, of the sample space. To simplify the notation, it will sometimes also be denoted as $x_i(t)$.
- $X(t_i, \lambda)$, when a time instant is fixed, it is a set of numbers, each one associated with each possible outcome of the random experiment. It is therefore a random variable (X). Thus, *at any fixed time instant t_i, a random process is a random variable.*

There are, as we have seen, several ways of interpreting a random process: as a set of signals, or as an indexed set of random variables, where the index is a time index, either in continuous time or in discrete time.

$$\{X(t_1), X(t_2), \ldots\}, \text{ or } \{X[n_1], X[n_2], \ldots\},$$

or in general as

$$\{X(t),\ t \in I\!R\}, \text{ or } \{X[n],\ n \in \mathbb{Z}\}.$$

Therefore, a random process can be defined as a set of random variables indexed by a certain time index (continuous or discrete).

- If the index is t, taking values in the continuous set of real numbers $t \in I\!R$, the process is a *continuous-time random process*
- If the index is the discrete set of values n, with $n \in \mathbb{Z}$, the process is a *discrete-time random process.*

For this reason, the notation often ignores the dependency on the output of the random experiment λ, starting to use the notation $X(t)$ or $X[n]$.

Next, random processes in continuous time will be studied, and later the main results will be extended, in a trivial way, to discrete time processes.

1.3.1 Description of a random process

A random process can be described by two types of descriptions:

- Analytical
- Statistical

The *analytical description* uses a compact analytical expression of the process. In this expression, the time index is included along with a set of random variables

$$X(t) = f(t, \boldsymbol{\theta}).$$

The vector $\boldsymbol{\theta} = \{\theta_1, \theta_2, \ldots, \theta_n\}$ is a vector that includes the random variables.

The analytic description includes the analytic expression involving the time index and the random variables, and the statistical description of the random variables included in $\boldsymbol{\theta}$.

Example

- $X(t) = A\cos(2\pi f_0 t + \theta)$, where A and f_0 are two constant values and θ is a uniform random variable in $[0, 2\pi)$. This is an analytical description of a continuous-time random process that could, for example, be used to model the output of an oscillator in a communications system (by assigning A and f_0 the values of the voltage amplitude and the oscillator frequency, respectively).
- $X[n] = A\theta$, where A is a constant value and θ is Bernoulli variable with $p = 0.5$. This is an analytical description of a discrete-time random process that can for example be used to statistically represent the transmission of a sequence of bits (by making the constant $A = 1$) with equiprobable ones and zeros.

The analytical description provides intuitive information about the random process, because it describes its performance in an analytical way. However, it is not always possible to have this type of description in real applications. In this case, an *statistical description* can be used. There are several types of statistical description. The most common are described below.

Statistical description

A *(complete) statistical description* of a random process $X(t)$ consists in knowing for any set of n time instants $\{t_1, t_2, \ldots, t_n\}$, for any value of n, the joint probability density function of the n random variables resulting from the evaluation of the random process at the n specified time instants, $\{X(t_1), X(t_2), \ldots, X(t_n)\}$

$$f_{X(t_1),X(t_2),\ldots,X(t_n)}(x_1, x_2, \ldots, x_n).$$

M-th order statistical description

This is a description similar to the previous one in which the maximum value of n is limited. In particular, this description consists of knowing, for every $n \leq M$ and every set of n instants of time $\{t_1, t_2, \ldots, t_n\}$, the joint probability density function of $\{X(t_1), X(t_2), \ldots, X(t_n)\}$.

In the analysis and design of communication systems, is common to use second order descriptions, $M = 2$, where the distribution is known for any pair of time instants (t_1, t_2)

$$f_{X(t_1),X(t_2)}(x_1, x_2).$$

Example

For a process $X(t)$, for any n and any $(t_1, t_2, \ldots, t_n) \in \mathbb{R}^n$, the PDF of the random variables $\{X(t_i)\}_{i=1}^n$ is a jointly Gaussian distribution, with zero mean and the followint covariance matrix

$$C_{i,j} = \mathrm{Cov}(X(t_i), X(t_j)) = \sigma^2 \min(t_i, t_j).$$

This is a complete statistical description of $X(t)$.

In this last example, although a complete statistical description of the process is provided, there is little information on how each realization of the process is. That is why in many cases a set of statistical averages is obtained to give us that information.

1.3.2 Statistic averages

As we saw for random variables, for random processes some statistical averages can be calculated based on expected values. In particular, the two most important statistics in the time domain are:

- Mean of the random process, $m_X(t)$.
- Autocorrelation function of the random process, $R_X(t_1, t_2)$.

Mean of a random process

The *mean* or *expectation (mathematical expectation)* of a random process $X(t)$ is a deterministic time function, $m_X(t)$. For each time instant t, the mean of the process is the mean of the random variable $X(t)$

$$m_X(t) = E[X(t)].$$

If the statistical description of the process is available, and if for any instant t the PDF $f_{X(t)}(x)$ is defined, this function can be computed from it as

$$m_X(t) = E[X(t)] = \int_{-\infty}^{\infty} x \; f_{X(t)}(x) \; dx.$$

Autocorrelation function of a random process

Another important statistical average of a random process is the *autocorrelation function*. This function is important because, as will be seen later, it is related to the frequency domain description of the random process, and to the power of the process. It is a second order statistic, as it depends on two time instants. Sometimes it is denoted as $R_{X,X}(t_1, t_2)$, although the usual notation is $R_X(t_1, t_2)$.

The autocorrelation function is defined as the mathematical expectation of the product of the random process evaluated at the two instants that are the arguments of the function, with the second random variable conjugated for complex processes

$$R_X(t_1, t_2) = E[X(t_1)X^*(t_2)],$$

Again, if the statistical description of the random process is available, the autocorrelation function can be obtained through the joint probability density function of the process at two instants as

$$R_X(t_1, t_2) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x_1 \; x_2^* \; f_{X(t_1),X(t_2)}(x_1, x_2) \; dx_1 \; dx_2.$$

In summary, two methods have been seen to describe random processes: analytical description and statistical description. In this case, you can have a complete or a M-th order statistical

description. It is common in communication systems to have a second-order description. This description is sometimes impractical in the sense that it does not give a clear idea of how the realizations of the process are. In this case, statistics are used that give averages over the executions. In particular, the mean and the autocorrelation function are interesting. As we will see later, in some cases the mean and the autocorrelation function provide a complete statistical description of the random process (this will be the case for Gaussian random processes).

Time autocorrelation function of a deterministic function

Due to the similarity of the names, the autocorrelation function of a random process is sometimes confused with the temporal autocorrelation function of a deterministic signal. The first is a statistical average that provides information about the statistics of a random process. The second is a deterministic function, applied to a deterministic signal, so the nature of these two functions is completely different, and it is important to understand the difference between them. As it will be useful later, the time autocorrelation function, also called the time ambiguity function, is defined below. For a given deterministic function $x(t)$, its time autocorrelation function is denoted as $r_x(t)$, and is defined as the convolution of that signal, $x(t)$, with its matched signal, $x^*(-t)$

$$r_x(t) = x(t) * x^*(-t).$$

For real signals, obviously, the complex conjugate operator is irrelevant, and therefore

$$r_x(t) = x(t) * x(-t).$$

In the frequency domain, if the Fourier transform of $x(t)$ is $X(j\omega)$, and taking into account that, due to the properties of the Fourier transform, the transform of the matched signal is $X^*(j\omega)$, the Fourier transform of the time autocorrelation function is

$$R_x(j\omega) = |X(j\omega)|^2.$$

The function $r_x(t)$ has some interesting properties, which will be used later. Some of them are:

- It is a hermitian function (symmetric for real signals), whose maximum is at zero.
- Allows us to calculate the energy of the signal $x(t)$, both in the time domain and in the frequency domain

  $$\mathcal{E}\{x(t)\} = r_x(0) = \frac{1}{2\pi}\int_{-\infty}^{\infty} R_x(j\omega)\, d\omega$$

  This property is evident considering the definition of energy (Parseval's relation)

  $$\mathcal{E}\{x(t)\} = \int_{-\infty}^{\infty} |x(t)|^2\, dt = \frac{1}{2\pi}\int_{-\infty}^{\infty} |X(j\omega)|^2\, d\omega$$

  and the properties of the Fourier transform: by its own definition, the value at zero in one domain is equal to the integral in the other domain.
- It is a translation invariant function of $x(t)$, i.e.

  $$y(t) = x(t - t_0) \rightarrow r_y(t) = r_x(t).$$

Figure 1.15 shows an example of some of these properties on a rectangular signal of duration T seconds.

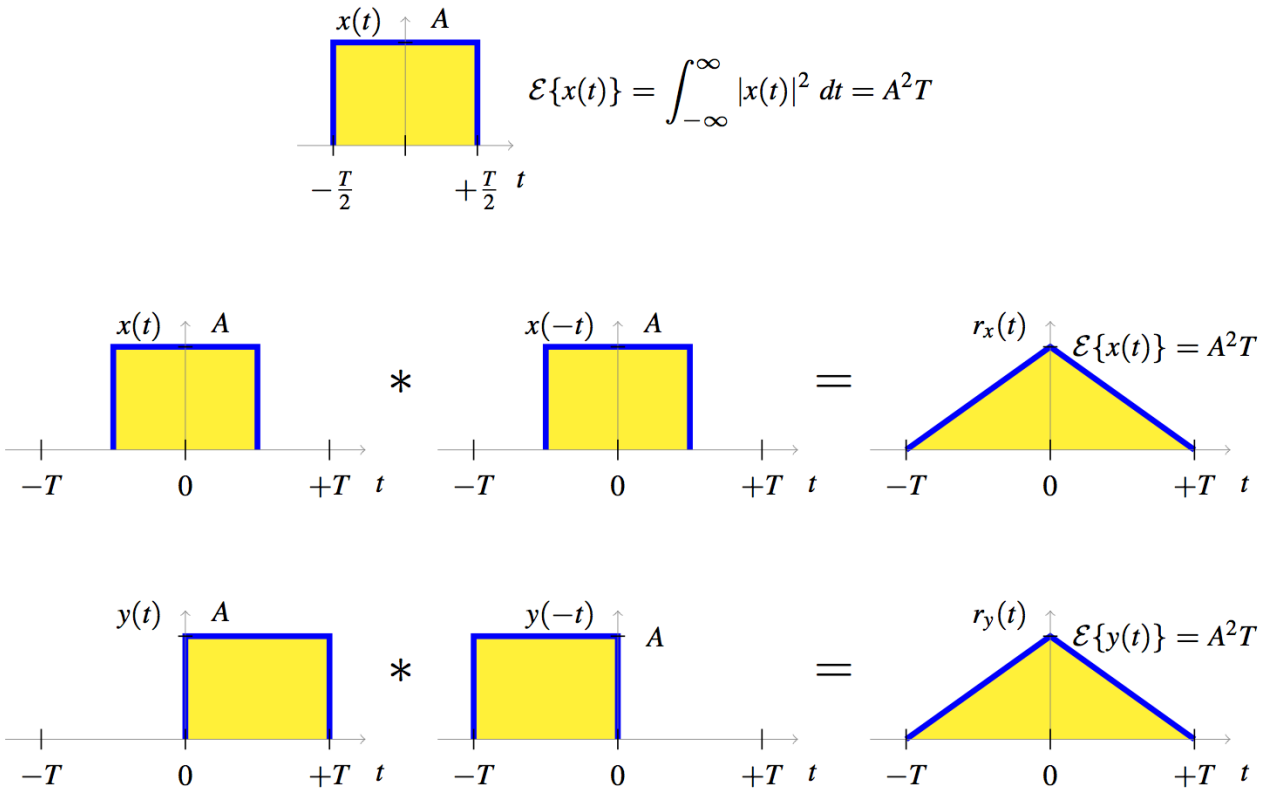


Figure 1.15: Example of the time autocorrelation function for a rectangular signal of duration T seconds, illustrating some of its properties.

1.3.3 Stationarity and ciclostationarity

The joint PDF of $\{X(t_1), X(t_2), \ldots, X(t_n)\}$ is

$$f_{X(t_1),X(t_2),\ldots,X(t_n)}(x_1, x_2, \ldots, x_n)$$

for any set of n time instants $\{t_1, t_2, \ldots, t_n\}$ and for any value of n.

In general, this joint distribution depends on the choice of the time reference. But there is a very important class of random processes in which that function is independent of the time, i.e. these processes have statistical properties that do not vary over time. These processes are called *stationary processes*.

There are two definitions of stationarity, in the strict sense and in the broad sense. Each of the definitions is shown below.

Strict sense stationarity

A process is *strict sense stationary* if for any set of n time instants $\{t_1, t_2, \ldots, t_n\}$, for any integer value n, and any value Δ

$$f_{X(t_1),X(t_2),\ldots,X(t_n)}(x_1, x_2, \ldots, x_n) = f_{X(t_1+\Delta),X(t_2+\Delta),\ldots,X(t_n+\Delta)}(x_1, x_2, \ldots, x_n).$$

The PDF of the random process in any set of n time instants does not depend on the specifit time intants but only on the relative difference among these n time instants $\{t_i\}_{i=1}^{n}$.

When this is only true for $n \leq M$, then the process is said to be *stationary of order M*.

Strict sense stationarity is a very strong constraint that very few real processes can meet. For this reason, a less restrictive definition of stationarity is often used.

Wide sense stationarity

A random process $X(t)$ is *wide sense stationary* (WSS) if the following conditions are satisfied:

1. The mean is a constant value that does not depend on time.

$$m_X(t) = E[X(t)] = m_X$$

2. The autocorrelation function does not depend explicitly on each of the two values of time, t_1 and t_2, but only on their difference

$$R_X(t_1, t_2) = R_X(t_1 - t_2) = R_X(\tau)$$

This expression emphasizes the fact that it depends only on the difference of time instants, the parameterization $t_1 = t + \tau$ and $t_2 = t$ is often used, so that it can be written as

$$R_X(t + \tau, t) = R_X(\tau).$$

From now on, when we talk about stationarity, we will refer to wide sense stationarity.

Autocorrelation function of stationary processes

The autocorrelation function of a real stationary process $X(t)$, $R_X(\tau)$, has the following properties:

1. $R_X(-\tau) = R_X(\tau)$. It is an even function.

$$R_X(\tau) = E[X(t)X(t-\tau)] = E[X(t-\tau)X(t)] = R_X(-\tau).$$

2. $|R_X(\tau)| \leq R_X(0)$. The maximum in module is obtained in $\tau = 0$.

$$E[(X(t) \pm X(t-\tau))^2] \geq 0$$

Expanding this result

$$E[X^2(t)] + E[X^2(t-\tau)] \pm 2E[X(t)X(t-\tau)] \geq 0.$$

$$R_X(0) + R_X(0) \pm 2R_X(\tau) \geq 0 \rightarrow R_X(0) \geq \pm R_X(\tau) \rightarrow R_X(0) \geq |R_X(\tau)|.$$

3. If for some T_o $R_X(T_o) = R_X(0)$ holds, then for every integer k

$$R_X(kT_o) = R_X(0).$$

The proof by induction can be found in [Proakis and Salehi, 2002].

4. It is a positive semidefinite function. Formally, this means that, for any function $g(t)$, it is true that
$$\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} g(t)\, R_X(t-s)\, g(s)\, dt\, ds \geq 0.$$
This property has an important practical implication, which is that the Fourier transform of the autocorrelation function is non-negative (it only takes values greater than or equal to zero, but never negative values). The proof of this property will be seen in a trivial way later, when the representation in the frequency domain of a random process is studied.

Ciclostationary processes

There is a special class of non-stationary processes that are closely related to stationary processes, the so-called cyclostationary processes. In this case, the statistical properties are not constant over time, they vary over time, but these variations are periodic in time. The definition of cyclostationarity of a random process (in the wide sense) reduces again to conditions on the mean and the autocorrelation function of the process. Specifically, a random process $X(t)$ is cyclostationary (in the wide sense) if its mean and its autocorrelation function are periodic with a certain period T_0, i.e., if it is satisfied that

1. $m_X(t+T_0) = m_X(t)$.
2. $R_X(t+\tau+T_0, t+T_0) = R_X(t+\tau, t)$, for all t and τ.

In the previous examples of random processes, the *Example I* corresponds to a stationary process, while the *Example II* corresponds to a cyclostationary process. The analytical expressions for the statistical parameters of the *Example I* are
$$m_X(t) = \frac{1}{2}$$
for the mean, which takes a constant value, and
$$R_X(t_1, t_2) = \frac{1}{4} + \frac{1}{2}\cos(\omega_0(t_1 - t_2))$$
for the autocorrelation function. If instead of parameterizing using two different instants t_1 and t_2 we use a parameterization with two time instants with a separation τ seconds, i.e. $t_1 = t + \tau$ and $t_2 = t$, the resulting expression is
$$R_X(t+\tau, t) = \frac{1}{4} + \frac{1}{2}\cos(\omega_0\tau),$$
where it can now be clearly seen that this function depends only on the separation between the time instants, τ, and not on the concrete instants (it does not depend on t). Therefore, it can be assured that the *Example I* is a stationary random process. Figure 1.16 shows the mean and autocorrelation function of this process.

In Figure 1.17 it can be seen that for a certain value of the difference between instants, τ, the different values of t are located in the plane t_1 vs t_2 along a straight line parallel to the straight line $t_1 = t_2$ (which is the particular case for a separation $\tau = 0$). In the case of the autocorrelation function of the *Example I*, it takes constant values along each of the straight lines that determines

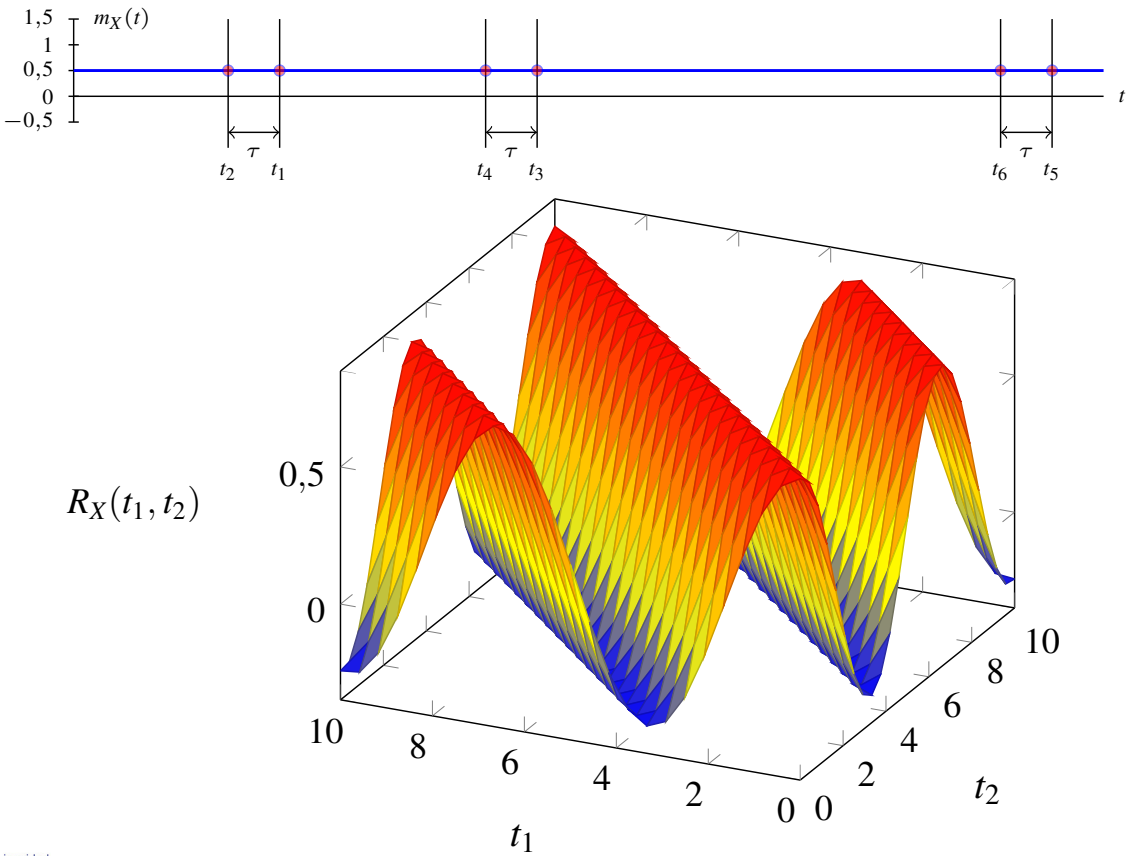


Figure 1.16: Example I: Mean and autocorrelation function.

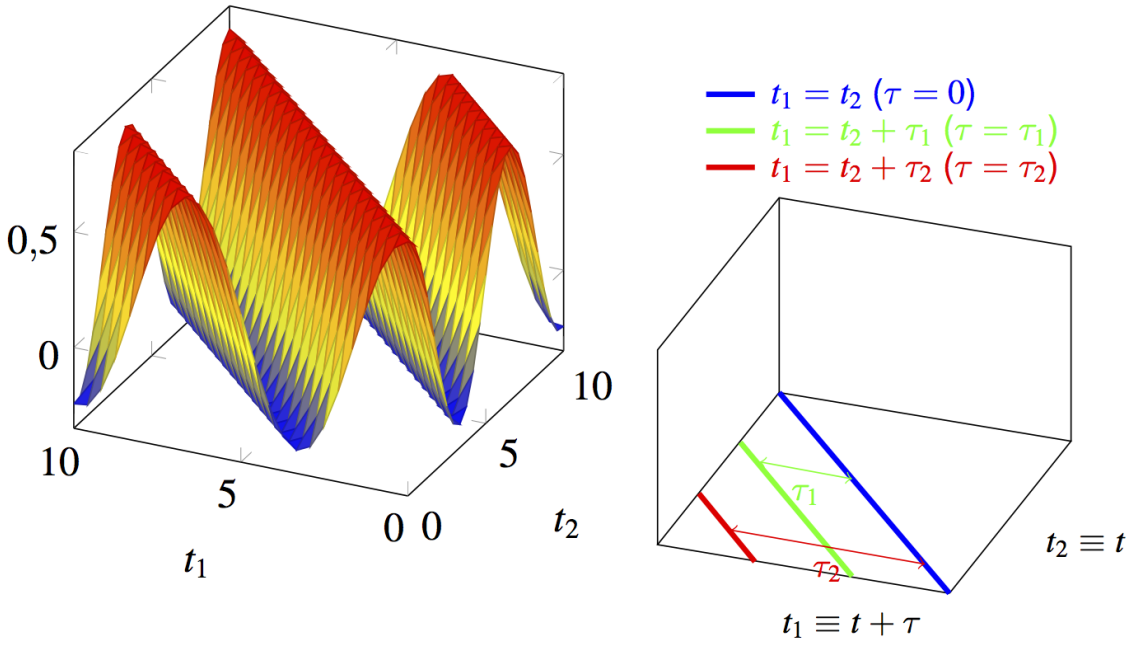


Figure 1.17: Example I: Autocorrelation function and illustration of the values for different separations between the time instants.

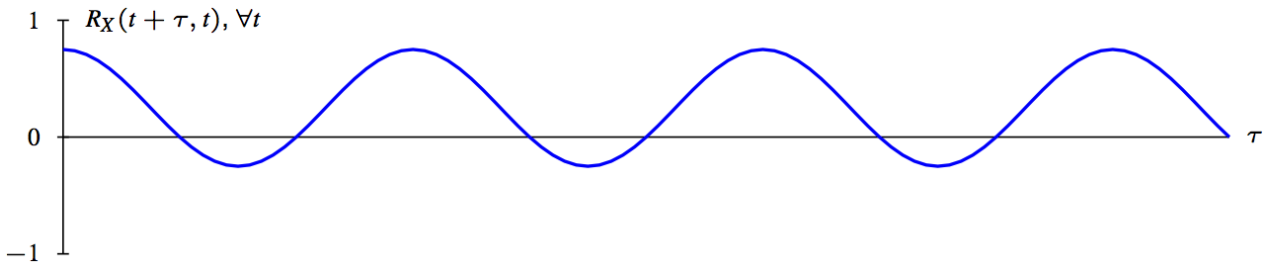


Figure 1.18: Example I: Autocorrelation function as a function of the difference between the two instants of time (parameter τ).

each value of τ, having only a dependence on the separation variable τ. Specifically, regardless of the value of t the autocorrelation function has a sinusoidal variation with the difference between the two time instants, τ, as shown in Figure 1.18.

In the case of *Example-II*, the mean of the process is

$$m_X(t) = \frac{1}{2}\cos(\omega_0 t)$$

which now depends on time, and the autocorrelation function is

$$R_X(t_1, t_2) = \frac{1}{4}\cos(\omega_0(t_1 - t_2)) + \frac{1}{4}\cos(\omega_0(t_1 + t_2))$$

which, parameterized as a function of the difference between the two temporary arguments, is written as

$$R_X(t+\tau, t) = \frac{1}{4}\cos(\omega_0 \tau) + \frac{1}{4}\cos(\omega_0(2t + \tau)).$$

Figure 1.19 shows the mean and the autocorrelation function of *Example II*. The autocorrelation function, like the mean, now depends on time t, although in both cases the dependency is periodic. Specifically, the autocorrelation function has a sinusoidal variation with the difference between the two time instants, τ, but that variation is different for different values of t, as shown in Figure 1.20. Therefore, it can be ensured that *Example-II* is a cyclostationary process.

The random processes in these two examples are one stationary, and the other cyclostationary. Naturally, there are also random processes that are neither stationary nor cyclostationary. Below is a very simple example.

Example

Random experiment: throwing a dice, with 6 possible outcomes

$$\lambda \in \{\lambda_1, \lambda_2, \lambda_3, \lambda_4, \lambda_5, \lambda_6\}$$

The random process is defined by defining each of the 6 signals associated with each possible output of the experiment.

$$X(t, \lambda_1) = \frac{1}{4}, \qquad X(t, \lambda_2) = -\frac{1}{4}$$

$$X(t, \lambda_3) = u(t-2) = \begin{cases} 1, & \text{si } t \geq 2 \\ 0, & \text{si } t < 2 \end{cases}$$

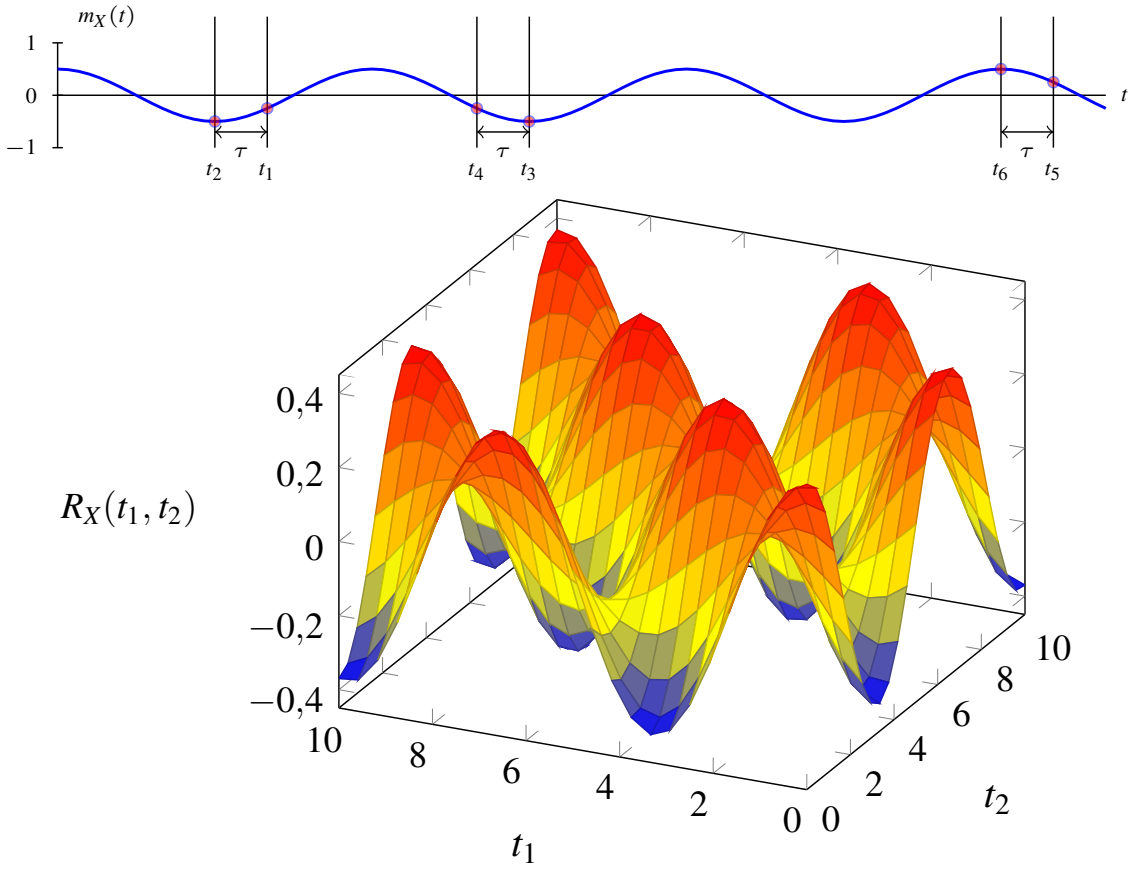


Figure 1.19: Example II: Mean and autocorrelation function.

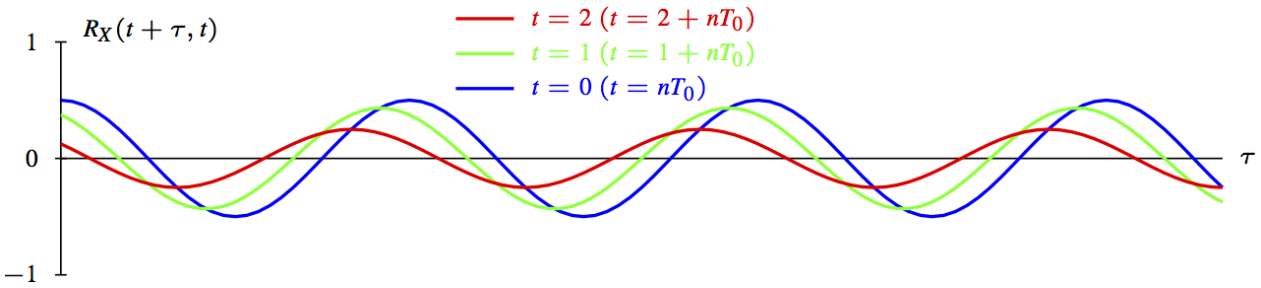


Figure 1.20: Example II: Autocorrelation function as a function of the difference between the two instants of time (parameter τ).

$$X(t,\lambda_4) = 1 - \frac{t}{5}$$

$$X(t,\lambda_5) = e^{-t}, \qquad X(t,\lambda_6) = \sin(\pi t)$$

From now on we will call this example *Example III*, and the 6 functions that make up the process are shown in Figure 1.21, together with the mean of the process. It is clear that the mean of the process depends on time, but it is not a periodic function, so it is a random process that is neither stationary nor cyclostationary.

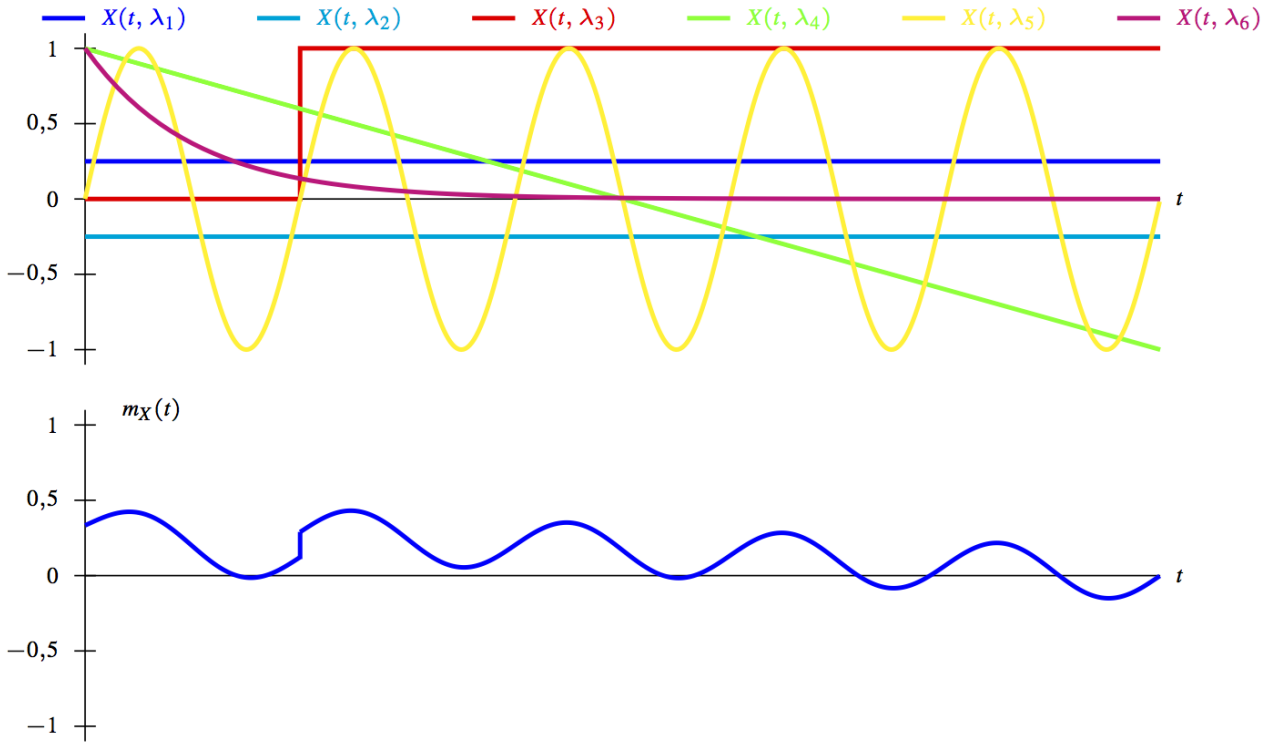


Figure 1.21: Example III: Signals and mean of the random process.

1.3.4 Ergodicity

To see what an ergodic process is, a simple example will be used. Suppose we have a process $X(t)$ that is stationary, and we have the various time functions of the process, $X(t,\lambda_i) \equiv x_i(t)$.

It is possible to calculate the mean of the process at an instant t_0 and at another instant t_1. Since the process is stationary both values coincide

$$m_X(t_0) = m_X(t_1) = m_X.$$

Suppose now that the time average of the realization of index i, $x_i(t)$, is denoted as m_i. If the value of this average for any realization coincides with m_X, that is, $m_X = m_i \ \forall i$, the process is an *ergodic process in the mean*. This idea can be extended to other statistics of the random process. Therefore, an ergodic process allows calculating statistical averages over performances of the random process from the time average of a single performance. Statistical averages are replaced by time averages.

Below is the formal definition of ergodicity.

Ergodic process

For a stationary random process $X(t)$ and for any function $g(x)$ two types of averages can be defined

1. For an instant t and different realizations of the process we have a random variable $X(t)$ with probability density function $f_{X(t)}(x)$ independent of t, since the process is stationary. The *statistical expectation* (*statistical average*) of the function $g(X)$ can be computed as

$$E[g(X(t))] = \int_{-\infty}^{\infty} g(x)\ f_{X(t)}(x)\ dx.$$

   This value is independent of t if the process is stationary.

2. If an individual realization of the process is taken, we have a deterministic time function $x(t, \lambda_i)$. We can compute the *time expectation* (or *time average*) of that function

$$< g(x) >_i = \lim_{T\to\infty} \frac{1}{T} \int_{-\frac{T}{2}}^{\frac{T}{2}} g(X(t, \lambda_i))\ dt.$$

   This value $< g(x) >_i$ does not depend on t but in general it depends on the realization, so that for each λ_i in general a different value for $< g(x) >_i$ is possible. If $< g(x) >_i$ is independent of i, that is, it is the same for all i and also

$$< g(x) >_i = E\left[g(X(t))\right],$$

   the process is *ergodic*.

Therefore, a stationary random process $X(t)$, is ergodic if for any function $g(x)$ and for any $\lambda_i \in \Omega$

$$\lim_{T\to\infty} \frac{1}{T} \int_{-\frac{T}{2}}^{\frac{T}{2}} g(X(t, \lambda_i))\ dt = E[g(X(t))].$$

This means that if all the time averages are equal to the statistical averages, the stationary process is ergodic.

Therefore, to estimate the statistics, mean and autocorrelation, of an ergodic stationary process, it is enough to have a single realization of it. From this realization, it is possible to obtain the statistical averages of interest through the time averages. For example, the mean and autocorrelation of the signal can be calculated using the appropriate $g(x)$ functions.

1.3.5 Power and energy of random processes

Two types of deterministic signals had been defined, power signals and energy signals. These definitions can be extended for realizations of random processes. If there is a realization $X(t, \lambda_i)$, which can be denoted as $x_i(t)$, the energy and power of the realization are defined respectively as

$$E_i = \int_{-\infty}^{\infty} |x_i(t)|^2\ dt,$$

and

$$P_i = \lim_{T\to\infty} \frac{1}{T} \int_{-\frac{T}{2}}^{\frac{T}{2}} |x_i(t)|^2 \ dt.$$

For each $\lambda_i \in \Omega$ there is a real number E_i and another P_i that denote energy and power respectively. Consequently, both energy and power are random variables that are denoted as $\mathcal{E}_X$ and $\mathcal{P}_X$ respectively.

Statistical averages can be defined on these random variables that will give an idea of the energy or power of the process. The averages that are usually defined are

- Energy of the random process $X(t)$: E_X.
- Power of the random process $X(t)$: P_X.

These averages are defined as

$$E_X = E[\mathcal{E}_X]$$
$$P_X = E[\mathcal{P}_X]$$

where

$$\mathcal{E}_X = \int_{-\infty}^{\infty} |X(t)|^2 \ dt,$$

and

$$\mathcal{P}_X = \lim_{T\to\infty} \frac{1}{T} \int_{-\frac{T}{2}}^{\frac{T}{2}} |X(t)|^2 \ dt.$$

In this case, as for deterministic signals

- A random process is energy-type if $E_X < \infty$
- A random process is power-type if $0 < P_X < \infty$

Taking these definitions into account, the energy of the process is obtained as

$$\begin{aligned} E_X &= E\left[\int_{-\infty}^{\infty} |X(t)|^2 \ dt\right] \\ &= \int_{-\infty}^{\infty} E\left[|X(t)|^2\right] \ dt \\ &= \int_{-\infty}^{\infty} R_X(t,t) \ dt. \end{aligned}$$

and the power of the process is given by

$$\begin{aligned} P_X &= E\left[\lim_{T\to\infty} \frac{1}{T} \int_{-\frac{T}{2}}^{\frac{T}{2}} |X(t)|^2 \ dt\right] \\ &= \lim_{T\to\infty} \frac{1}{T} \int_{-\frac{T}{2}}^{\frac{T}{2}} E\left[|X(t)|^2\right] \ dt \\ &= \lim_{T\to\infty} \frac{1}{T} \int_{-\frac{T}{2}}^{\frac{T}{2}} R_X(t,t) \ dt. \end{aligned}$$

It can be observed that in both cases, the result depends on the autocorrelation function of the process, evaluated at the same instant of time, that is, $R_X(t,t)$.

For *stationary random processes* $R_X(t,t) = R_X(0)$, is independent of t, and hence

$$P_X = R_X(0),$$

and

$$E_X = \int_{-\infty}^{\infty} R_X(0)\, dt.$$

For the process to be energy-type, $E_X < \infty$ must be fulfilled. This is only possible for the case $R_X(0) = E[X^2(t)] = 0$, which means that for all t and for any realization $X(t) = 0$. Thus, for the case of stationary random processes, only the power-type processes are of practical interest, and in that case the power of the random process can be obtained by evaluating the autocorrelation function, $R_X(\tau)$ at zero ($\tau = 0$).

Finally, if the process is, in addition to being stationary, ergodic, then $\mathcal{P}$ is no longer really a random variable, since all realizations have the same power, which is precisely the power of the process

$$P_i = P_X = R_X(0).$$

1.3.6 Multidimensional (multiple) random processes

As in the case of random variables, it is possible to work with several random (stochastic) processes, defined on the same probability space, simultaneously. Also, when working with communications systems, this comes naturally. For example, when the input of a system is modeled with a random process $X(t)$, we have its output associated with the system, passing through an LTI system. Each realization $X(t,\lambda_i)$ has an associated output

$$X(t,\lambda_i) \to Y(t,\lambda_i) = X(t,\lambda_i) * h(t)$$

This can be interpreted as that for each $\lambda_i \in \Omega$ there are associated two temporary signals $X(t,\lambda_i)$ and $Y(t,\lambda_i)$: these are two random processes defined on the same probability space.

Independence and uncorrelation of random processes

Two random processes $X(t)$ and $Y(t)$ are *independent* if for any pair of time instants, t_1 and t_2, the random variables $X(t_1)$ and $Y(t_2)$ are independent.

Similarly, two random processes $X(t)$ and $Y(t)$ are *uncorrelated* if for any pair of time instants, t_1 and t_2, the random variables $X(t_1)$ and $Y(t_2)$ are uncorrelated.

As for random variables, independence implies uncorrelation, but the reverse is not true: uncorrelation does not imply independence in general.

Cross-correlation

The *cross-correlation function* between two random processes $X(t)$ and $Y(t)$ is defined as

$$R_{X,Y}(t_1,t_2) = E[X(t_1)\, Y^*(t_2)].$$

In general, by the very definition of the cross-correlation function, we have the following relationship between the cross-correlation functions

$$R_{X,Y}(t_1, t_2) = R^*_{Y,X}(t_2, t_1).$$

For real random processes

$$R_{X,Y}(t_1, t_2) = R_{Y,X}(t_2, t_1).$$

Jointly stationarity

Two random processes $X(t)$ and $Y(t)$ are *jointly stationary* (in the wide sense) if the following conditions hold:

a) $X(t)$ and $Y(t)$ are both individually stationary.

b) The cross-correlation function $R_{X,Y}(t_1, t_2)$, depends only on the difference between the two time instants, $\tau = t_1 - t_2$, and can be therefore denoted as

$$R_{X,Y}(t_1, t_2) = R_{X,Y}(t_1 - t_2) = R_{X,Y}(\tau),$$

As in the case of stationarity, this condition is sometimes written using the parameterization $t_1 = t + \tau$ and $t_2 = t$, such that

$$R_{X,Y}(t + \tau, t) = R_{X,Y}(\tau).$$

1.3.7 Random processes in the frequency domain

The usefulness of the frequency representation of signals in systems analysis is well known, since the relationships between signals are in many cases simpler in the frequency domain than in the time domain.

When looking for a suitable frequency representation for random processes, one might first think of trying to define the Fourier transform for each time function of the process, $X(t, \lambda_i)$, and redefining the process by a function that depends on frequency ω instead of t, resulting in a set of frequency domain transforms associated with each possible outcome of the random experiment $X(j\omega, \lambda_i)$, where

$$X(j\omega, \lambda_i) = \mathcal{FT}\{X(t, \lambda_i)\},$$

and $\mathcal{FT}\{\cdot\}$ represents the Fourier transform.

One of the potential problems of this frequency representation of the process is that it is possible that not all of the time functions that are part of the process, $X(t, \lambda_i)$, have a defined Fourier transform.

In this section, we will see how to apply frequency domain analysis techniques to work with random processes, to eventually arrive at an appropriate frequency domain representation for random processes, called the *Power Spectral Density* (PSD).

Power spectral density of random processes

The power spectral density of a random process is a natural extension of the definition of power spectral density for deterministic signals.

To define the power spectral density, the power spectral density of each process signal, $X(t, \lambda_i)$, is first defined and then averaged for all signals. To ensure the existence of the Fourier transform, the truncated functions of duration T seconds are defined

$$X^{[T]}(t, \lambda_i) = \begin{cases} X(t, \lambda_i), & |t| < T/2 \\ 0, & \text{in other case} \end{cases},$$

or by extending the use of the compact notation $x_i(t) \equiv X(t, \lambda_i)$,

$$x_i^{[T]}(t) = \begin{cases} x_i(t), & |t| < T/2 \\ 0, & \text{in other case} \end{cases}.$$

In this way it is guaranteed that the truncated signals are energy signals, as they have a limited duration and therefore their squared modulus is integrable. Therefore, they have a well-defined Fourier transform, which is denoted

$$X_i^{[T]}(j\omega) = \mathcal{FT}\left\{x_i^{[T]}(t)\right\} = \int_{-\infty}^{\infty} x_i^{[T]}(t)\, e^{-j\omega t}\, dt = \int_{-T/2}^{T/2} x_i(t)\, e^{-j\omega t}\, dt.$$

For energy signals, the *energy-spectral density* is $|X_i^{[T]}(j\omega)|^2$. We can then define the *power spectral density* as the energy-spectral density per time unit

$$\frac{\left|X_i^{[T]}(j\omega)\right|^2}{T}.$$

Finally, increasing T arbitrarily, we have the power spectral density of each signal that is part of the random process.

$$S_{X_i}(j\omega) = \lim_{T\to\infty} \frac{|X_i^{[T]}(j\omega)|^2}{T}.$$

For each frequency ω there is a random variable, since there is a value for each possible outcome of the random experiment. It therefore seems logical to define the power spectral density of the process as the average of these random variables

$$S_X(j\omega) \stackrel{def}{=} E\left[\lim_{T\to\infty} \frac{|X^{[T]}(j\omega)|^2}{T}\right] = \lim_{T\to\infty} \frac{E\left[\left|X^{[T]}(j\omega)\right|^2\right]}{T}.$$

This definition allows us to have an intuitive notion of the meaning of the representation, i.e., the mean value of the squared modulus of the frequency responses of all the signals in the process. However, the literal application of this definition to the calculation of the power spectral density is, in most cases, rather involved.

Fortunately, there is a theorem that relates the power spectral density to the autocorrelation function of the random process, which greatly simplifies obtaining these densities.

Wiener-Khinchin Theorem

If for any finite value τ and any interval $\mathcal{A}$, of length $|\tau|$, the autocorrelation of the random process satisfies

$$\left| \int_{\mathcal{A}} R_X(t+\tau, t) dt \right| < \infty,$$

then the power spectral density of $X(t)$ is the Fourier transform of the time average of the autocorrelation function, that is

$$S_X(j\omega) = \mathcal{FT}\left\{ < R_X(t+\tau, t) > \right\},$$

where the time average of the autocorrelation function is

$$< R_X(t+\tau, t) > \stackrel{def}{=} \lim_{T\to\infty} \frac{1}{T} \int_{-T/2}^{T/2} R_X(t+\tau, t) dt.$$

The proof, which can be found for example in [Proakis and Salehi, 2002], p. 179, is reproduced below.

Taking into account that

$$S_X(j\omega) = \lim_{T\to\infty} \frac{E[|X_T(j\omega)|^2]}{T},$$

and that

$$X^{[T]}(j\omega) = \int_{-\frac{T}{2}}^{\frac{T}{2}} X(t)\, e^{-j\omega t}\, dt,$$

introducing this expression, we have

$$\begin{aligned} S_X(j\omega) &= \lim_{T\to\infty} \frac{1}{T} E\left[\int_{-\frac{T}{2}}^{\frac{T}{2}} X(s)\, e^{-j2\pi fs}\, ds \int_{-\frac{T}{2}}^{\frac{T}{2}} X(t)\, e^{+j2\pi ft}\, dt \right] \\ &= \lim_{T\to\infty} \frac{1}{T} \int_{-\frac{T}{2}}^{\frac{T}{2}} \int_{-\frac{T}{2}}^{\frac{T}{2}} R_X(s,t) e^{-j2\pi f(s-t)} dt ds. \end{aligned}$$

Now, the inverse Fourier transform is obtained to show that it is precisely $< R_X(t+\tau, t) >$

$$\begin{aligned} \mathcal{FT}^{-1}\left\{S_X(j\omega)\right\} &= \lim_{T\to\infty} \frac{1}{T} \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{+j\omega\tau} \int_{-\frac{T}{2}}^{\frac{T}{2}} \int_{-\frac{T}{2}}^{\frac{T}{2}} R_X(s,t)\, e^{-j\omega(s-t)}\, dt\, ds\, d\omega \\ &= \lim_{T\to\infty} \frac{1}{T} \frac{1}{2\pi} \int_{-\frac{T}{2}}^{\frac{T}{2}} \int_{-\frac{T}{2}}^{\frac{T}{2}} R_X(s,t)\, ds\, dt \int_{-\infty}^{\infty} e^{j\omega[\tau-(s-t)]}\, d\omega. \end{aligned}$$

Given that the inverse Fourier transform of a constant is a delta

$$\frac{1}{2\pi} \int_{-\infty}^{\infty} e^{j\omega[\tau-(s-t)]}\, d\omega = \delta(\tau - s + t),$$

and including this result in the previous expression, we have

$$\mathcal{FT}^{-1}\left\{S_X(j\omega)\right\} = \lim_{T\to\infty} \frac{1}{T} \int_{-\frac{T}{2}}^{\frac{T}{2}} \int_{-\frac{T}{2}}^{\frac{T}{2}} R_X(s,t)\, \delta(\tau - s + t)\, ds\, dt.$$

Taking into account that

$$\int_{-\frac{T}{2}}^{\frac{T}{2}} R_X(s,t)\delta(\tau-s+t)ds = \begin{cases} R_X(t+\tau,t), & -\frac{T}{2} < t+\tau < \frac{T}{2} \\ 0, & \text{otherwise} \end{cases},$$

then

$$\mathcal{FT}^{-1}\{S_X(j\omega)\} = \lim_{T\to\infty}\frac{1}{T}\int_{-\frac{T}{2}}^{\frac{T}{2}} \left\{ \begin{cases} R_X(t+\tau,t), & -\frac{T}{2} < t+\tau < \frac{T}{2} \\ 0, & \text{otherwise} \end{cases} \right\} dt$$

This expression can be rewritten as

$$\mathcal{FT}^{-1}\{S_X(j\omega)\} = \begin{cases} \lim_{T\to\infty}\frac{1}{T}\int_{-\frac{T}{2}}^{\frac{T}{2}-\tau} R_X(t+\tau,t)\,dt, & \tau > 0 \\ \lim_{T\to\infty}\frac{1}{T}\int_{-\frac{T}{2}-\tau}^{\frac{T}{2}} R_X(t+\tau,t)\,dt, & \tau < 0 \end{cases}$$

or equivalently

$$\mathcal{FT}^{-1}\{S_X(j\omega)\} = \begin{cases} \lim_{T\to\infty}\frac{1}{T}\left[\int_{-\frac{T}{2}}^{\frac{T}{2}} R_X(t+\tau,t)dt - \int_{\frac{T}{2}-\tau}^{\frac{T}{2}} R_X(t+\tau,t)\,dt\right], & \tau > 0 \\ \lim_{T\to\infty}\frac{1}{T}\left[\int_{-\frac{T}{2}}^{\frac{T}{2}} R_X(t+\tau,t)\,dt - \int_{-\frac{T}{2}}^{\frac{T}{2}-\tau} R_X(t+\tau,t)\,dt\right], & \tau < 0 \end{cases}.$$

Since the integral in a segment of length τ is bounded (it is the condition of the statement of the theorem), when $T\to\infty$ the second term is negligible, and therefore

$$S_X(j\omega) = \mathcal{FT}\left\{\lim_{T\to\infty}\frac{1}{T}\int_{-\frac{T}{2}}^{\frac{T}{2}} R_X(t+\tau,t)\,dt\right\},$$

which concludes the proof.

The theorem also includes a couple of corollaries that simplify the calculation of $S_X(j\omega)$ for stationary and cyclostationary random processes.

Corolary 1 If $X(t)$ is a stationary process and $\tau R_X(\tau) < \infty$ for all $\tau < \infty$, then

$$S_X(j\omega) = \mathcal{FT}\{R_X(\tau)\}.$$

The proof is immediate, since $R_X(\tau)$ only depends on τ. In this case

$$< R_X(t+\tau,t) >= R_X(\tau).$$

Corolary 2 If $X(t)$ is cyclostationary and if

$$\left|\int_0^{T_o} R_X(t+\tau,t)dt\right| < \infty,$$

then

$$S_X(j\omega) = \mathcal{FT}\{\widetilde{R}_X(\tau)\},$$

where $\widetilde{R}_X(\tau)$ is the one-period time average the autocorrelation function

$$\widetilde{R}_X(\tau) = \frac{1}{T_0}\int_{-T_0/2}^{T_0/2} R_X(t+\tau, t)\, dt.$$

T_0 is the period of the cyclostationary process. The proof is immediate, because the autocorrelation in periodic.

$$\lim_{T\to\infty}\frac{1}{T}\int_{-T/2}^{T/2} R_X(t+\tau, t)\, dt = \frac{1}{T_0}\int_{-T_0/2}^{T_0/2} R_X(t+\tau, t)\, dt.$$

Apart from these two corollaries, the following properties can be extracted from the Wiener-Khinchin theorem:

(1) The power of the random process can be obtained by integrating the power spectral density,

$$P_X = \frac{1}{2\pi}\int_{-\infty}^{\infty} S_X(j\omega)\, d\omega.$$

This result matches the previous expression obtained from the autocorrelation function

$$P_X = \lim_{T\to\infty}\frac{1}{T}\int_{-\frac{T}{2}}^{\frac{T}{2}} R_X(t,t)\, dt.$$

It must be taken into account that the power can also be calculated in the time domain from the autocorrelation function.

- For stationary processes
$$P_X = R_X(0).$$
- For cyclostationary processes
$$P_X = \tilde{R}_X(0).$$

These conditions are given by the definition of the Fourier transform. For stationary processes, the inverse Fourier transform of the power spectral density is the autocorrelation function

$$R_X(\tau) = \frac{1}{2\pi}\int_{-\infty}^{\infty} S_X(j\omega)\, e^{j\omega\tau}\, d\omega,$$

that evaluated at $\tau = 0$ is

$$R_X(0) = \frac{1}{2\pi}\int_{-\infty}^{\infty} S_X(j\omega)\, d\omega.$$

The same applies in the case of cyclostationary processes to $\tilde{R}_X(0)$.

(2) For *stationary and ergodic* processes, the power spectral density of each signal is equal to the power spectral density of the random process. In this case, the power spectral density of each signal is the Fourier transform of the autocorrelation of that signal, which is defined as

$$R_{x_i(t)}(\tau) = \lim_{T\to\infty}\frac{1}{T}\int_{-T/2}^{T/2} x_i(t)x_i(t-\tau)dt.$$

Therefore

$$S_{x_i(t)}(j\omega) = \mathcal{FT}\left\{R_{x_i(t)}(\tau)\right\}.$$

As the process is ergodic, the statistical average coincides with the temporary one and therefore

$$\lim_{T\to\infty}\frac{1}{T}\int_{-T/2}^{T/2} x_i(t)x_i(t-\tau)dt = R_X(\tau),$$

which means that

$$S_X(j\omega) = \mathcal{FT}\left\{R_X(\tau)\right\} = S_{x_i(t)}(j\omega).$$

(3) The power spectral density was initially defined as

$$\lim_{T\to\infty}\frac{E[|X^{[T]}(j\omega)|^2]}{T}.$$

From this definition it is obvious that $S_X(j\omega)$ is an even function of ω, real and nonnegative: $R_X(\tau)$ is real and even, as we already knew that it happened for stationary processes. Furthermore we have that it is nonnegative and this implies that the autocorrelation is *positive semidefinite*, i.e., to say that

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(t)\ R_X(t-s)\ g(s)\ dt\ ds \geq 0,$$

for any function $g(x)$. So these features of the power spectral density come from the properties of the autocorrelation function of a stationary processes.

1.3.8 Stationary random processes and linear systems

In Section 1.3.6 it has been seen that the output of a time invariant linear system whose input is a random process is itself a random process, as shown in Figure 1.22

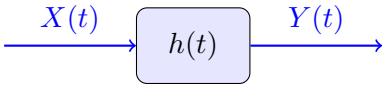


Figure 1.22: A random process is filtered with a time-invariant linear system.

Next, we will analyze the properties of the output process, $Y(t)$, based on the knowledge of the input process, $X(t)$. It is assumed that the input process is real and stationary and that the system is a real linear and time invariant system. In particular, the following questions are considered:

1. Under what conditions is the output process stationary?
2. Under what conditions are the input and output processes jointly stationary?
3. How can the mean and autocorrelation of the output process and the cross-correlation between the input and output processes be obtained?

The following theorem answers these questions.

Theorem: A random process, $X(t)$, is stationary, with mean m_X and autocorrelation function $R_X(\tau)$. The process passes through a linear and time invariant system with impulse response $h(t)$. In this case, *the input and output processes, $X(t)$ and $Y(t)$, are jointly stationary*, and

$$m_Y = m_X \int_{-\infty}^{\infty} h(t)dt,$$

$$R_Y(\tau) = R_X(\tau) * h(\tau) * h(-\tau) = R_X(\tau) * r_h(\tau)$$

$$R_{X,Y}(\tau) = R_X(\tau) * h(-\tau).$$

$$R_{Y,X}(\tau) = R_{X,Y}(-\tau) = R_X(\tau) * h(\tau).$$

In the expression for the autocorrelation function of the output process, $r_h(t)$ denotes the temporal ambiguity function (also called the temporal autocorrelation function) of the channel impulse response

$$r_h(t) = h(t) * h(-t).$$

From the previous expressions it can be seen that the relations

$$R_Y(\tau) = R_{X,Y}(\tau) * h(\tau) = R_{Y,X}(\tau) * h(-\tau).$$

Proof: It follows from the convolution expression, which relates the input and output of a linear system

$$Y(t) = \int_{-\infty}^{\infty} X(s)h(t-s)\, ds$$

Therefore

$$\begin{aligned} m_Y(t) &= E\left[\int_{-\infty}^{\infty} X(s)h(t-s)\, ds\right], \\ &= \int_{-\infty}^{\infty} E[X(s)]h(t-s)\, ds, \\ &= \int_{-\infty}^{\infty} m_X h(t-s)\, ds, \\ &\overset{u=t-s}{=} m_X \int_{-\infty}^{\infty} h(u)\, du. \end{aligned}$$

On the other hand, the cross-correlation between $X(t)$ and $Y(t)$ is

$$\begin{aligned} R_{X,Y}(t_1,t_2) &= E[X(t_1)Y(t_2)], \\ &= E\left[X(t_1)\int_{-\infty}^{\infty} X(s)h(t_2-s)\, ds\right], \\ &= \int_{-\infty}^{\infty} E[X(t_1)X(s)]h(t_2-s)\, ds, \\ &= \int_{-\infty}^{\infty} R_X(t_1-s)h(t_2-s)\, ds, \\ &\overset{u=s-t_2}{=} \int_{-\infty}^{\infty} R_X(t_1-t_2-u)h(-u)\, du, \\ &= \int_{-\infty}^{\infty} R_X(\tau-u)h(-u)\, du, \\ &= R_X(\tau) * h(-\tau). \end{aligned}$$

Similarly, the cross-correlation between $Y(t)$ and $X(t)$ is

$$\begin{aligned} R_{Y,X}(t_1,t_2) &= E[Y(t_1)\ X(t_2)] \\ &= E\left[\int_{-\infty}^{\infty} X(s)\ h(t_1-s)\ ds\ X(t_2)\right] \\ &= \int_{-\infty}^{\infty} E[X(s)\ X(t_2)]\ h(t_2-s)\ ds \\ &= \int_{-\infty}^{\infty} R_X(s-t_2)\ h(t_1-s)\ ds \\ &\overset{u=s-t_2}{=} \int_{-\infty}^{\infty} R_X(u)\ h(t_1-t_2-u)\ du \\ &= \int_{-\infty}^{\infty} R_X(u)\ h(\tau-u)\ du \\ &= R_X(\tau) * h(\tau) \end{aligned}$$

Finally, using the previous result

$$\begin{aligned} R_Y(t_1,t_2) &= E[Y(t_1)Y(t_2)], \\ &= E\left[\left(\int_{-\infty}^{\infty} X(s)h(t_1-s)\ ds\right) Y(t_2)\right], \\ &= \int_{-\infty}^{\infty} E[X(s)Y(t_2)]h(t_1-s)\ ds, \\ &= \int_{-\infty}^{\infty} R_{X,Y}(s-t_2)h(t_1-s)\ ds, \\ &\overset{u=s-t_2}{=} \int_{-\infty}^{\infty} R_{X,Y}(u)h(t_1-t_2-u)\ du, \\ &= R_{X,Y}(\tau) * h(\tau), \\ &= R_X(\tau) * h(-\tau) * h(\tau). \end{aligned}$$

Equivalent expressions in the frequency domain

Previously we have seen what happens for the statistical averages of the output process of a linear and time invariant system when there is a stationary process at its input. Next, the relationships between the input and output process statistics in the frequency domain are analyzed. To do this, it is only necessary to transfer the expressions from the time domain (obtained above) to the frequency domain. The relation is obtained immediately if it is taken into account that

$$\mathcal{FT}\{h(-\tau)\} = H^*(j\omega), \text{ and } \int_{-\infty}^{\infty} h(t)\ dt = H(0).$$

The second expression is obvious if we take into account that

$$H(j\omega) = \int_{-\infty}^{\infty} h(t)\ e^{-j\omega t}\ dt \overset{\omega=0}{=} \int_{-\infty}^{\infty} h(t)dt = H(0).$$

Thus, the relations of the statistics in the frequency domain are

$$m_Y = m_X\ H(0),$$

for the mean, and applying the Fourier transform to the expression relating autocorrelations of the input and output processes, the power spectral density of the output process is

$$S_Y(j\omega) = S_X(j\omega)\ |H(j\omega)|^2.$$

The first equation says that the average is only affected by the continuous response of the system, that is, by the component of $\omega = 0$, $H(0)$. And the second one says that, regarding the power spectral density, the phase of the system is irrelevant and only its module matters.

It is also possible to define a frequency domain relationship for cross-correlation. Define the *cross-spectral density*, $S_{X,Y}(j\omega)$ as

$$S_{X,Y}(j\omega) \stackrel{def}{=} \mathcal{FT}\{R_{X,Y}(\tau)\}.$$

In this case,

$$S_{X,Y}(j\omega) = S_X(j\omega)\ H^*(j\omega),$$

and taking into account that $R_{Y,X}(\tau) = R_{X,Y}(-\tau)$

$$S_{Y,X}(j\omega) = S^*_{X,Y}(j\omega) = S_X(j\omega)\ H(j\omega).$$

It is interesting to note that although the spectral densities of the processes X and Y, $S_X(j\omega)$ and $S_Y(j\omega)$, are non-negative real functions, the cross-spectral densities $S_{X,Y}(j\omega)$ and $S_{Y,X}(j\omega)$ can be, in general, complex functions. Figure 1.23 represents by means of block diagrams the relationship between the different frequency representations.

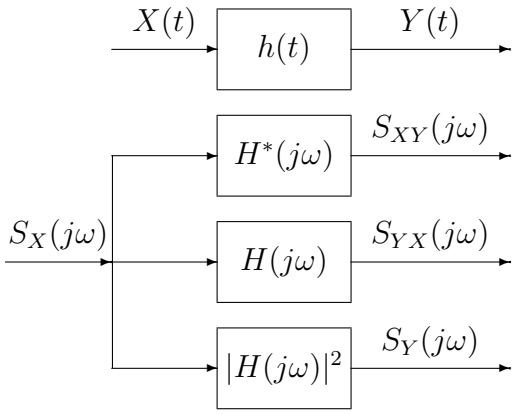


Figure 1.23: Schematic representation of input/output relationships for power spectral densities and cross-spectral densities.

Extension of the previous results to cyclostationary processes

If the random process at the input of a linear and time invariant system is cyclostationary, some of the previous results can be easily extended. In particular it is trivial to do the extension

$$m_Y(t) = m_X(t) \int_{-\infty}^{\infty} h(t)\ dt = m_X(t)\ H(0).$$

From the definition of the power spectral density, it is also trivial to obtain the relation

$$S_Y(j\omega) = S_X(j\omega)\ \left|H(j\omega)\right|^2.$$

Given the relationship of the power spectral density to the average over one period of the autocorrelation function, the relationship is also evident

$$\widetilde{R}_Y(\tau) = \widetilde{R}_X(\tau) * h(\tau) * h(-\tau).$$

The relationship between autocorrelation functions is a bit more complex, but from simple calculations it is possible to arrive at the expression

$$R_Y(t_1, t_2) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} R_X(s, u)\ h(t_1 - s)\ h(t_2 - u)\ ds\ du.$$

1.3.9 Sum of random processes

In practice, the sum of two random processes is often encountered. For example, in the case of a communications system, noise is added to the communications signal.

Suppose we have a random process $Z(t) = X(t) + Y(t)$, where $X(t)$ and $Y(t)$ are jointly stationary. The mean, autocorrelation and power spectral density for $Z(t)$ are obtained below.

The mean of process $Z(t)$ is obtained as follows

$$m_Z = E[Z(t)] = E[X(t) + Y(t)] = E[X(t)] + E[Y(t)] = m_X + m_Y.$$

The autocorrelation of the random process $Z(t)$ is

$$\begin{aligned} R_Z(t+\tau, t) &= E[Z(t+\tau)Z(t)] \\ &= E[(X(t+\tau) + Y(t+\tau))(X(t) + Y(t))] \\ &= E[X(t+\tau)X(t)] + E[X(t+\tau)Y(t)] + E[Y(t+\tau)X(t)] + E[Y(t+\tau)Y(t)] \\ &= R_X(\tau) + R_{X,Y}(\tau) + R_{Y,X}(\tau) + R_Y(\tau). \end{aligned}$$

As can be seen, the process is stationary, since the mean is constant, and the autocorrelation function depends only on the difference between tht time instants. Rearranging terms, the above expression becomes

$$R_Z(\tau) = R_X(\tau) + R_Y(\tau) + R_{X,Y}(\tau) + R_{Y,X}(\tau).$$

The autocorrelation function of the sum process is equal to the sum of the autocorrelation functions of the two processes plus the two crossed-correlations.

The power spectral density is obtained by the Fourier transform of the autocorrelation function. Taking the Fourier transform on both sides of the above expression

$$S_Z(j\omega) = S_X(j\omega) + S_Y(j\omega) + S_{X,Y}(j\omega) + S_{Y,X}(j\omega),$$

and taking into account that $S_{Y,X}(j\omega) = S^*_{X,Y}(j\omega)$

$$S_Z(j\omega) = S_X(j\omega) + S_Y(j\omega) + 2\mathrm{Re}[S_{X,Y}(j\omega)].$$

Thus, the power spectral density is the sum of the power spectral densities of each individual process plus a third term that depends on the cross-spectral density of the two processes.

When the processes are uncorrelated, taking into account that the The relationship between covariance and autocorrelation is,

$$\mathrm{Cov}(X(t+\tau), Y(t)) = E[(X(t+\tau) - m_X)\ (Y(t) - m_Y)] = R_{X,Y}(\tau) - m_X\ m_Y,$$

and that for uncorrelated processes the covariance is null, it is clear that

$$R_{X,Y}(\tau) = m_X m_Y.$$

In this case, if at least one of the two processes has zero mean, $R_{X,Y} = 0$, which implies that $S_{X,Y}(j\omega) = 0$ and therefore

$$R_Z(\tau) = R_X(\tau) + R_Y(\tau),$$

and

$$S_Z(j\omega) = S_X(j\omega) + S_Y(j\omega).$$

In the case of the sum of the signal and noise, typically the noise is uncorrelated with the signal and has zero mean. In this case, this relation can be applied.

1.4 Thermal noise model: white and Gaussian processes

Gaussian processes and white processes play a very important role in communication systems. Because of their relevante, these processes are introduced in this section.

Gaussian processes are important for two reasons:

1. Thermal noise, produced by the random movement of electrons due to thermal agitation, presents a Gaussian distribution. This type of noise is present in any electronic device and is the most important in many communication systems.

   The explanation about why thermal noise is Gaussian can be found in the central limit theorem. Thermal noise is due to the random motion of electrons, and current is the sum of multiple electrons. If it is assumed that each electron behaves independently, we have the sum of a number of random variables *i.i.d.*, with which, in the end, its distribution is Gaussian, $\mathcal{N}(m_X, \frac{\sigma_X^2}{n})$.

2. Gaussian processes provide a good model for some sources of information, so Gaussian processes make their analysis possible.

Therefore, the analysis of Gaussian processes will allow us to analyze the effect of thermal noise and the characteristics of some sources of information.

White processes are also important in the modeling of noise in a communications system, since, as will be seen later, the spectral characteristics of thermal noise are very similar to those of a white process.

1.4.1 Gaussian random processes

A random process $X(t)$ is a *Gaussian process* if for every set of n time instants $\{t_1, t_2, \ldots, t_n\}$ and for any value of n, the n random variables resulting from evaluating the process in those n time instants $\{X(t_i)\}_{i=1}^n$, have a jointly Gaussian distribution.

This means that for any time instant t_0 the random variable $X(t_0)$ is Gaussian and for each pair of instants t_1 and t_2 the random variables $(X(t_1), X(t_2)$ have a jointly Gaussian distribution.

One of the characteristics of Gaussian processes is that their full statistical description depends only on the vector of means $\boldsymbol{\mu}$ and the covariance matrix $\boldsymbol{C}$. Because of this, the following theorem can be formulated.

Theorem: For Gaussian processes, the knowledge of the mean, $m_X(t)$, and of the autocorrelation function, $R_X(t_1, t_2)$, provides a complete statistical description of the process.

This theorem implies that it is not necessary to know the vector $\boldsymbol{\mu}$ or the matrix $\boldsymbol{C}$ for every n and every set of times $\{t_i\}_{i=1}^n$, but it is enough to know $m_X(t)$ and $R_X(t_1, t_2)$. It is always possible to construct $\boldsymbol{\mu}$ and $\boldsymbol{C}$ from these for any set of time instants $\{t_i\}_{i=1}^n$.

- For $X(t_i)$, $\Rightarrow \mu_i = m_X(t_i)$.
- For $(X(t_i), X(t_j))$, $\Rightarrow C_{i,j} = \text{Cov}(X(t_i),X(t_j)) = R_X(t_i, t_j) - m_X(t_i)m_X(t_j)$.

Another advantage of Gaussian processes is their behavior in linear systems. The following theorem describes this behavior which is of great importance.

Theorem: If a Gaussian process $X(t)$ goes over a linear and time invariant (LTI) system, the resulting system, $Y(t)$, is also a Gaussian process.

To prove it, we use one of the properties of jointly Gaussian processes. The aim is to prove that for all n, the random variables $(X(t_1), X(t_2), \ldots, X(t_n))$ are jointly Gaussian. In general, for anny time instant t_i

$$Y(t_i) = \int_{-\infty}^{\infty} X(\tau)h(t_i - \tau)d\tau.$$

This integral can be interpreted as a sum taken to the limit, where $X(t)$ is multiplied by the different values of the impulse response $h(t)$. Specifically, this integral is equal to

$$Y(t_i) = \lim_{N\to\infty} \lim_{\Delta\to 0} \sum_{k=-N}^{N} X(k\Delta)h(t_i - k\Delta).$$

This expression can be seen as a linear combination of a set of jointly Gaussian random variables, $\{X(k\Delta)\}_{k=-N}^N$. Therefore, now

$$\begin{cases} Y(t_1) = \lim\limits_{N\to\infty} \lim\limits_{\Delta\to 0} \sum\limits_{k=-N}^{N} X(k\Delta)h(t_1 - k\Delta) \\ Y(t_2) = \lim\limits_{N\to\infty} \lim\limits_{\Delta\to 0} \sum\limits_{k=-N}^{N} X(k\Delta)h(t_2 - k\Delta) \\ \vdots \\ Y(t_n) = \lim\limits_{N\to\infty} \lim\limits_{\Delta\to 0} \sum\limits_{k=-N}^{N} X(k\Delta)h(t_n - k\Delta) \end{cases}.$$

This expression states the linear combination of the random variables $\{X(k\Delta)\}_{k=-N}^N$, which form $Y(t)$ are jointly Gaussian. And it has been seen in Section 1.2.6 any linear combination of jointly Gaussian random variables forms a set of jointly Gaussian random variables.

This property of Gaussian processes is very important as it means that the type of process that is at the output of a system when the input is Gaussian is known. For any other type of process, in general, it can be very difficult to know this.

All definitions made so far are for Gaussian processes in general. Below are some conditions for stationary processes.

Theorem: For Gaussian processes, stationarity in the strict sense and in the wide sense are equivalent.

This property is due to the fact that Gaussian processes have a complete statistical description that only depends on the mean $m_X(t)$ and the autocorrelation function $R_X(t_1, t_2)$.

Theorem: For Gaussian, stationary, and zero-mean processes, a sufficient condition for the ergodicity of the process $X(t)$ is

$$\int_{-\infty}^{\infty} |R_X(\tau)| d\tau < \infty.$$

This is something that simplifies the ergodicity analysis of this type of process.

Jointly Gaussian random processes

The processes $X(t)$ and $Y(t)$ are *jointly Gaussian*, if for any value of n and m, and any pair of sets of time instants $\{t_1, t_2, \ldots, t_n\}$ and $\{\tau_1, \tau_2, \ldots, \tau_m\}$, the $n + m$ random variables

$$\{X(t_1), X(t_2), \ldots, X(t_n), Y(\tau_1), Y(\tau_2), \ldots, Y(\tau_m)\},$$

they have a jointly Gaussian distribution (of dimension $n + m$).

Jointly Gaussian random processes have a very important property that is stated in the following theorem.

Theorem: For jointly Gaussian processes, uncorrelation and independence are equivalent.

1.4.2 White random processes

A *white process* is a random process that has equal power for all frequencies. The term refers to the analogy with the case of white light, which is made up of the sum of all colors.

By definition, a random proces is white if it has a constant (flat) power spectral density

$$S_X(j\omega) = C.$$

Therefore, the autocorrelation function for a stationary white process is

$$R_X(\tau) = \mathcal{FT}^{-1}\{C\} = C\,\delta(\tau).$$

This means that for any $\tau \neq 0$ $R_X(\tau) = 0$. And that implies that the random variables $X(t_1)$ and $X(t_2)$, $\forall t_1 \neq t_2$, are uncorrelated. If additionally, the process is Gaussian, which as will be seen later is the case of the usual statistical model for thermal noise, this means that the random variables are also independent. Figure 1.24 shows the power spectral density of a white process, and the autocorrelation function for a stationary white process.

From the definition of a white random process, its power is infinite, since

$$P_X = \frac{1}{2\pi}\int_{-\infty}^{\infty} S_X(j\omega)\, d\omega = \frac{1}{2\pi}\int_{-\infty}^{\infty} C\, d\omega = \infty$$

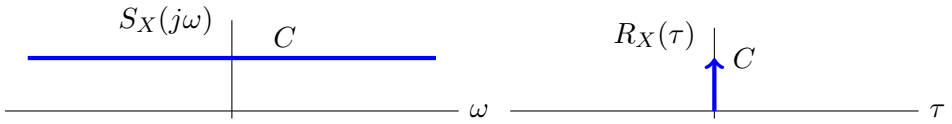


Figure 1.24: Power spectral density of a white process, and autocorrelation function for a stationary white process.

or if it is calculated in the time domain for a stationary random process

$$P_X = R_X(0) = C\,\delta(0) = \infty.$$

Filtering a white process

It has been seen that when a Gaussian process is transmitted over a linear and time-invariant system, the resulting process is still Gaussian. However, the same is not true for the "white" condition. When a white process is transmitted over a linear and time-invariant system, the resulting process is generally not white. Specifically, the power spectral density is determined by the frequency response of the system

$$S_Y(j\omega) = S_X(j\omega)\,|H(j\omega)|^2 = C\,|H(j\omega)|^2.$$

Except in the case of a trivial all-pass filter ($h(t) = \alpha\delta(t)$, or $H(j\omega) = \alpha$, i.e., an amplifier or attenuator), this response will not be constant, so the process $Y(t)$ will not be white.

The autocorrelation function of the filtered process is

$$R_Y(\tau) = R_X(\tau) * h(\tau) * h(-\tau) = R_X(\tau) * r_h(\tau),$$

where $r_h(\tau)$ is the time ambiguity function of $h(\tau)$. Considering the form of $R_X(\tau)$ for a white process,

$$R_Y(\tau) = C\,r_h(\tau).$$

This means that the power of the process is

$$P_Y = R_Y(0) = C\,r_h(0).$$

Taking into account that the value of the time ambiguity function of a signal at zero provides the energy of the signal

$$P_Y = C\,\mathcal{E}\{h(t)\}.$$

Therefore, the power of a filtered white process is no longer infinite, but is related to the energy of the filter.

1.4.3 Thermal noise model

The usual model for thermal noise is that of a stationary, ergodic, white, Gaussian random process. Generally the noise will be denoted as $n(t)$, so the notation $R_n(\tau)$ and $S_n(j\omega)$ will be used to represent the autocorrelation function and the power spectral density of the noise process. In the case of thermal noise, the constant C that defines the value of the power spectral density is usually denoted as $N_0/2$, as shown in the Figure 1.25.

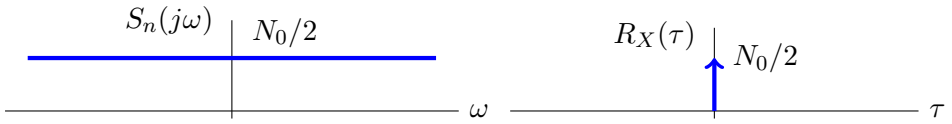


Figure 1.25: Power spectral density and autocorrelation function of a stationary white process modeling thermal noise.

Obviously, no physical process can have infinite power. Therefore, a white process does not exist as such. The importance of white processes in practice is due to thermal noise being modeled as a white process for a wide range of frequencies. Quantum mechanical analysis of thermal noise says that the power spectral density of white noise is

$$S_n(j\omega) = \frac{h\omega}{4\pi(e^{\frac{h\omega}{2\pi kT}} - 1)},$$

where
- h: Planck constant (6.6×10^{-34} Jules $\times$ second).
- k: Boltzmann constant (1.38×10^{-23} Jules/$^{\text{o}}$Kelvin).
- T: Temperature in Kelvin degrees.
- ω: Angle frequency, in radians/s (2π times the linear frequency).

This power spectral density has its maximum at $\omega = 0$, where it takes the value $kT/2$, and tends to zero as ω tends to infinity. However, the descent is very slow. For example, at $T = 290$ Kelvin degrees, $S_n(j\omega)$ drops to about 90% of its maximum for $\omega \approx 2\pi$ (2×10^{12}) rad/s, which is above the frequencies commonly used by communications systems. Figure 1.26 plots $S_n(j\omega)$, with a zoom (right figure) in the range of 200 GHz.

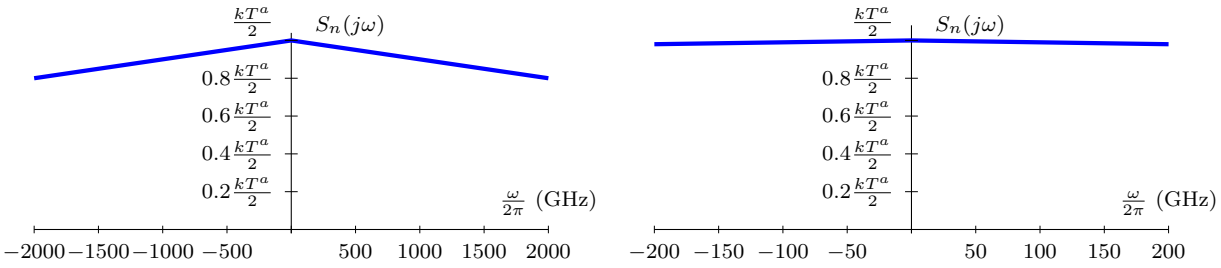


Figure 1.26: Power spectral density of thermal noise.

For this reason, thermal noise is modeled as a white process with $C = \frac{N_0}{2}$, with $N_0 = kT$ Watts/Hz.

From now on, the model for the thermal noise is a random process with the following characteristics:

- Stationary.
- Ergodic.
- Zero mean, $m_n = 0$.
- Autocorrelation function

$$R_n(\tau) = \frac{N_0}{2}\,\delta(\tau).$$

- Power spectral density

$$S_n(j\omega) = \frac{N_0}{2}.$$

NOTE: It is easy to see that a Gaussian stationary process with such an autocorrelation function is ergodic, since

$$\int_{\infty}^{\infty} |R_n(\tau)| \; d\tau = \frac{N_0}{2}.$$

1.4.4 Filtered noise and noise equivalent bandwidth

As seen above, the power of a white random process is infinite. In the case of thermal noise, although it is not infinite, it is relatively high. This power is limited by filtering. Denoting by $Z(t)$ the random process resulting from filtering the thermal noise process with a linear time-invariant filter $h(t)$, the power spectral density of this process is

$$S_Z(j\omega) = S_n(j\omega) \; |H(j\omega)|^2 = \frac{N_0}{2} \; |H(j\omega)|^2 .$$

The power of $Z(t)$ can be obtained by integrating $S_Z(j\omega)$

$$P_Z = \frac{1}{2\pi} \int_{-\infty}^{\infty} S_Z(j\omega) \; d\omega = \frac{N_0}{2} \underbrace{\frac{1}{2\pi} \int_{-\infty}^{\infty} |H(j\omega)|^2 \; d\omega}_{\mathcal{E}\{h(t)\}} = \frac{N_0}{2} \, \mathcal{E}\{h(t)\}.$$

The frequency responses of ideal filters with bandwidth B Hz (or $W = 2\pi B$ rad/s), either low-pass filters or band-pass filters with central frequency f_c Hz (or $\omega_c = 2\pi f_c$ rad/s), are shown in Figure 1.27.

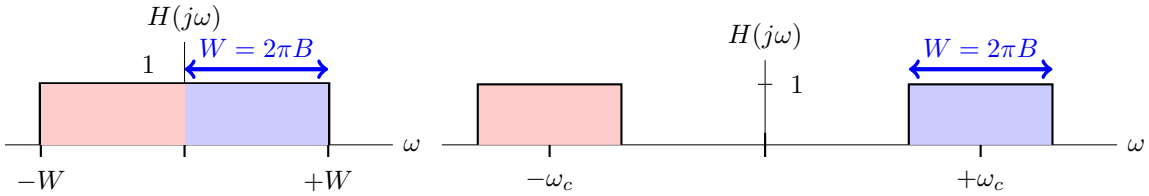


Figure 1.27: Frequency response of ideal filters, low pass and band pass.

For both low pass and bandpass ideal filters, it is easy to check that

$$\mathcal{E}\{h(t)\} = \frac{1}{2\pi} \int_{-\infty}^{\infty} |H(j\omega)|^2 \; d\omega = 2B.$$

Therefore the power of the filtered noise $Z(t)$ is

$$P_Z = N_0 \; B.$$

If the ideal filters do not have unit gain, but power gain G, their frequency response is that of Figure 1.28.

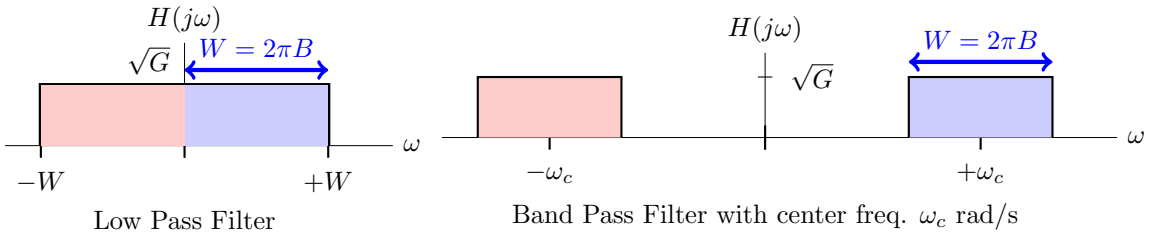


Figure 1.28: Frequency response of ideal filters, low pass and band pass, with a power gain G (voltage gain $\sqrt{G}$).

In that case, the energy of the filters is

$$\mathcal{E}\{h(t)\} = \frac{1}{2\pi}\int_{-\infty}^{\infty} |H(j\omega)|^2 \, d\omega = 2BG,$$

and therefore the power of the filtered noise is

$$P_Z = N_0 \, B \, G.$$

Both with unit gain and with an arbitrary power gain G, the power of the filtered noise is calculated very easily from the previous expressions.

When non-ideal filters are used, it is in many cases impractical for users of a certain system to have to measure the energy of the filter (theoretically, by integrating the squared modulus of the filter response, or using equipment). It would be much more convenient to be able to apply an expression similar to the one that is used with ideal filters. For this reason, the so-called *noise-equivalent bandwidth* of a system is used. The noise-equivalent bandwidth, which is denoted as B_{eq}, is used to obtain the noise power as

$$P_Z = N_0 \, B_{eq} \, G_{eq},$$

where $G_{eq} = H_{max}^2$, and H_{max} is the maximum value of $H(j\omega)$. Thus, G_{eq} denotes the equivalent gain of the filter. Given the equivalent noise bandwidth of the filter, B_{eq}, and its equivalent gain G_{eq}, it is very simple Calculate the noise power at the output of the system. Usually, manufacturers measure these values and publish them in the data sheets so that it can be used by users.

Comparing the expressions for the filtered noise power given by B_{eq} and by $\mathcal{E}\{h(t)\}$,

$$\frac{N_2}{2}\,\mathcal{E}\{h(t)\} = N_0 \, B_{eq} \, G_{eq}$$

it is straightforward to obtain the noise-equivalent bandwidth as

$$B_{eq} = \frac{\mathcal{E}\{h(t)\}}{2\,G_{eq}} = \frac{\dfrac{1}{2\pi}\displaystyle\int_{-\infty}^{\infty} |H(j\omega)|^2 \, d\omega}{2\,G_{eq}} \text{ Hz}.$$

Bearing in mind that the frequency response of real systems is symmetric with respect to the origin, it follows that

$$\int_{-\infty}^{\infty} |H(j\omega)|^2 \, d\omega = 2\int_{0}^{\infty} |H(j\omega)|^2 \, d\omega,$$

and the equivalent noise bandwidth can also be calculated as

$$B_{eq} = \frac{\frac{1}{2\pi}\int_0^\infty |H(j\omega)|^2\, d\omega}{G_{eq}} \text{ Hz.}$$

The interpretation of the noise-equivalent bandwidth would be that an ideal low-pass filter, of amplitude H_{max} and bandwidth B_{eq} Hz, allows to pass the same noise power than the characterized system. Figure 1.29 illustrates this interpretation.

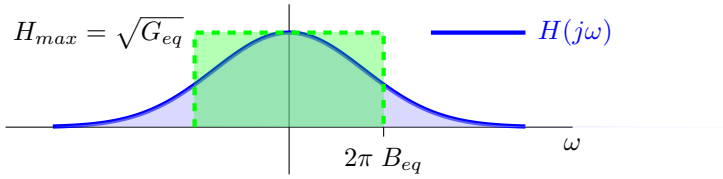


Figure 1.29: Illustration of the meaning of the equivalent noise bandwidth of a system.

Example

Calculate the noise-equivalent bandwidth of a filter with the following frequency response

$$|H(j\omega)| = \begin{cases} \sqrt{1+\frac{\omega}{W}}, & \text{if } -W \le \omega < 0 \\ \sqrt{1-\frac{\omega}{W}}, & \text{if } 0 \le \omega \le W \\ 0, & \text{in other case,} \end{cases}$$

where W is the bandwidth in rad/s, i.e., $W = 2\pi B$, where B is the bandwidth in Hz.

In this case, the squared modulus of the frequency response of the filter is a triangle between -W and W,

$$|H(j\omega)|^2 = \Lambda\left(\frac{\omega}{2W}\right) = \begin{cases} 1+\frac{\omega}{W}, & \text{if } -W \le \omega < 0 \\ 1-\frac{\omega}{W}, & \text{if } 0 \le \omega \le W \\ 0, & \text{in other case.} \end{cases}$$

It is straightforward to obtain $H_{max} = \max |H(j\omega)|^2 = 1$, the value for $\omega = 0$. On the other hand, taking into account the symmetry of the filter response

$$\begin{aligned} \frac{1}{2\pi}\int_{-\infty}^{\infty} |H(j\omega)|^2\, d\omega &= \frac{1}{\pi}\int_0^\infty |H(j\omega)|^2\, d\omega \\ &= \frac{1}{\pi}\int_0^W 1-\frac{\omega}{W}\, d\omega \\ &= \frac{1}{\pi}\times\frac{W}{2} = B. \end{aligned}$$

Therefore,

$$B_{eq} = \frac{B}{2\times 1} = \frac{B}{2} \text{ Hz.}$$

Example

Calculate the noise-equivalent bandwidth of a low pass RC filter.

The frequency response of an RC filter is

$$H(j\omega) = \frac{1}{1+j\omega\tau},$$

where the constant τ is equal to $\tau = RC$. The module of this answer is

$$|H(j\omega)| = \frac{1}{\sqrt{1+\omega^2\tau^2}}.$$

In this case it is clear that $H_{max} = 1$, for $\omega = 0$. Regarding the energy of the filter

$$\begin{aligned}\frac{1}{2\pi}\int_{-\infty}^{\infty}|H(j\omega)|^2\,d\omega = \frac{1}{\pi}\int_0^{\infty}|H(j\omega)|^2\,d\omega = \frac{1}{\pi}\int_0^{\infty}\frac{1}{1+\omega^2\tau^2}\,d\omega \\ \overset{u=\omega\tau}{=} \frac{1}{\pi}\int_0^{\infty}\frac{1}{1+u^2}\,\frac{du}{\tau} \\ = \frac{1}{\pi\tau}\int_0^{\infty}\frac{1}{1+u^2}\,du = \frac{1}{\pi\tau}\underbrace{\text{arctg}(u)\Big|_0^{\infty}}_{\frac{\pi}{2}} = \frac{1}{2\tau}.\end{aligned}$$

Finally,

$$B_{eq} = \frac{\frac{1}{2\tau}}{2\times 1} = \frac{1}{4\tau} = \frac{1}{4RC}\text{ Hz}.$$

1.4.5 Signal to noise ratio

We have already seen how to calculate the power of random processes. A particular case is the case in which there is the signal plus a noise component and both signals are modeled with different processes. In this case it is useful to calculate the signal to noise ratio

$$\frac{S}{N} = \frac{\text{Power of the signal}}{\text{Power of the noise}}.$$

In many cases this ratio is expressed in decibels.

$$\frac{S}{N}(\text{dB}) = 10\log_{10}\frac{S}{N}.$$

If a signal is modeled with a process $X(t)$, and thermal noise is added to it, as has been said before, at the input of any receiver a filter is used to limit the noise power. The filtered signal is denoted as $Y(t)$ and the filtered noise is $Z(t)$, as illustrated in Figure 1.30.

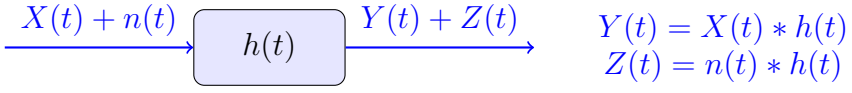


Figure 1.30: Filtering of a signal to which thermal noise has been added.

In this situation, it is possible to define two signal-to-noise rations: one at the input of the filter, and the other at the output of the filter, although as we will see now only one of them really makes sense. Before filtering, the signal to noise ratio it is

$$\left.\frac{S}{N}\right|_{in} = \frac{P_X}{P_n}, \quad \left.\frac{S}{N}\right|_{in}(\text{dB}) = 10\log_{10}\frac{P_X}{P_n}\text{ dB},$$

and after filtering it is

$$\left.\frac{S}{N}\right|_{out} = \frac{P_Y}{P_Z}, \quad \left.\frac{S}{N}\right|_{out} (\text{dB}) = 10 \log_{10} \frac{P_Y}{P_Z} \text{ dB}.$$

The power of the signal at the input of the filter can be calculated in several ways, for example by integrating its power spectral density

$$P_X = \frac{1}{2\pi} \int_{-\infty}^{\infty} S_X(j\omega)\, d\omega.$$

The power of thermal noise, as previously seen, is infinite.

$$P_n = \frac{1}{2\pi} \int_{-\infty}^{\infty} S_n(j\omega)\, d\omega = \int_{-\infty}^{\infty} \frac{N_0}{2}\, d\omega = \infty,$$

so the signal-to-noise ratio at the input is zero, or equivalently, $-\infty$ dB

$$\left.\frac{S}{N}\right|_{in} = \frac{P_X}{P_n} = \frac{P_X}{\infty} = 0, \quad \left.\frac{S}{N}\right|_{in} (\text{dB}) = 10 \log_{10} \frac{P_X}{P_n} = -\infty \text{ dB}.$$

This result makes evident the need to filter to limit the power of thermal noise. At the output of the filter, the power of the signal can be obtained by integrating its power spectral density

$$P_Y = \frac{1}{2\pi} \int_{-\infty}^{\infty} S_Y(j\omega)\, d\omega = \frac{1}{2\pi} \int_{-\infty}^{\infty} S_X(j\omega)\, |H(j\omega)|^2\, d\omega.$$

For noise, depending on the type of filter, the power can be calculated in several ways. If it is an ideal filter with bandwidth B Hz

$$P_Z = N_0\, B,$$

being $N_0 = K\, T$. If the ideal filter has a power gain G then

$$P_Z = N_0\, B\, G.$$

On the other hand, if we have a non-ideal filter whose noise-equivalent bandwidth and its (equivalent) gain are known, the power of the filtered noise is

$$P_Z = N_0\, B_{eq}\, G_{eq}.$$

Finally, if we have a non-ideal filter and its noise-equivalent bandwidth is unknown, the noise power can be calculated (if its response is known) as

$$P_Z = \frac{N_0}{2}\, \frac{1}{2\pi} \int_{-\infty}^{\infty} |H(j\omega)|^2\, d\omega = \frac{N_0}{2} \int_{-\infty}^{\infty} |h(t)|^2\, dt = \frac{N_0}{2}\, \mathcal{E}\{h(t)\}.$$

In any case, this power P_Z will be finite, which gives rise to a non-zero signal-to-noise ratio.

1.5 Sampling of band-limited random processes

The definition of a band-limited random process is the natural extension of the definition of a band-limited signal. A random process of bandwidth B Hz has the property

$$S_X(j\omega) = 0, \quad \forall |\omega| \geq 2\pi\, B.$$

As already seen, for band-limited signals the sampling theorem allows a signal to be sampled without information loss. This theorem says that to be able to perfectly reconstruct the original signal, the sampling frequency must be at least twice the bandwidth of the signal

$$f_m \geq 2B \rightarrow T_m = \frac{1}{f_m} \leq \frac{1}{2B}.$$

In this case the signal is reconstructed as

$$x(t) = 2BT_m \sum_{k=-\infty}^{\infty} x(kT_m)\ \text{sinc}(2B(t-kT_m)).$$

For $T_m = \frac{1}{2B}$, this expression can be simplified as

$$x(t) = \sum_{k=-\infty}^{\infty} x(kT_m)\ \text{sinc}(2B(t-kT_m)).$$

It makes sense to think that the sampling theorem can be extended to random processes. The following theorem justifies this intuition.

Theorem: If $X(t)$ is a band-limited process, with $S_X(j\omega) = 0$ for $\omega \geq W = 2\pi B$, taking a sampling interval $T_m = \frac{1}{2B} = \frac{\pi}{W}$

$$E\left[\left(X(t) - \sum_{k=-\infty}^{\infty} X(kT_m)\ \text{sinc}(2B(t-kT_m))\right)^2\right] = 0.$$

To prove this, we expand this expression

$$\begin{aligned}
&E\left[\left(X(t) - \sum_{k=-\infty}^{\infty} X(kT_m)\ \text{sinc}(2B(t-kT_m))\right)^2\right] \\
=&E[X^2(t)] - 2\sum_{k=-\infty}^{\infty} E[X(t)X(kT_m)]\ \text{sinc}(2B(t-kT_m)) \\
&+ \sum_{k=-\infty}^{\infty}\sum_{u=-\infty}^{\infty} E[X(kT_m)X(uT_m)]\ \text{sinc}(2B(t-kT_m))\,\text{sinc}(2B(t-uT_m)) \\
=&R_X(0) - 2\sum_{k=-\infty}^{\infty} R_X(t-kT_m)\ \text{sinc}(2B(t-kT_m)) \\
&+ \sum_{k=-\infty}^{\infty}\sum_{u=-\infty}^{\infty} R_X((k-u)T_m)\ \text{sinc}(2B(t-kT_m))\,\text{sinc}(2B(t-uT_m)).
\end{aligned}$$

For the last term, we can make the change of variable $m = u - k$ and it remains

$$\begin{aligned}
&\sum_{k=-\infty}^{\infty}\sum_{m=-\infty}^{\infty} R_X(mT_m)\ \text{sinc}(2B(t-kT_m))\,\text{sinc}(2B(t-kT_m-mT_m)) \\
&\qquad = \sum_{k=-\infty}^{\infty} \text{sinc}(2B(t-kT_m)) \sum_{m=-\infty}^{\infty} R_X(mT_m)\ \text{sinc}(2B(t-kT_m-mT_m)),
\end{aligned}$$

where the property $R_X(-mT_m) = R_X(mT_m)$ has been used.

Since $X(t)$ is band limited, its autocorrelation is also band limited, so that

$$R_X(t) = \sum_{k=-\infty}^{\infty} R_X(kT_m)\ \text{sinc}(2B(t - kT_m)),$$

and therefore

$$\sum_{m=-\infty}^{\infty} R_X(mT_m)\ \text{sinc}(2B(t - kT_m - mT_m)) = R_X(t - kT_m).$$

Substituting this expression, we get

$$E\left[\left(X(t) - \sum_{k=-\infty}^{\infty} X(kT_m)\ \text{sinc}(2B(t - kT_m))\right)^2\right]$$
$$= R_X(0) - \sum_{k=-\infty}^{\infty} R_X(t - kT_m)\ \text{sinc}(2B(t - kT_m)).$$

And it can be checked that the last term is $R_X(0)$, which completes the proof.

Since it is the expectation of the square error that vanishes, in this case it is said that the sampling theorem holds in the mean square sense, or that $X(t)$ is equal in the mean square sense (MSS) to the expression of the sampling theorem for $X(kT_m)$

$$X(t) \overset{MSS}{=} \sum_{k=-\infty}^{\infty} X(kT_m)\ \text{sinc}(2B(t - kT_m)).$$

Another interesting property is that the samples of the random process are only uncorrelated if the process has a constant (flat) power spectral density in the band, that is, if

$$S_X(j\omega) = \begin{cases} C, & |\omega| < W \\ 0, & \text{in other case} \end{cases}.$$

Chapter 2

Analog Modulations

This chapter introduces the most commonly used analog modulations: amplitude modulations and phase modulations. Their main characteristics in both the time and frequency domains are presented, and the effect of noise on each modulation is analyzed.

2.1 Introduction to the concept of modulation

An analog signal is a continuous-time signal with a continuous range of possible values (continuous-time continuous signal). The output of most information sources corresponds to this type of signal. Voice and video are two important examples of analog sources. Figure 2.1 shows a voice signal. With this type of signal, the information is stored in the waveform of the signal itself. Therefore, an analog communication system must attempt to transmit this waveform as faithfully as possible to the system endpoint.

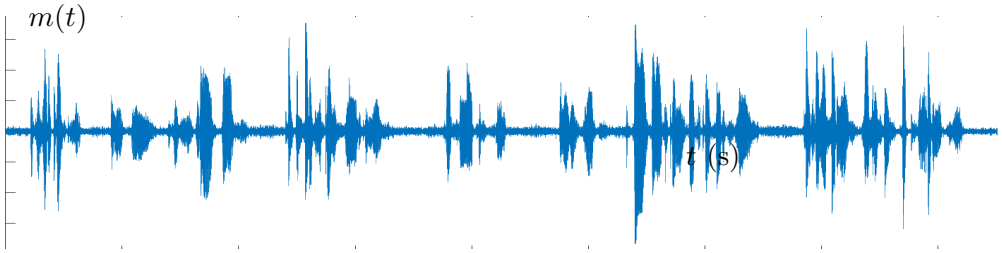


Figure 2.1: Example of a voice signal.

The general trend consists of sampling and quantizing the signals (analog-to-digital conversion, or A/D conversion) and transmitting them through a digital communications system, in order to reconstruct them at the receiver (digital-to-analog conversion, or D/A conversion). However, some analog communication systems still exist today, and they are still usefull in some applications with specific requirements. Therefore, it is necessary to analyze these systems.

The transmission of analog signals can be carried out basically in two different ways:

1. Baseband or unmodulated transmission: The information signal is transmitted directly, without any modification.

2. Modulated transmission: The information signal is modified, the spectrum of the signal is shifted, it will be centered around a certain center or carrier frequency ω_c rad/s. The shape of the spectrum or the bandwidth of the signal may or may not be modified, as shown in Figure 2.2 in examples A and B, respectively.

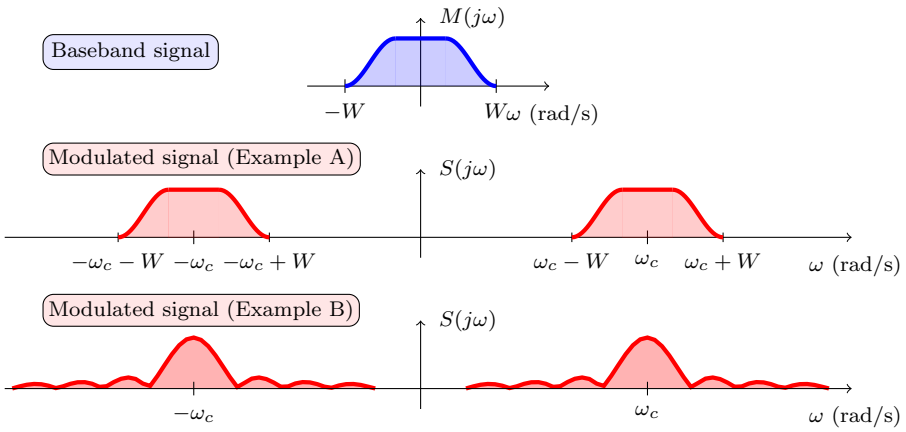


Figure 2.2: Different modes of transmission of analog information (frequency domain view).

The most frequent option is the transmission of a modulated signal. The process of modifying the information signal to produce another signal with different spectral characteristics but containing the same information is generally called modulation. The new generated signal is called the modulated signal, while the information signal is called the modulating signal. Usually, the transmission is done by modulating a signal called carrier. In general, the analog information signal is stored in the amplitude, frequency, or phase of a sinusoidal carrier. This is called amplitude, frequency, or phase modulation.

The modulation of an analog signal serves one of the following purposes:

1. To shift the spectrum of the original signal to adapt it to the channel characteristics, that is, to bring it into the region of the spectrum (frequency band) where the channel behaves better (ideally, in such a way that a distortionless transmission is produced; in practice, where the distortion is as small as possible).

2. To expand the bandwidth of the transmitted signal to reduce the effect of the noise during the transmission.

3. To accommodate the simultaneous transmission of different signals or sources of information on the same channel, which is called *multiplexing* or *multiple-access*, depending on the scenario. The spectrum of the different signals can be modulated to shift their spectrum into non-overlapping frequency bands. This type of simultaneous transmission is called Frequency Division Multiplexing (FDM), or Frequency Division Multiple-Access (FDMA). The basic idea of an FDM system is illustrated in Figure 2.3.

   In this example, there are three signals that share the same frequency band. If the three signals are transmitted directly over the same medium, it would not be possible to separate them later. On the other hand, if the frequency range of each signal is shifted by modulation and the spectrum of each of them is moved to a frequency band in such a way that there is no overlap between the spectrum of the three modulated signals, the simultaneous transmission of the three signals is possible. At the receiver, after filtering each of the three signals, and

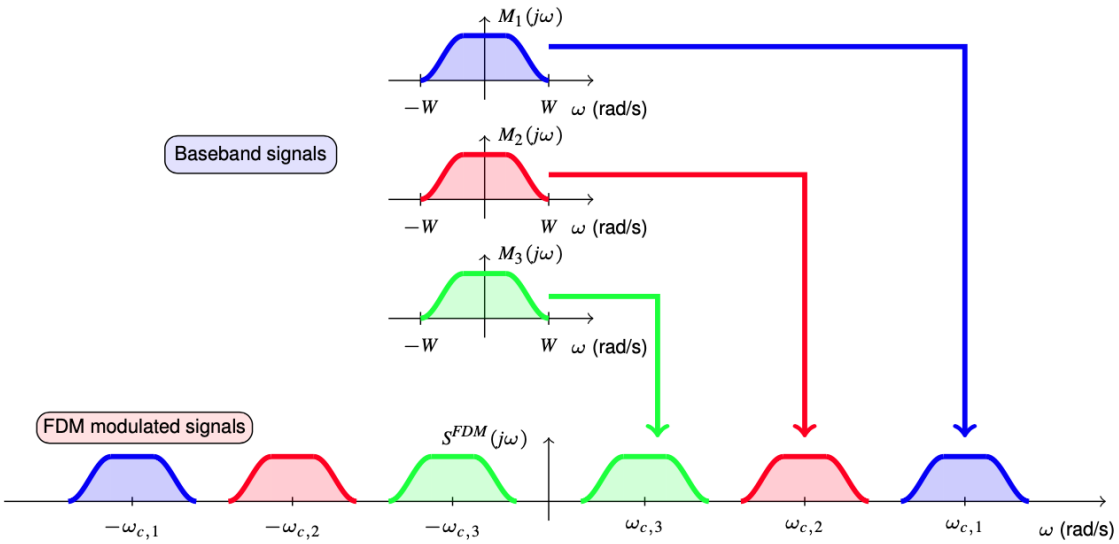


Figure 2.3: Simple example of multiplexing three signals by frequency division.

restoring the spectrum of each signal to its original frequency range, each of the three signals can be recovered separately (demultiplexing), as illustrated in Figure 2.4. A classic example of the application of FDM is commercial radio broadcasting, where every radio station has assigned a specific frequency band. All stations transmit simultaneously in their non-overlapping frequency bands, and a user can tune a given station by selecting its corresponding reference frequency, as shown in Figure 2.5, to filter and demodulate the signal that is transmitted in its assigned frequency band.

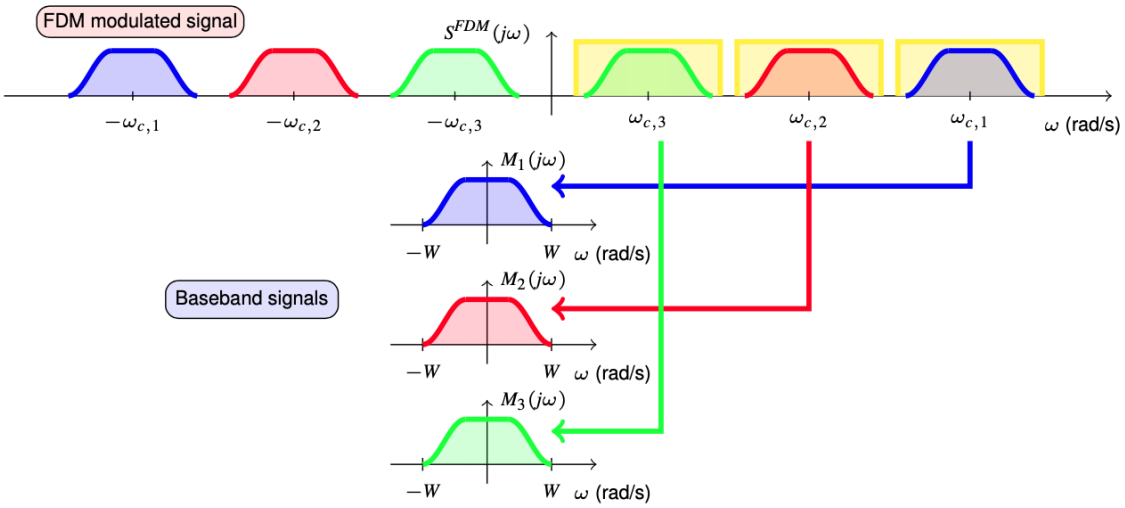


Figure 2.4: Demultiplexing three signals in an FDM transmission scheme.

The first and third purposes are achieved with the three types of modulations mentioned above (amplitude, frequency or phase), while the second is only produced with the so-called *angle modulations*, which are phase and frequency modulations.

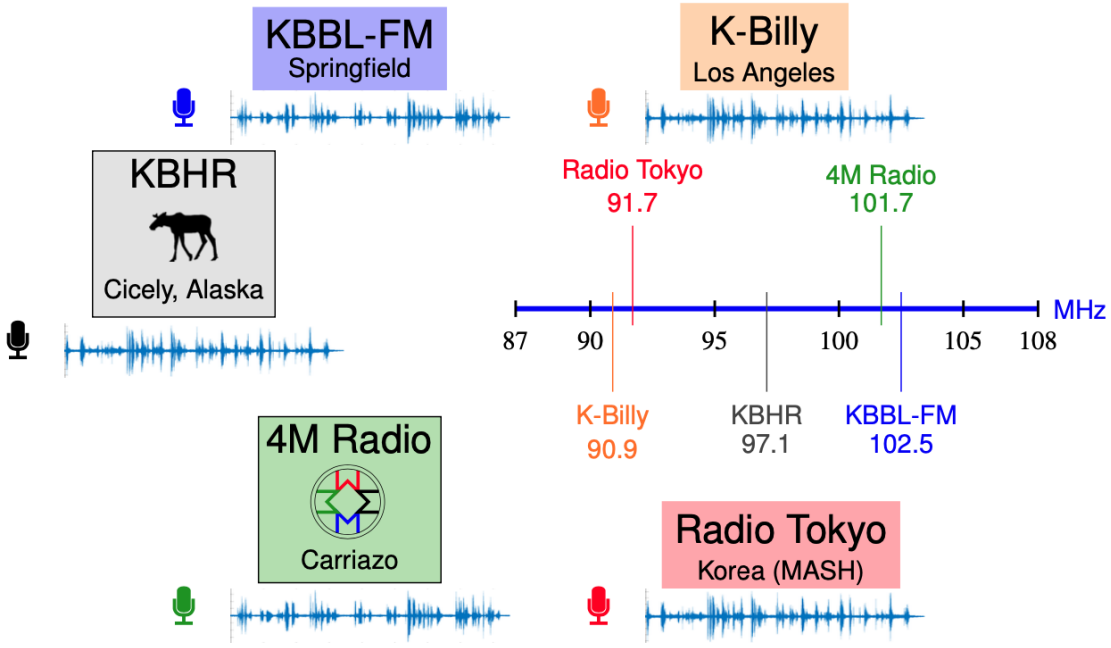


Figure 2.5: Commercial radio broadcasting as a classic example of FDM.

2.1.1 Basic notation and modulating signal models

In this section we will establish some aspects of the notation that will be used throughout the chapter.

The analog information signal to be transmitted, or modulating signal, will be denoted alternatively as $m(t)$ or $M(t)$: in the presentation and the analysis of the different modulations, in some cases a deterministic signal $m(t)$ is assumed; in other cases, it is considered a random signal whose statistical parameters are known, which is modeled with a random process $M(t)$.

In the first case, a deterministic modulating signal, the following characteristics are assumed:

1. It is a baseband signal with bandwidth B Hz or $W = 2\pi B$ rad/s; that is, its Fourier transform, $M(j\omega)$, is zero, $M(j\omega) = 0$ for $|\omega| > W$ rad/s.
2. It is a power-type signal with power

$$P_m = \lim_{T\to\infty} \frac{1}{T} \int_{T/2}^{T/2} |m(t)|^2 dt.$$

When considering a random modulating signal, this signal will be characterized by a stationary random process $M(t)$ with the following characteristics:

1. It is a Wide Sense Stationary (WSS) random process.
2. It has a known autocorrelation function $R_M(\tau)$.
3. The power spectral density is $S_M(j\omega)$ (related to the autocorrelation function via the Fourier transform).

4. It is a baseband and band-limited random process, with bandwidth B Hz or $W = 2\pi B$ rad/s; that is, $S_M(j\omega) = 0$ for $|\omega| > W$ rad/s.
5. It is a power-type process with power P_M, which can be obtained from the above functions as
$$P_M = R_M(0) = \frac{1}{2\pi}\int_{-\infty}^{\infty} S_M(j\omega)\, d\omega.$$

The signal is transmitted through a communications channel by storing it on a sinosoidal carrier
$$c(t) = A_c \cos(2\pi f_c t + \phi_c) = A_c \cos(\omega_c t + \phi_c),$$
where A_c is the amplitude, f_c is the frequency in Hz (ω_c is the frequency in rad/s) and ϕ_c is the phase of the carrier signal.

The modulating signal, $m(t)$, is said to modulate the carrier signal in amplitude, frequency, or phase, if the amplitude, frequency, or phase depends on $m(t)$. In any case, the effect of the modulation is *to convert the baseband modulating signal, $m(t)$, into a band-pass signal whose spectrum is around the frequency of the carrier signal, f_c Hz.* Summarizing, there are the following types of analog modulations:

1. Amplitude Modulations (AM)
   The carrier amplitude varies in time as a function of the modulating signal.
   $$A_c \rightarrow A_c(t) = f(m(t)).$$
2. Angle modulations
   The angle value of the carrier varies in time as a function of the modulating signal
   a) Phase Modulation (PM)
      The phase of the carrier varies in time as a function of the modulating signal.
      $$\phi_c \rightarrow \phi_c(t) = f(m(t)).$$
   b) Frequency Modulation (FM)
      The instantaneous frequency of the carrier varies in time as a function of the modulating signal
      $$f_i(t) = f_c \rightarrow f_i(t) = f(m(t)).$$
      $f_i(t)$: instantaneous frequency of the carrier signal at instant t.

2.2 Amplitude Modulations (AM)

In an amplitude modulation the modulating signal $m(t)$ is stored on the amplitude of the sinusoidal carrier $c(t)$, which instead of having a constant value A_c will vary in time as a function of $m(t)$. There are different variants of amplitude modulations. In this chapter we will analyze the following:

1. AM: Conventional AM modulation (or double sideband AM modulation with carrier).
2. DSB: Double SideBand modulation (without carrier).
3. SSB: Single SideBand modulation.
4. VSB: Vestigial SideBand modulation.

2.2.1 Conventional AM

A conventional AM signal consists of the sum of two components: the carrier signal and a double sidedband signal, which consists of the product between the modulating signal and the carrier. The general analytical expression for the modulated signal is therefore

$$s(t) = \underbrace{A_c \cos(\omega_c t + \phi_c)}_{\text{Carrier } c(t)} + \underbrace{m(t) \times A_c \cos(\omega_c t + \phi_c)}_{\text{Double SideBand (DSB): } m(t)\times c(t)},$$

which can also be written as

$$s(t) = A_c \left[1 + m(t)\right] \cos(\omega_c t + \phi_c). \tag{2.1}$$

The new varying amplitude of the sinusoidal carrier is $A_c \left[1 + m(t)\right]$. In most cases it is useful to impose the constraint $|m(t)| \leq 1$ such that this amplitude $A_c[1 + m(t)]$ is always positive, since in that case the modulating signal $m(t)$ is directly reflected in the envelope of the modulated signal, and the demodulation of the signal will be easier. If for some value of t the modulating signal is $m(t) < -1$, the modulated signal is *overmodulated*, and the necessary process for demodulation becomes more complicated, since in the intervals in which $m(t) < -1$ the envelope becomes proportional to $-m(t)$ (see Figure 2.6).

To avoid overmodulation $m(t)$ can be scaled so that the amplitude is always less than unity. The most common way to do it is by introducing a normalization and the so-called modulation index.

Taking into account the dynamic range of the modulating signal, and assuming that $-C_M \leq m(t) \leq +C_M$, the normalized modulating signal $m_n(t)$ is defined as

$$m_n(t) = \frac{m(t)}{\max |m(t)|} = \frac{m(t)}{C_M}.$$

From this normalized signal, the modulating signal with modulation index a is defined as

$$m_a(t) = a \times m_n(t).$$

The scale factor a is called *modulation index*, and it is a positive value. Now, the modulated signal AM with a modulation index a is defined by replacing in the expression (2.1) the modulating signal $m(t)$ by the modulating signal with modulation index a, $m_a(t)$, i.e.

$$s(t) = c(t) + m_a(t) \times c(t) = A_c \left[1 + m_a(t)\right] \cos(\omega_c t + \phi_c).$$

To avoid overmodulation, the modulation index is in the range

$$0 < a \leq 1.$$

Figure 2.7 shows an example of the waveform for a modulation index $a = \frac{1}{2}$. You can see how the information signal is printed in the envelope of the modulated signal, and how its amplitude varies between $A_c(1 - a)$ and $A_c(1 + a)$, in this case between $\frac{A_c}{2}$ and $\frac{3A_c}{2}$.

If the modulation index is modified for the same modulating signal, and $a = \frac{3}{4}$ is used, the resulting modulated signal is the one shown in Figure 2.8. Again the modulating signal (information) is printed in the signal envelope, but now the amplitude of the signal varies in a larger range,

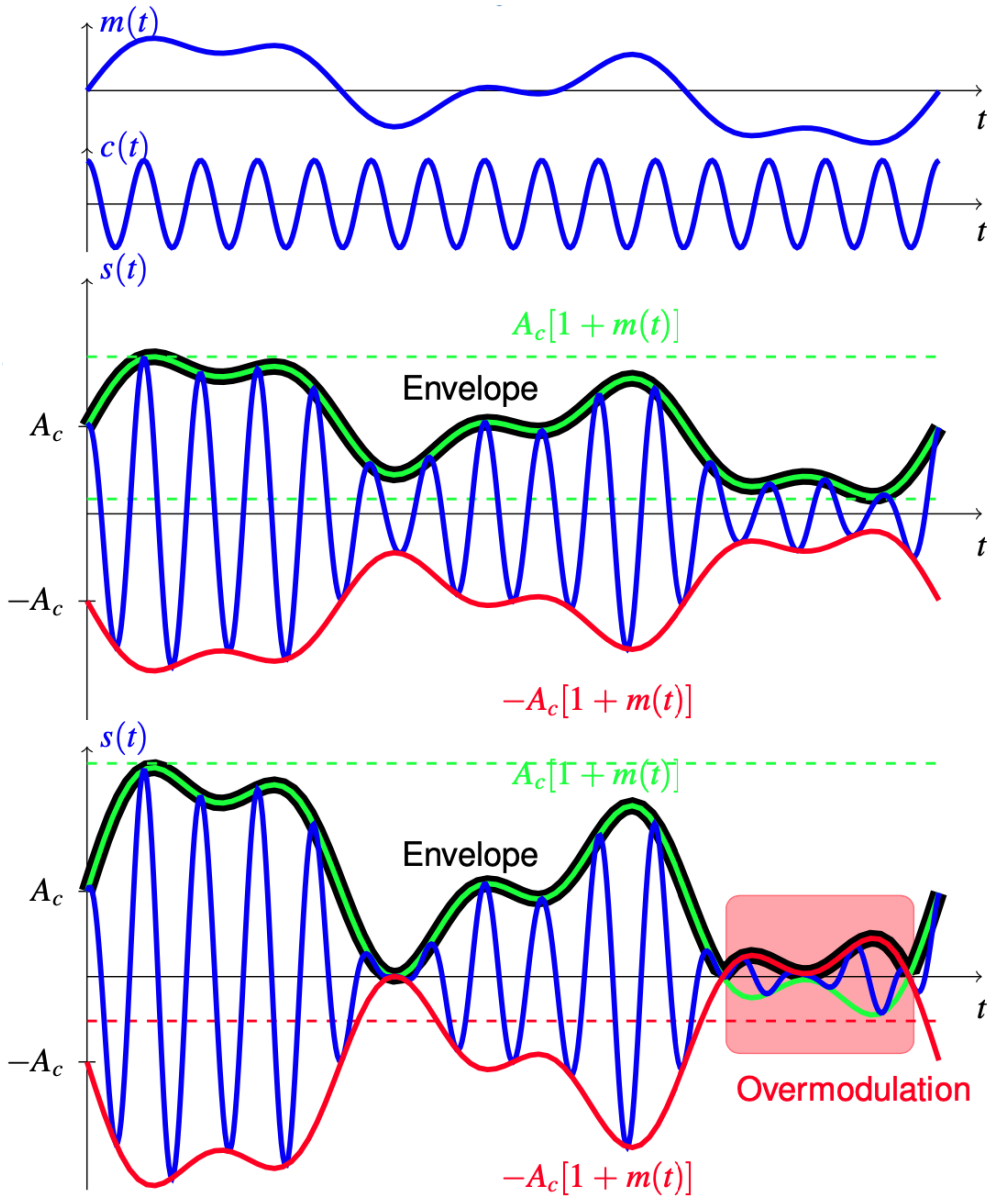


Figure 2.6: Example of modulating signal, $m(t)$, carrier signal, $c(t)$, and modulated signal, $s(t)$, for a conventional AM without (above) and with (below) overmodulation.

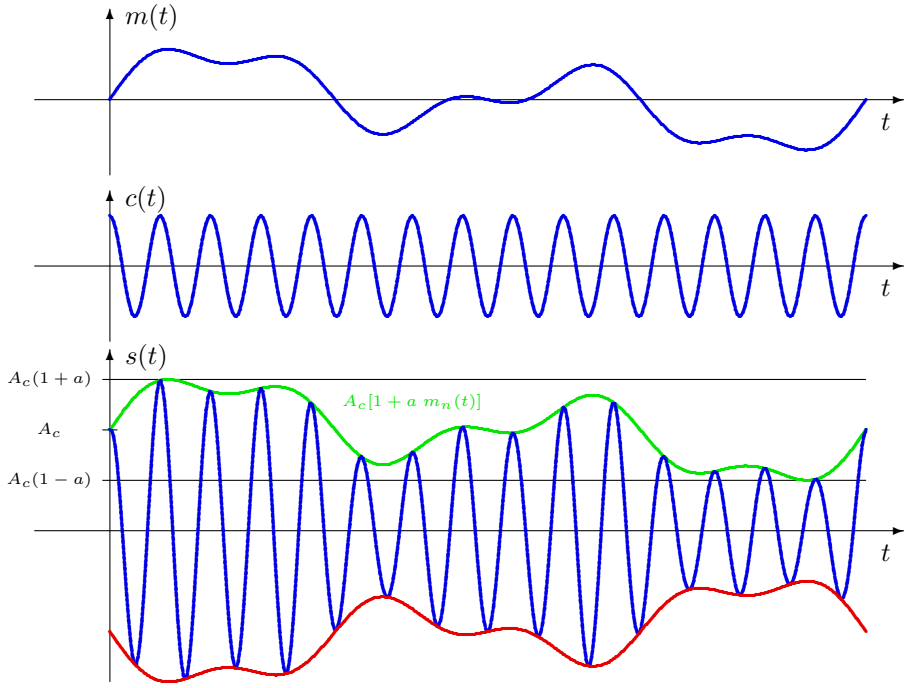


Figure 2.7: Example of a conventional AM signal with modulation index $a = \frac{1}{2}$.

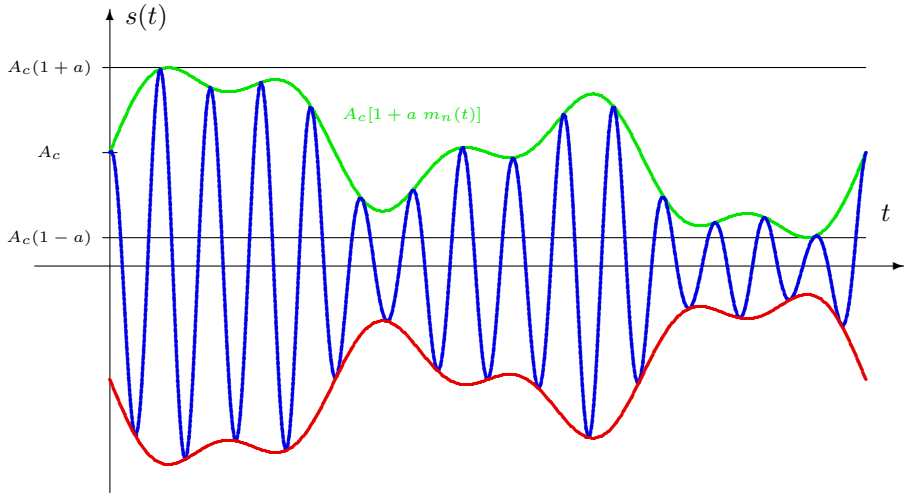


Figure 2.8: Example of a conventional AM signal with modulation index $a = \frac{3}{4}$.

between $\frac{A_c}{4}$ and $\frac{7A_c}{4}$. With respect to the previous case, now the smallest values of the envelope are closer to zero.

If the modulation index were to take a value greater than 1, for example $a = \frac{3}{2}$, now the amplitude term $A_c[1 + m_a(t)]$ can take negative values. the resulting modulated signal is the one shown in Figure 2.9. It can be seen that now the signal envelope no longer contains the shape of the modulating signal (information), and that where the amplitude term $A_c[1 + m_a(t)] < 0$ there is a 180^{o} phase shift in the sinusoid of the modulated signal, and the envelope is proportional to $-m(t)$. This makes impossible to demodulate the signal from the envelope in this case.

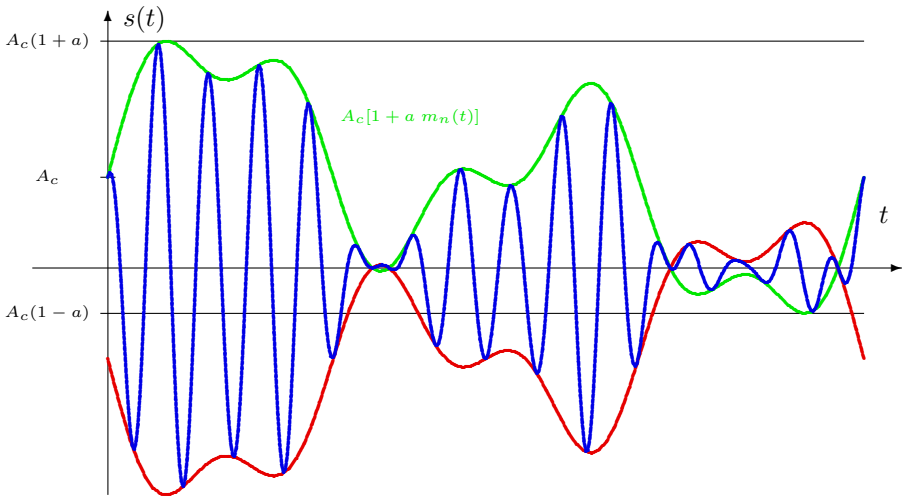


Figure 2.9: Example of a conventional AM signal with modulation index $a = 1.5$.

Spectrum of the conventional AM signal - Deterministic case

In this section, the frequency response of the modulated signal will be obtained when considering a deterministic modulating signal $m(t)$ with Fourier transform $M(j\omega)$. The signal is bandlimited, with $M(j\omega) = 0$ for $|\omega| > W = 2\pi B$ rad/s. It is convenient to remember that $m_a(t) = a\ m_n(t)$, with $m_n(t) = \frac{1}{C_M}\ m(t)$. In the frequency domain this means that $M_a(j\omega) = a\ M_n(j\omega) = \frac{a}{C_M}\ M(j\omega)$. Considering that the modulated signal is

$$s(t) = A_c \cos(\omega_c t + \phi_c) + m_a(t) \times A_c \cos(\omega_c t + \phi_c)$$

applying the basic property of the Fourier transform that a product of signals in time becomes a convolution of their Fourier transforms in the frequency domain, and taking into account that the Fourier transform of a sinusoid is two deltas, and that the phase term in the sinusoid implies a complex exponential in the frequency domain, the Fourier transform of the modulated signal is

$$\begin{aligned} S(j\omega) =& \mathcal{FT}\{A_c \cos(\omega_c t + \phi_c)\} + \frac{1}{2\pi}\ \mathcal{FT}\{m_a(t)\} * \mathcal{FT}\{A_c \cos(\omega_c t + \phi_c)\} \\ =& A_c\ \pi\ [\delta(\omega - \omega_c)\ e^{j\phi_c} + \delta(\omega + \omega_c)\ e^{-j\phi_c}] \\ &+ \frac{A_c}{2}\ [\underbrace{M_a(j\omega - j\omega_c)}_{\frac{a}{C_M}\ M(j\omega - j\omega_c)}\ e^{j\phi_c} + \underbrace{M_a(j\omega + j\omega_c)}_{\frac{a}{C_M}\ M(j\omega + j\omega_c)}\ e^{-j\phi_c}] \end{aligned}$$

If this expression is analyzed, the following conclusions are reached

- Modulus of the Fourier transform $S(j\omega)$
  - Two deltas, in $-\omega_c$ and in $+\omega_c$
    - Amplitude $A_c\pi$
  - Replicas of the shape of $M(j\omega)$ shifted $-\omega_c$ and $+\omega_c$
    - Scale factor $\frac{aA_c}{2C_M}$
- Phase of the Fourier transform
  - The carrier phase introduces the term $e^{-j\phi_c}$
    - Constant phase term
- Bandwidth of the modulated signal

$$W_{AM} = 2\,W \text{ rad/s}, \qquad B_{AM} = 2\,B \text{ Hz}$$

One of the most important characteristics of the conventional AM modulation is that its bandwith is twice the bandwidth of the modulating signal. Figures 2.10 and 2.11 show two examples of the Fourier transform of the modulated signal, $S(j\omega)$, for two particular cases of frequency response of the modulating signal, $M(j\omega)$. It can be easily seen that for a modulating signal of bandwidth W rad/s, the replica of $M(j\omega)$ that is shifted to ω_c in the modulated signal, $M(j\omega - j\omega_c)$, has a support of $2W$ rad/s, independently of the shape of $M(j\omega)$.

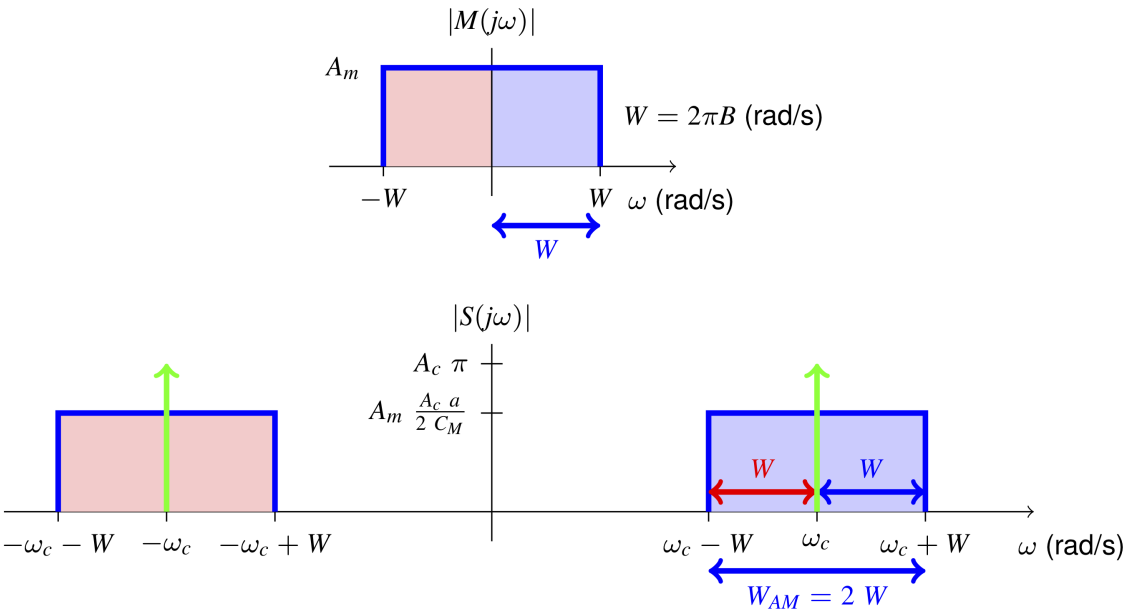


Figure 2.10: An example of the Fourier transform of a conventional AM modulated signal.

Statistical analysis of conventional AM modulation

We now consider that the modulating signal is a random signal whose statistics are known, which is modeled by a stationary random process, $M(t)$, with the characteristics defined in Section 2.1.1

$$M(t),\ \text{stationary, with } m_M = 0,\ R_M(\tau),\ S_M(j\omega),\ \text{ and power } P_M.$$

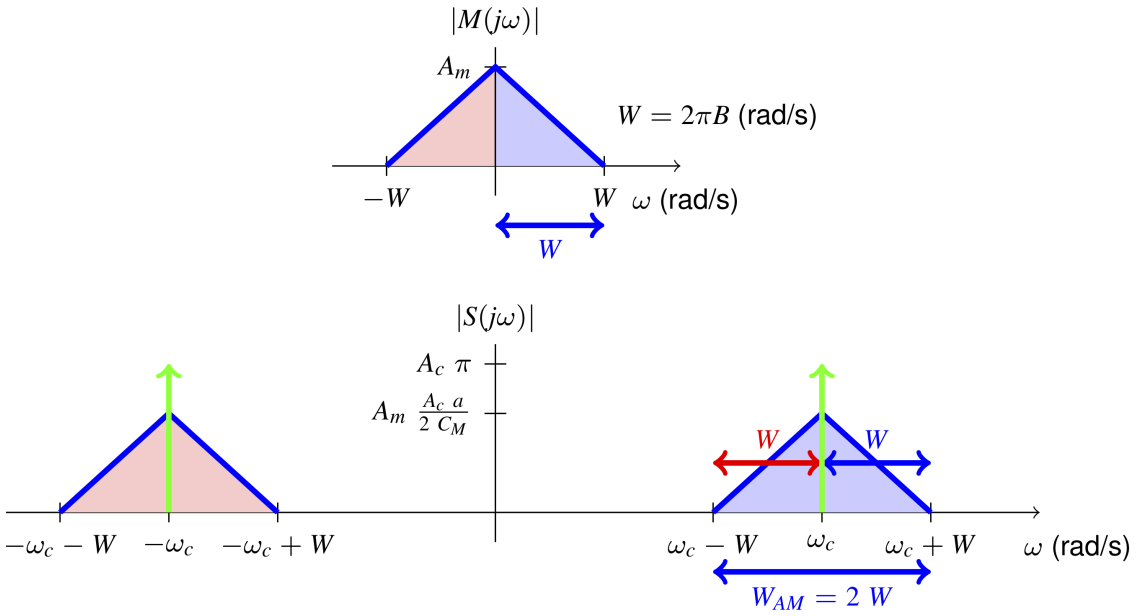


Figure 2.11: Another example of the Fourier transform of the conventional AM modulated signal.

Firstly, the statistical parameters of the random process that models the modulated signal will be obtained. It is defined as

$$S(t) = A_c[1 + M_a(t)]\cos(\omega_c t + \phi_c),$$

where the random process $M_a(t)$ models the modulating signal with modulation index a, which is given by $M_a(t) = a\ M_n(t) = \frac{a}{C_M}\ M(t)$. The mean of this random process $S(t)$ is

$$m_S(t) = E[S(t)] = A_c[1 + E[M_a(t)]]\cos(\omega_c t + \phi_c) = A_c\cos(\omega_c t + \phi_c),$$

since $E[M_a(t)]$ is the mean of $M_a(t)$, and if $M_a(t) = a\ M_n(t) = \frac{a}{C_M}\ M(t)$, then the mean of $M_a(t)$ is $E[M_a(t)] = \frac{a}{C_M}\ E[M(t)] = 0$.

The autocorrelation function is

$$\begin{aligned} R_S(t+\tau, t) =& E[S(t+\tau) \times S(t)] \\ =& A_c^2\ E\left[\underbrace{(1 + M_a(t+\tau))(1 + M_a(t))}_{1+M_a(t)+M_a(t+\tau)+M_a(t+\tau)\times M_a(t)}\right]\cos(\omega_c(t+\tau) + \phi_c)\cos(\omega_c t + \phi_c) \\ =& \frac{A_c^2}{2}\left[1 + R_{M_a}(\tau)\right]\left[\cos(\omega_c\tau) + \cos(\omega_c(2t+\tau) + 2\phi_c)\right]. \end{aligned}$$

We have taken into account the linearity of the mathematical expectation operator, that $E[1] = 1$, and the fact that since $M_a(t)$ is a stationary random process, $E\left[M_a(t)\right] = E\left[M_a(t+\tau)\right] = 0$, and that $E\left[M_a(t) \times M_a(t+\tau)\right]$ is the definition of the autocorrelation function of the random process $M_a(t)$, i.e, $R_{M_a}(\tau)$. Moreover, the following trigonometric equality has been used

$$\cos(a) \times \cos(b) = \frac{1}{2}\cos(a-b) + \frac{1}{2}\cos(a+b).$$

Clearly, both the mean and the autocorrelation are periodic functions of period $T_m = \frac{2\pi}{\omega_c} = \frac{1}{f_c}$ for the mean and $T_R = \frac{2\pi}{2\omega_c} = \frac{1}{2f_c}$ for the autocorrelation function. The common period is $T = T_m$.

Therefore the process is a *cyclostationary* random process with period T. Thus, to characterize it, it is necessary to calculate the time average of the autocorrelation over a period, which is

$$\begin{aligned}\tilde{R}_S(\tau) =& \frac{1}{T}\int_{-T/2}^{T/2} R_S(t+\tau,t)\,dt = \frac{A_c^2}{2}\,\left[1+R_{M_a}(\tau)\right]\,\cos(\omega_c\tau)\\ =& \frac{A_c^2}{2}\,\left[1+\frac{a^2}{C_M^2}R_M(\tau)\right]\,\cos(\omega_c\tau).\end{aligned}$$

In this case, it has been taken into account that the integral over the variable t in the cosine of frequency $2\omega_c$ rad/s is made over 2 complete periods of the sinusoid, and therefore it is zero. Also keep in mind that if $M_a(t) = a\ M_n(t) = \frac{a}{C_M}\ M(t)$, then $R_{M_a}(\tau) = a^2\ R_{M_n}(\tau) = \frac{a^2}{C_M^2}\ R_M(\tau)$. Therefore, $S_{M_a}(j\omega) = \frac{a^2}{C_M^2}\ S_M(j\omega)$ y $P_{M_a} = \frac{a^2}{C_M^2}\ P_M$.

Now the power spectral density is the Fourier transform of this time average of the autocorrelation function

$$\begin{aligned}S_S(j\omega) = \mathcal{FT}\{\tilde{R}_S(\tau)\} =& \frac{A_c^2}{2}\,\pi\,\left[\delta(\omega-\omega_c)+\delta(\omega+\omega_c)\right]\\ &+\frac{A_c^2}{4}\,\left[S_{M_a}(j\omega-j\omega_c)+S_{M_a}(j\omega+j\omega_c)\right]\\ =& \frac{A_c^2}{2}\,\pi\,\left[\delta(\omega-\omega_c)+\delta(\omega+\omega_c)\right]\\ &+\frac{A_c^2}{4}\,\left[\frac{a^2}{C_M^2}\,S_M(j\omega-j\omega_c)+\frac{a^2}{C_M^2}\,S_M(j\omega+j\omega_c)\right].\end{aligned}$$

And the power of the process can be calculated as

$$\begin{aligned}P_S =& \tilde{R}_S(0) = \frac{1}{2\pi}\int_{-\infty}^{\infty} S_S(j\omega)\,d\omega\\ =& \frac{A_c^2}{2}\,[1+R_{M_a}(0)]\\ =& \frac{A_c^2}{2}\,[1+P_{M_a}] = \frac{A_c^2}{2}\,\left[1+\frac{a^2}{C_M^2}\,P_M\right].\end{aligned}$$

Analyzing these expressions, the following conclusions are reached:

- Bandwidth of the conventional AM signal

  $$W_{AM} = 2\,W \text{ rad/s}, \qquad B_{AM} = 2\,B \text{ Hz}$$

  The power spectral density is composed of
  - Two deltas, located at $-\omega_c$ and at $+\omega_c$
    - Amplitude $\frac{A_c^2}{2}\pi$
  - Two replicas of $S_M(j\omega)$, shifted $-\omega_c$ and $+\omega_c$
    - Scale factor $\left(\frac{aA_c}{2C_M}\right)^2$
- Power of conventional AM modulation

  $$P_S = \tilde{R}_S(0) = \frac{A_c^2}{2}\,\left[1+R_{M_a}(0)\right] = \frac{A_c^2}{2}\,\left[1+P_{M_a}\right] = \frac{A_c^2}{2}\,\left[1+\frac{a^2}{C_M^2}\,P_M\right]$$

  Two components can be distinguished:

– Power of the carrier: $\frac{A_c^2}{2}$

– Power of the double sideband component: $\left(\frac{A_c^2}{2}\frac{a^2}{C_M^2}\right) \times P_M$

The carrier power is not useful power from the point of view of information transmission (although it is useful because, as we will see, it allows the use of a simple receiver).

Figures 2.12 and 2.13 show two examples of power spectral density of the modulated signal, for two particular cases of the power spectral density of the modulating signal.

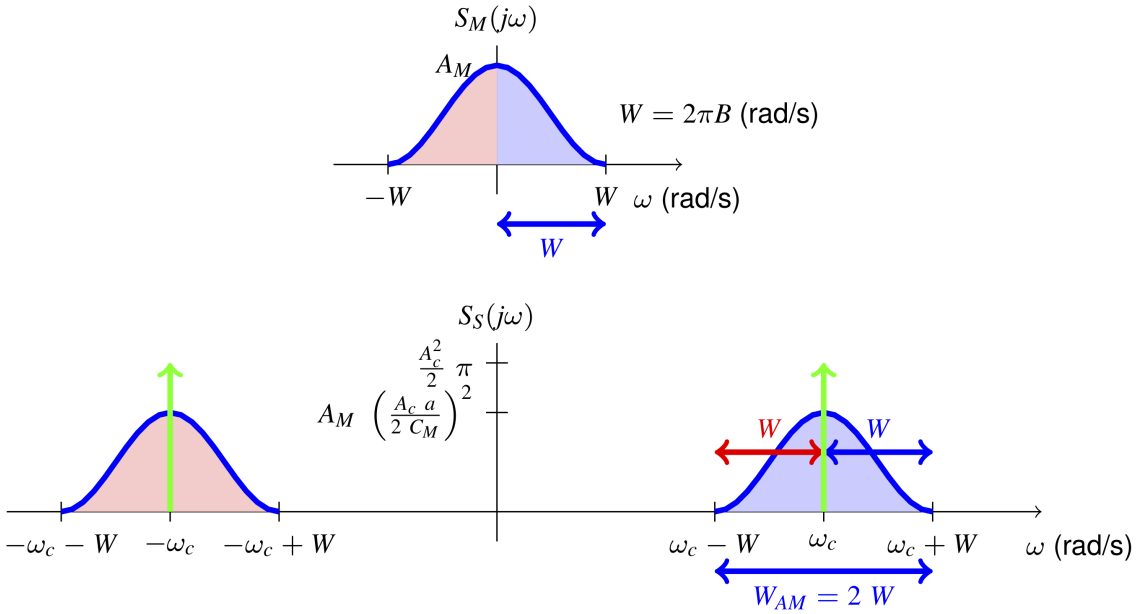


Figure 2.12: An example of power spectral density of the conventional AM modulated signal.

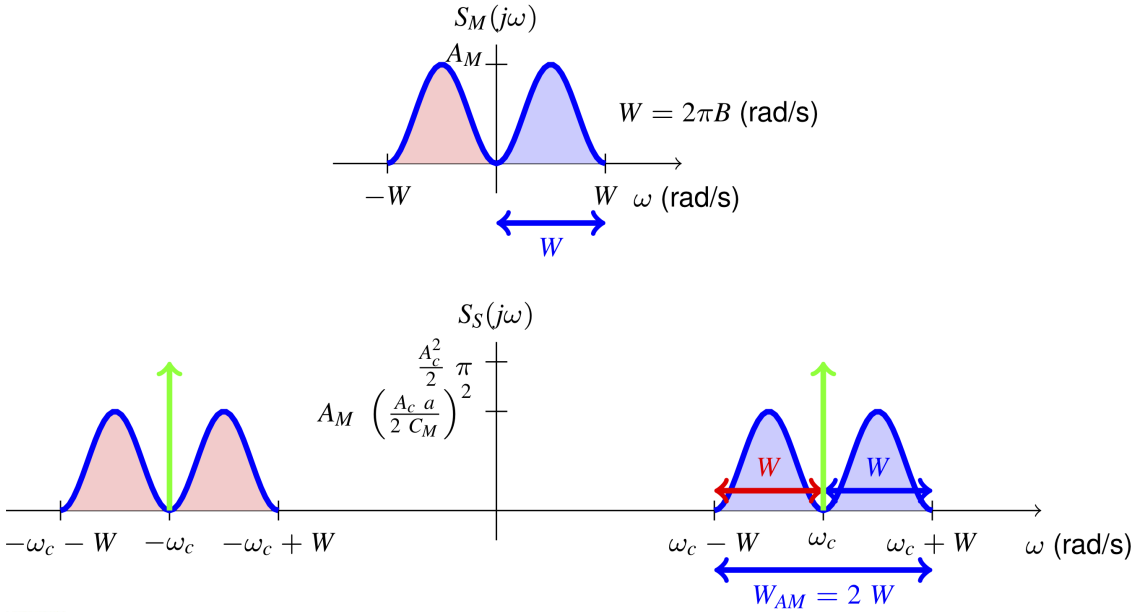


Figure 2.13: Another example of power spectral density of the conventional AM modulated signal.

It can be easily seen that the bandwidth of the modulated signal is twice the bandwidth of the modulating signal, independently of the shape of the power spectral density of the modulating signal.

Demodulation of conventional AM modulation

The most important advantage of this modulation lies in how it can be demodulated:

1. Since the envelope is proportional to the modulating signal, a simple envelope detector allows to recover the information.
2. It does not need a synchronous or coherent demodulator, although it can be used as well (this receiver will be seen later, when discussing the double sideband modulation).

Since $|m_a(t)| < 1$, the envelope is proportional to $1 + m_a(t) > 0$, and therefore proportional to $m(t)$. This allows the receiver to be implemented by means of a simple envelope detector, which can be implemented by means of a rectifier and a low-pass filter with the cut-off frequency matched to the signal bandwidth, B Hz.

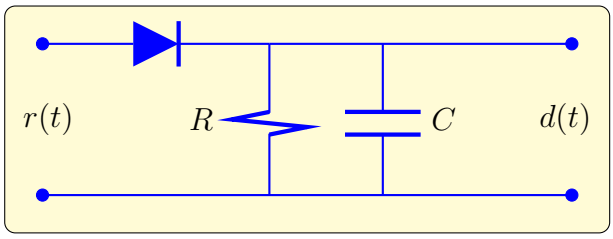


Figure 2.14: Envelope detection of a conventional AM signal.

This demodulator recovers the signal with a gain factor and with a DC term that can be easily removed.

$$d(t) \approx A_c \left[1 + \frac{a}{C_M} m(t)\right].$$

The simplicity of this demodulator makes this modulation the one used for AM radio broadcasting. This is because

1. Receivers are very simple, and there are millions of them.
2. Although it is not efficient in terms of transmitted power, there are few transmitters, which limits this problem in practice.

As a summary of the characteristics of this modulation, the following aspects could be cited:

- Drawbacks of the conventional AM modulation:
  - Low power efficiency:
    - Power is spent in the transmission of the carrier (which does not contain information by itself).

  - Low spectral efficiency:
    * The bandwidth of the modulated signal is twice the bandwidth of the modulating signal.

- Fundamental advantage of the conventional AM modulation
  - If $a \leq 1$, there is no overmodulation and the signal envelope is proportional to $1 + m_a(t) \geq 0$, from which $m(t)$ can be extracted:
    * Mean removal and scaling.
  - Simple receiver: envelope detector:
    * No need for a synchronous demodulator.

2.2.2 Double Sideband (DSB), no carrier

This technique suppresses the carrier of the conventional AM modulation so that the problem of power efficiency thereof is eliminated. Its mathematical expression is

$$s(t) = m(t) \times c(t) = m(t) \times A_c \cos(2\pi f_c t + \phi_c).$$

Removing the carrier causes the signal envelope to no longer contain the waveform of the modulating signal. As we will see later, the demodulation requires a more complex receiver. Figure 2.15 shows an example of a modulated signal.

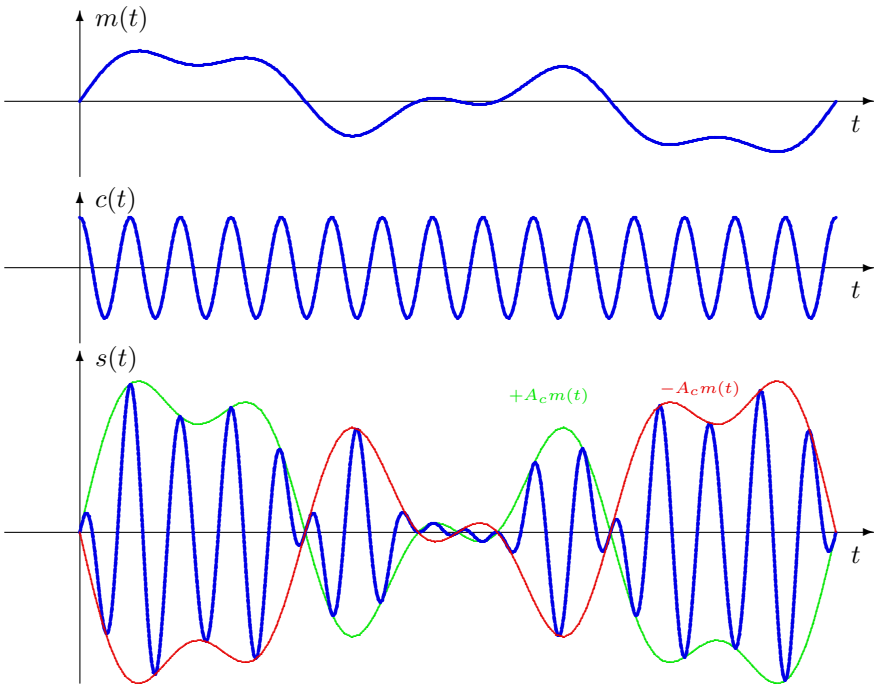


Figure 2.15: Example of a double sideband modulated signal.

DSB spectrum - Deterministic case

A deterministic signal $m(t)$ with Fourier transform $M(j\omega)$ is considered. The spectrum of the DSB modulated signal is

$$\begin{aligned} S(j\omega) =& \frac{1}{2\pi}\, \mathcal{FT}\{m(t)\} * \mathcal{FT}\{A_c\, \cos(\omega_c t + \phi_c)\} \\ =& \frac{A_c}{2}\, [M(j\omega - j\omega_c)\, e^{j\phi_c} + M(j\omega + j\omega_c)\, e^{-j\phi_c}]. \end{aligned}$$

With respect to conventional AM modulation, it can be seen that:

- The deltas of conventional AM modulation disappear.
- The scaling of the replicas of $M(t)$ is different, since there is no normalization in this case, so that the factors a and C_M disappear.

The name of the modulation refers to the fact that two sidebands appear in the signal spectrum, lower ($|w| < w_c$) and upper ($|w| > w_c$), each of them being symmetric with respect to the other:

1. $|\omega| > \omega_c$: Upper sideband
2. $|\omega| < \omega_c$: Lower sideband

Figures 2.16 ans 2.17 show two examples of Fourier transform of the modulated signal, for two particular cases of the frequency response of the modulating signal.

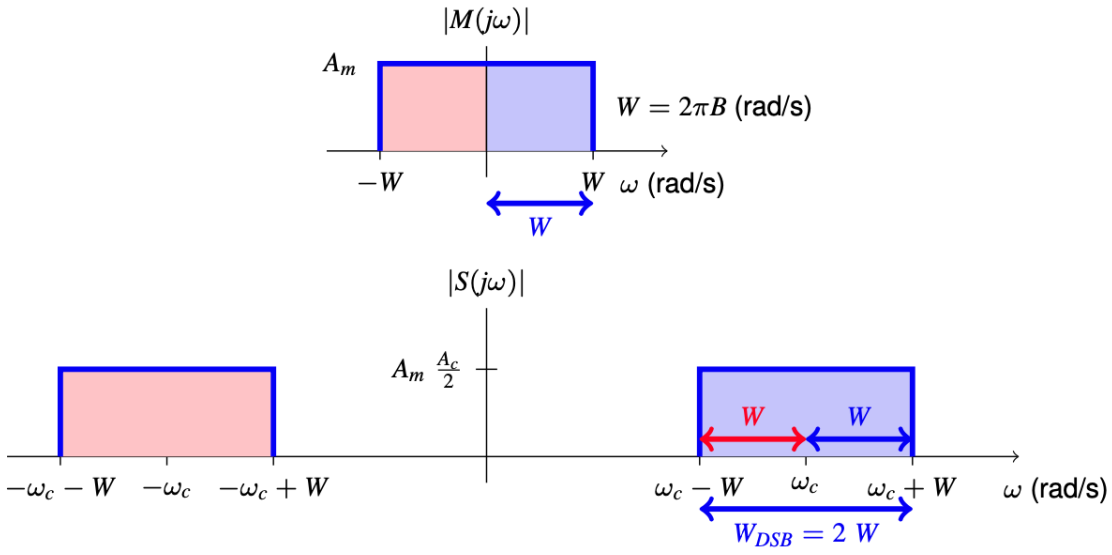


Figure 2.16: An example of a Fourier transform of the double sideband modulated signal.

Clearly, the spectrum of the DSB signal occupies twice the bandwidth of the modulating signal.

$$W_{DSB} = 2W \text{ rad/s, or alternatively } B_{DSB} = 2B \text{ Hz.}$$

Therefore, spectrally the DSB is still just as inefficient as the conventional AM modulation. DSB spectrum also contains two sidebands. It should be noted that each one of the sidebands contains all the information of the signal, it has all the frequency components of it.

With respect to the power, the carrier has been suppressed, so the deltas in $\pm f_c$ do not appear, and power is not wasted in a non informative component.

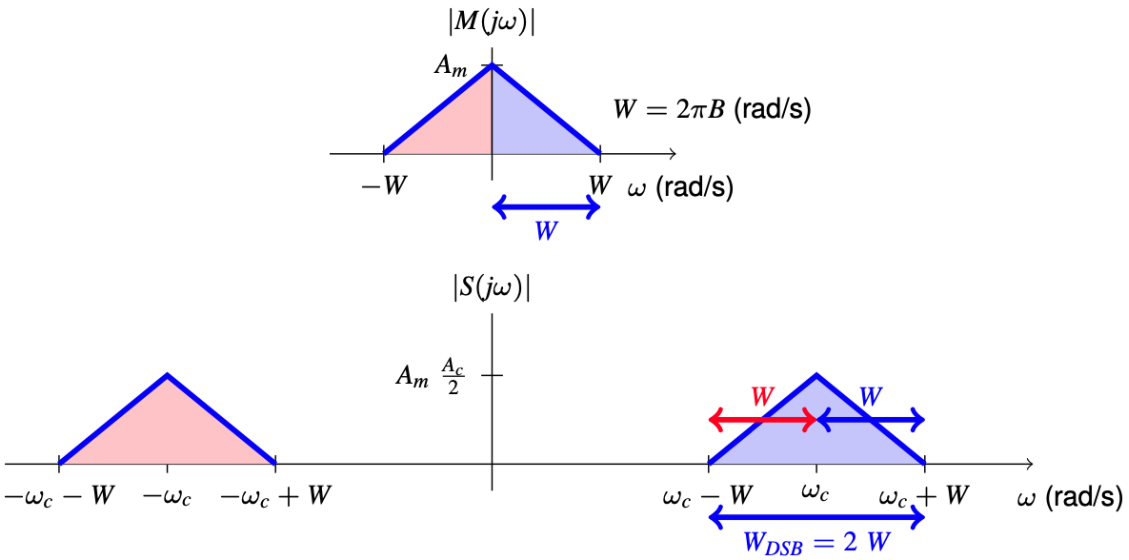


Figure 2.17: Another example of a Fourier transform of the double sideband modulated signal.

Statistical analysis of double sideband modulation

To carry out this analysis, stochastic processes are used again. The message, or modulating signal, is modeled as a wide sense stationary stochastic process $M(t)$, with the characteristics described in Section 2.1.1. The modulated signal is modeled by the random process defined as

$$S(t) = M(t) \times c(t) = A_c M(t) \cos(\omega_c t + \phi_c).$$

The mean of the modulated signal is

$$m_S(t) = E[S(t)] = A_c \; E[M(t)] \; \cos(\omega_c t + \phi_c) = 0.$$

And the autocorrelation function

$$\begin{aligned} R_S(t+\tau, t) =& A_c^2 \; E[M(t+\tau) \; M(t)] \; \cos(\omega_c(t+\tau) + \phi_c) \; \cos(\omega_c t + \phi_c) \\ =& \frac{A_c^2}{2} \; R_M(\tau) \; \left[\cos(\omega_c \tau) + \cos(\omega_c(2t+\tau) + 2\phi_c)\right]. \end{aligned}$$

Although the mean is constant, the autocorrelation function is periodic with period $T = \frac{1}{2f_c}$. Therefore the process is a *cyclostationary* random process. Thus, to characterize it, it is necessary to calculate the time average (in one cycle) of the autocorrelation function

$$\tilde{R}_S(\tau) = \frac{1}{T} \int_{-T/2}^{T/2} R_S(t+\tau, t) \; dt = \frac{A_c^2}{2} \; R_M(\tau) \; \cos(\omega_c \tau).$$

Again it has been taken into account that the integral in a cycle of a sinusoid is zero. The power spectral density is obtained through the Fourier transform

$$S_S(j\omega) = \mathcal{FT}\left\{\tilde{R}_S(\tau)\right\} = \frac{A_c^2}{4} \; \left[S_M(j\omega - j\omega_c) + S_M(j\omega + j\omega)\right].$$

The power of the modulated signal is

$$P_S = \tilde{R}_S(0) = \frac{A_c^2}{2} R_M(0) = \frac{A_c^2}{2} P_M.$$

Now there is no longer a power term associated with the carrier, so this modulation is more efficient than conventional AM modulation in terms of power.

Figures 2.18 and 2.19 show two examples of power spectral density of the modulated signal, for two particular cases of power spectral density of the modulating signal.

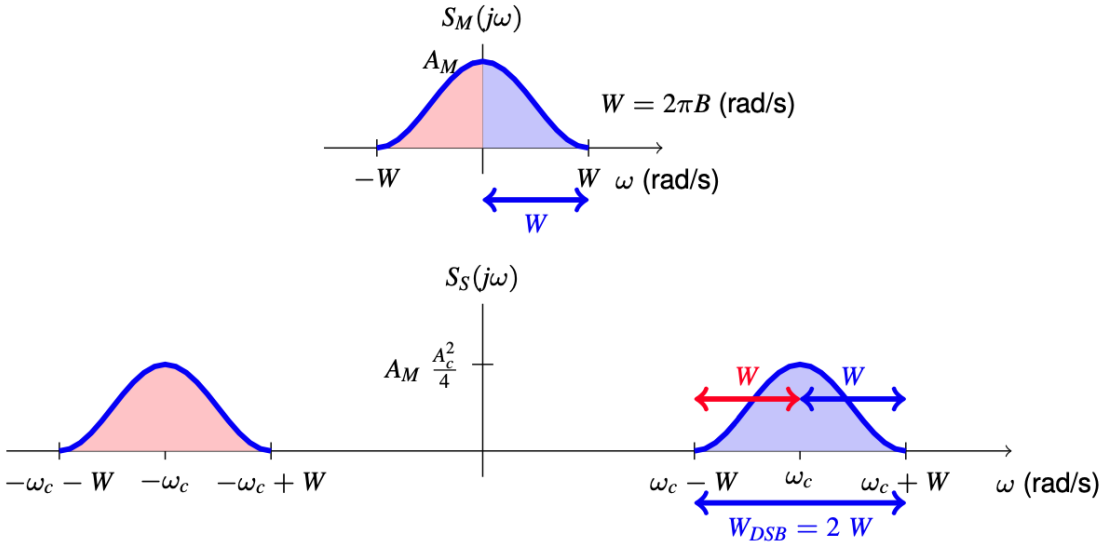


Figure 2.18: An example of the power spectral density of a double sideband modulated signal.

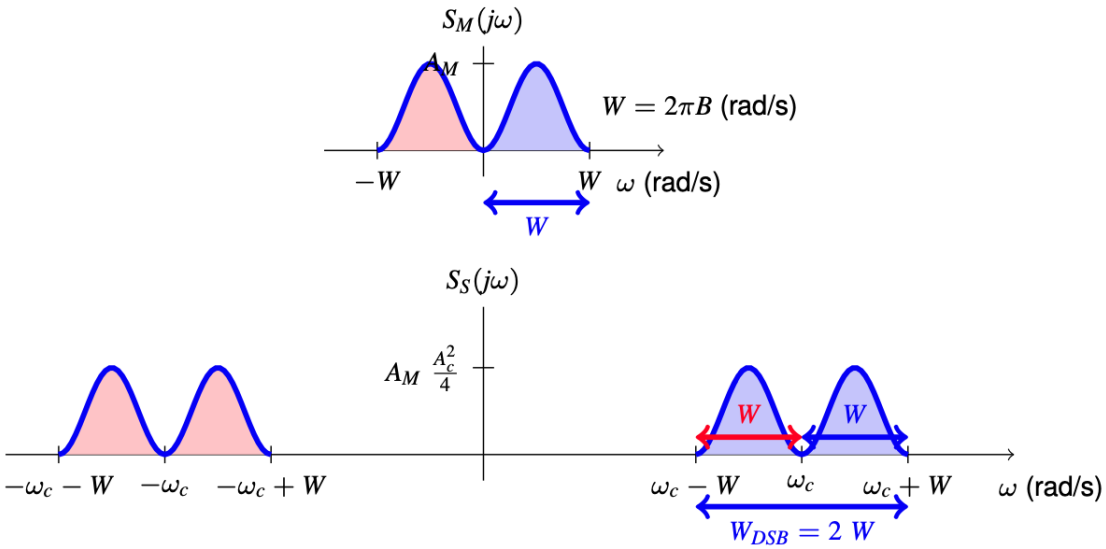


Figure 2.19: Another example of double sideband modulated signal power spectral density.

Demodulation of DSB signals

The suppression of the carrier, as said above, and as it can be seen for instance in Figure 2.15, makes the shape of the signal envelope no longer proportional to the modulating signal, so it is not possible to use an envelope detector. In this case it is necessary to use a synchronous receiver or coherent receiver. This type of receiver is shown in Figure 2.20, where the abbreviation LPF stands for *Low Pass Filter*. In this case the filter has a bandwidth of B Hz.

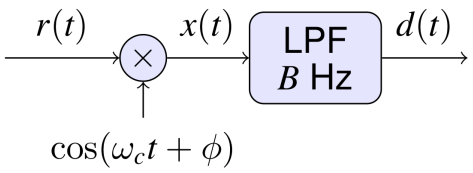


Figure 2.20: Synchronous or coherent demodulator for a double sideband signal.

Optimum performance is obtained with a synchronous or coherent receiver, which means that the carrier phase of the receiver is the same as the carrier phase that was used in the transmitter (in the generation of the modulated signal), i.e.

$$\phi = \phi_c.$$

As we will see later, if this condition is not satisfied (non-synchronous or non-coherent receiver), there is an attenuation of the received signal and therefore a loss of signal-to-noise ratio and performance.

Next, we will proceed to the analysis of the operation of the receiver. Initially, it is assumed that the signal does not suffer any distortion during its transmission (ideal situation), so that the received signal is equal to the transmitted modulated signal.

$$r(t) = s(t) = A_c m(t) \cos(\omega_c t + \phi_c).$$

The demodulated signal before filtering, $x(t)$, is

$$\begin{aligned} x(t) =& r(t) \times \cos(\omega_c t + \phi) \\ =& A_c m(t) \cos(\omega_c t + \phi_c) \cos(\omega_c t + \phi) \\ =& \frac{A_c}{2}\, m(t)\ \left[\cos(\phi - \phi_c) + \cos(2\omega_c t + \phi_c + \phi)\right]. \end{aligned}$$

The low-pass filter removes high-frequency components, these component located arround $2\omega_c$, so the filtered output is

$$d(t) = \frac{A_c}{2}\, m(t)\ \cos(\phi_c - \phi).$$

Figure 2.21 shows a frequency interpretation of the demodulation process. Bearing in mind that the product with a sinusoid produces two replicas of the spectrum (each with half the amplitude of the sinusoid), one shifted ω_c to the right, and another ω_c to the left

$$X(j\omega) = \frac{1}{2} R(j\omega - j\omega_c) + \frac{1}{2} R(j\omega + j\omega_c),$$

the contribution of both terms reconstructs the spectrum of the signal and adds a high-frequency component (centered at $2\omega_c$ that is removed with low-pass filtering).

As the goal is to recover a signal proportional to $m(t)$ with the highest possible amplitude, obviously the best option is to use a coherent receiver ($\phi = \phi_c$). In this case, the output of the coherent receiver is

$$d(t) = \frac{A_c}{2}\, m(t).$$

In case of using a non-coherent receiver, a factor $\cos(\phi_c - \phi)$ appears multiplying the desired signal. This term is an attenuation term. Multiple values can be given for the angle difference. Some illustrative cases are:

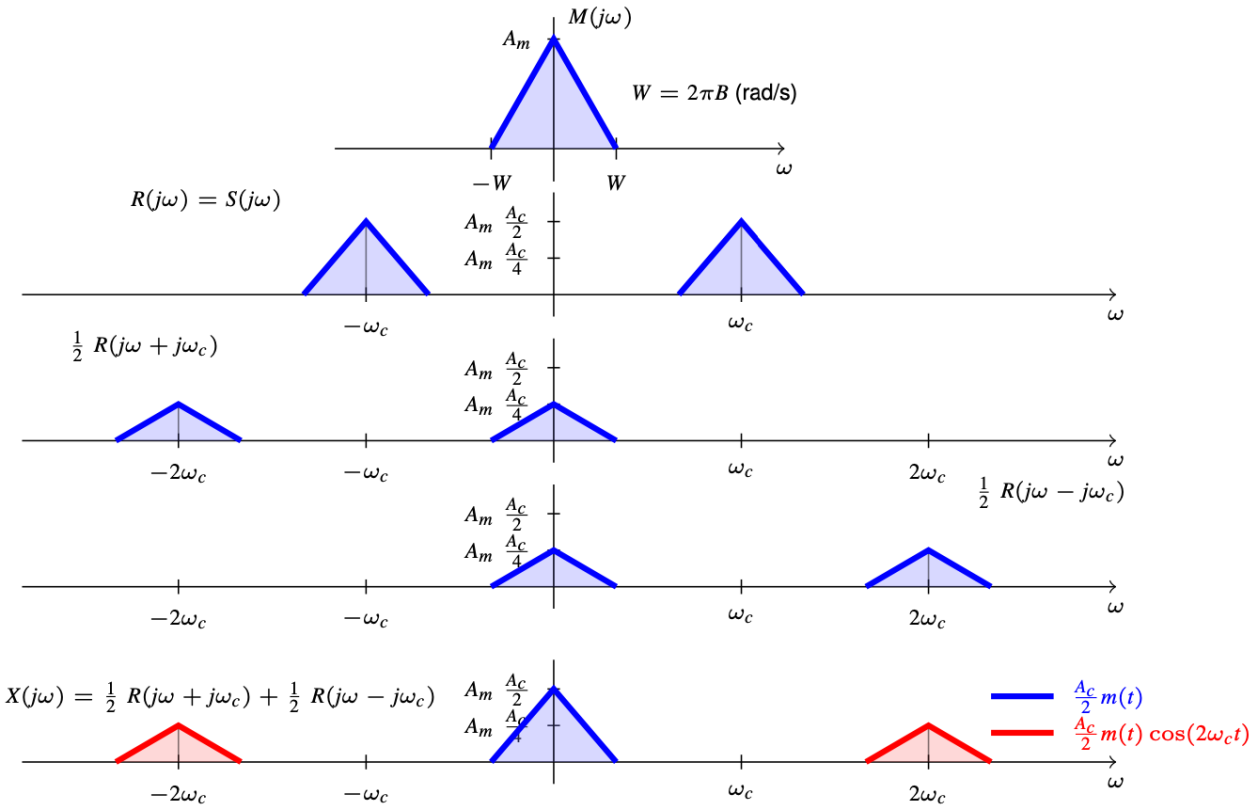


Figure 2.21: Frequency interpretation of the demodulation process of a double sideband signal.

1. $\phi = \phi_c$. This is the *ideal* case, in which the cosine is equal to 1.

2. $\phi_c - \phi = 45^{\text{o}}$. The amplitude is reduced by a factor $\sqrt{2}$, which means that the power is cut in half.

3. $\phi_c - \phi = 90^{\text{o}}$. The modulating signal $m(t)$ disappears.

This indicates the need to have a *synchronous demodulation* or *coherent demodulation* or simply coherent.

To generate a sinusoidal at the receiver that is locked in phase with the carrier that was used to generate the received signal, there are two options

1. To transmit a pilot tone (a low amplitude carrier). This tone is extracted at the receiver with a narrow band filter tuned to frequency ω_c. This option has the disadvantage that power efficiency is lost, since a part of the power of the transmitted signal is used in the generation of the pilot, which does not actually contain information.

2. To introduce a Phase-Locked Loop (PLL), a device that makes it possible to recover the phase of the carrier from the received signal. This alternative makes the receiver more complex and with a higher cost.

2.2.3 Single Sideband (SSB) modulation

In double-sideband modulation and conventional AM, both sidebands are present, each of which contains all the information of the transmitted or modulated signal due to the symmetry property of the frequency response of real signals. The use of both bands is redundant and wastes an important resource such as bandwidth. Single-sideband modulation transmits a single sideband, halving the bandwidth of conventional AM and double-sideband modulation.

$$W_{SSB} = W \text{ rad/s}, \qquad B_{SSB} = B \text{ Hz}.$$

In this case, the bandwidth of the modulated signal coincides with the bandwidth of the modulating signal.

Figure 2.22 shows the frequency response of this type of modulation and the corresponding bandwidth.

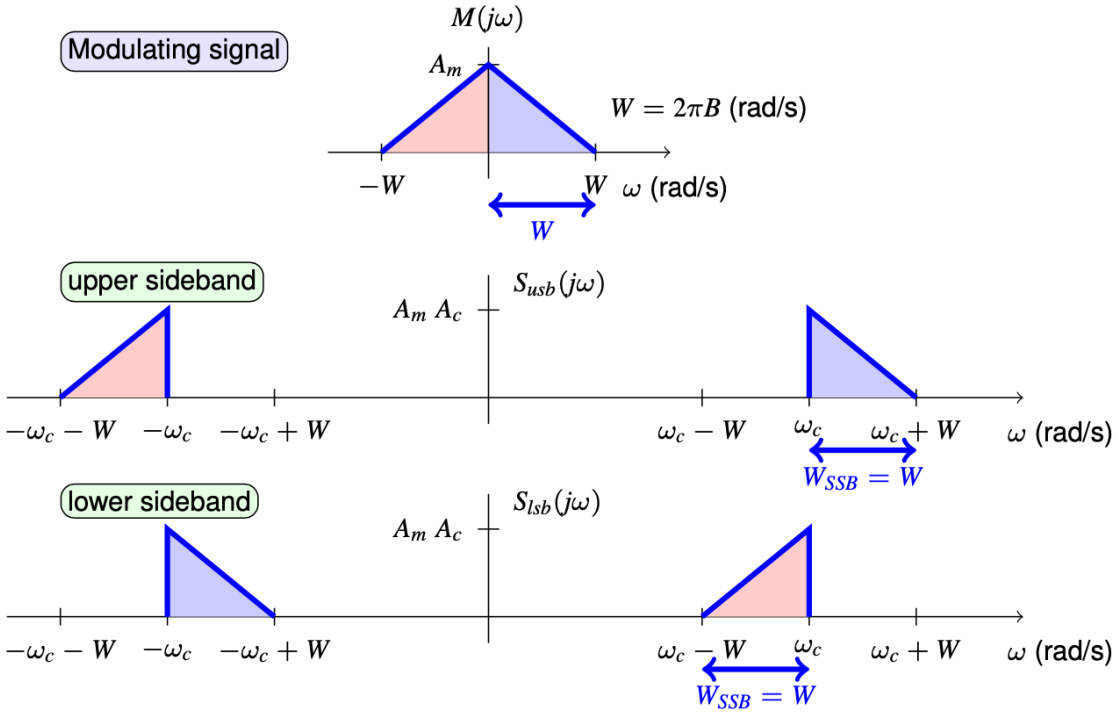


Figure 2.22: Spectrum and bandwidth of single sideband: upper sideband and lower sideband signals.

The simplest option for the generation of this type of signals is through direct filtering; in this case a double sideband signal is generated, and then one of the two sidebands is filtered out:

- SSB upper Sideband (USB): frequencies $|\omega| < \omega_c$ are removed
- SSB lower sideband (LSB): frequencies $|\omega| > \omega_c$ are removed

The scheme of this SSB signal generation method is shown in Figure 2.23. Note that the intermediate signal $s_D(t)$ is a double sideband signal but with double amplitude (with respect to the DSB).

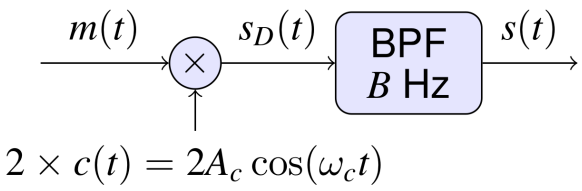


Figure 2.23: Generation of a single sideband (SSB) signal by direct filtering.

The generic frequency response of the single sideband, upper sideband, and lower sideband filters will be

$$H_{usb}(j\omega) = \begin{cases} 1, & \text{if } |\omega| \geq \omega_c \\ 0, & \text{if } |\omega| < \omega_c \end{cases} \text{ and } H_{lsb}(j\omega) = \begin{cases} 0, & \text{if } |\omega| > \omega_c \\ 1, & \text{if } |\omega| \leq \omega_c \end{cases}$$

Figure 2.24 shows the frequency response of the signals generated with this option.

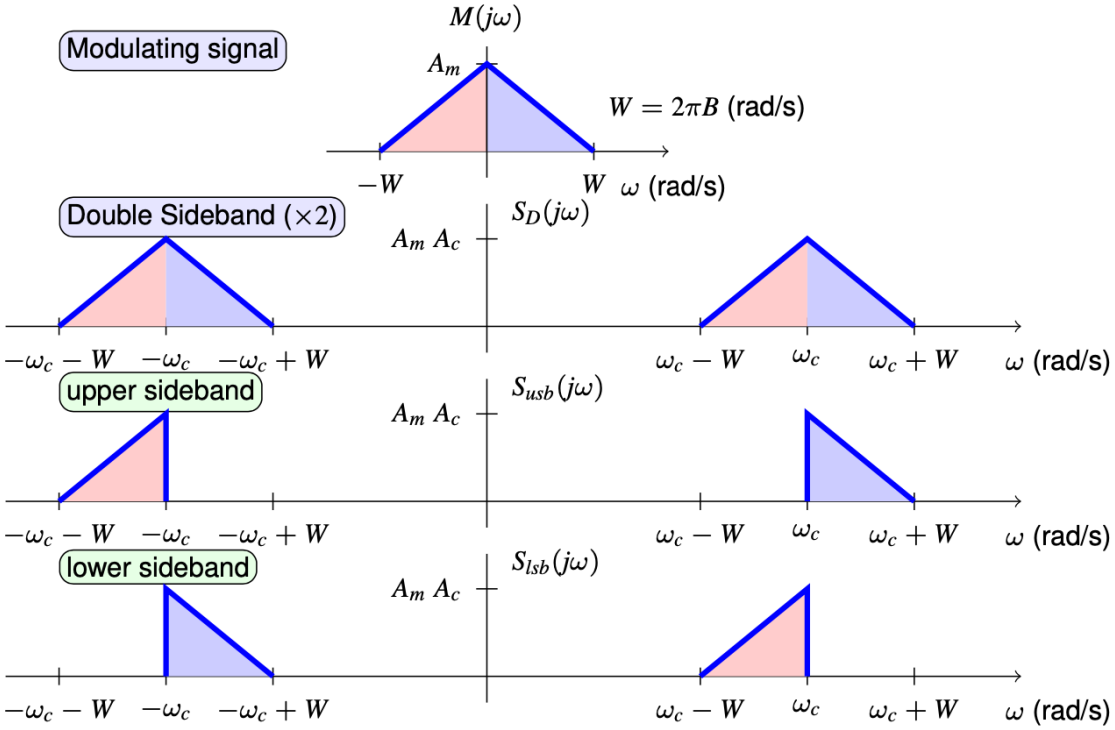


Figure 2.24: Spectrum example of the single sideband modulation: upper sideband (sup) and lower sideband (inf) signals, and the intermediate double sideband signal that is generated before filtering.

There is another alternative for the generation of the single sideband signal. As we will demonstrate below, the SSB signal has the analytical expression

$$s(t) = A_c\, m(t)\ \cos(\omega_c t + \phi_c) \mp A_c\, \hat{m}(t)\ \sin(\omega_c t + \phi_c),$$

where the negative sign corresponds to the upper sideband, the positive sign to the lower sideband, and $\hat{m}(t)$ is the Hilbert transform of the signal $m(t)$.

The Hilbert transform of a signal is the signal obtained by filtering the original signal with a

Hilbert transformer

$$\hat{m}(t) = m(t) * h_{Hilbert}(t),$$

where a Hilbert transformer is a linear filter with impulse response

$$h_{Hilbert}(t) = \frac{1}{\pi t}$$

and frequency response

$$H_{Hilbert}(j\omega) = \begin{cases} -j, & \omega > 0 \\ +j, & \omega < 0 \\ 0, & \omega = 0 \end{cases}.$$

This allows signal generation using a Hilbert transformer and two quadrature oscillators (90 degrees phase shift, such as a cosine and a sine). This scheme for generation, shown in Figure 2.25, is called the Hartley modulator.

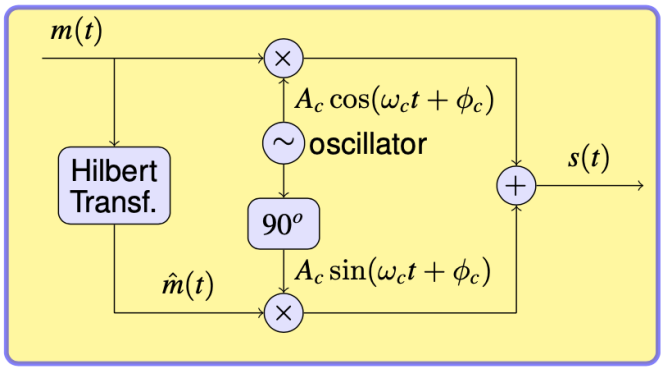


Figure 2.25: Generation of a single sideband (SSB) signal using a Hilbert transformer and two quadrature carriers (Hartley modulator).

Analytical Expression of the Modulated Signal - Upper Sideband

Figure 2.26 shows the frequency response of the signals generated with this option. In this case it can be seen that the frequency response of the single sideband filter for upper sideband can be written as a function of the step function $u(x)$, specifically:

$$H_{usb}(j\omega) = u(\omega - \omega_c) + u(-(\omega + \omega_c))$$

The frequency response of the DSB signal with double amplitude, $s_D(t)$, is

$$S_D(j\omega) = A_c\left[M(j\omega - j\omega_c) + M(j\omega + j\omega_c)\right]$$

Looking at Figure 2.26, the frequency response of the upper sideband signal can be written as

$$\begin{aligned} S_{usb}(j\omega) =& S_D(j\omega)\ H_{usb}(j\omega) \\ =& A_c M(j\omega)\ u(\omega)|_{\omega=\omega-\omega_c} + A_c M(j\omega)\ u(-\omega)|_{\omega=\omega+\omega_c}, \end{aligned}$$

that is, the sum of two terms: the product of the frequency response of the modulating signal and a step function, shifted to $+\omega_c$, and the product of the frequency response and the modulating signal and a frequency reversed step function, shifted to $-\omega_c$, in both cases with the scale factor A_c.

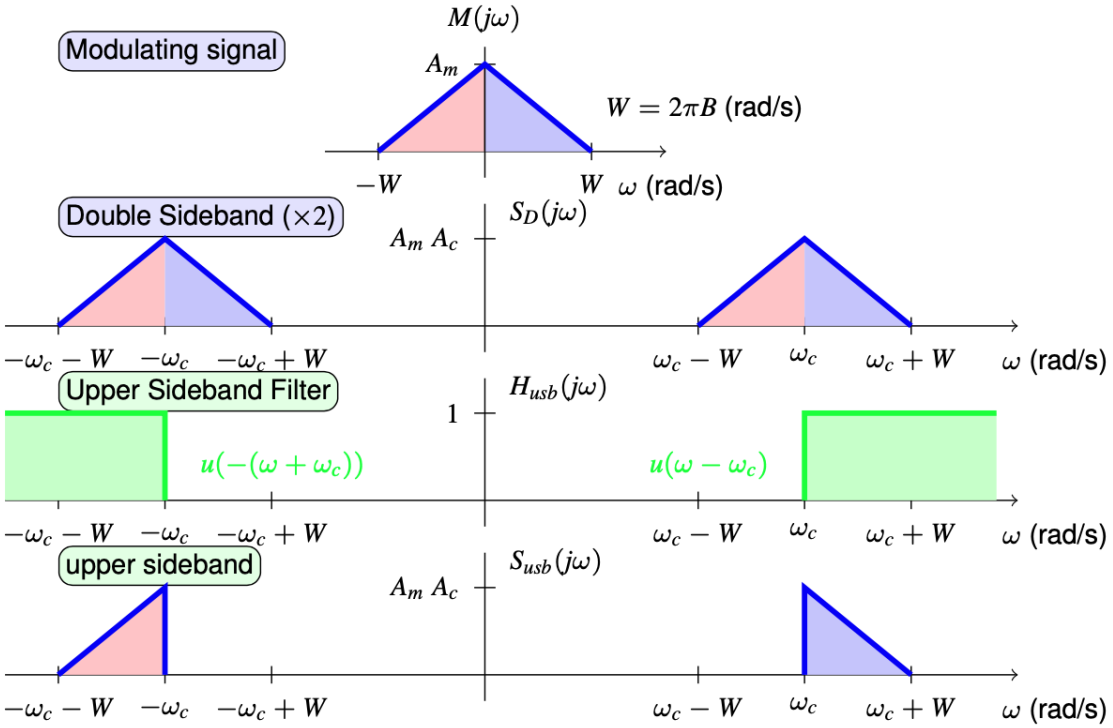


Figure 2.26: Representation of the generation of an upper sideband signal in the frequency domain.

To obtain the response in the time domain, all we have to do is to calculate the inverse Fourier transform, for which the following expressions (properties of the Fourier transform and Euler's formulas for sinusoids) will be useful:

$$\mathcal{FT}\left\{\frac{1}{2}\delta(t) + \frac{j}{2\pi t}\right\} = u(\omega),$$

$$\mathcal{FT}\left\{\frac{1}{2}\delta(t) - \frac{j}{2\pi t}\right\} = u(-\omega),$$

$$\mathcal{FT}\left\{x(t)\, e^{j\omega_c t}\right\} = X(j\omega - j\omega_c)$$

$$\cos(\omega_c t) = \frac{e^{+j\omega_c t} + e^{-j\omega_c t}}{2}$$

$$\sin(\omega_c t) = \frac{e^{+j\omega_c t} - e^{-j\omega_c t}}{2j} = j\frac{e^{-j\omega_c t} - e^{+j\omega_c t}}{2}$$

Taking these relationships into account, it is easy to calculate the inverse Fourier transform of $S_{usb}(j\omega)$

$$\begin{aligned} s_{usb}(t) =& A_c\, m(t) * \left[\frac{1}{2}\delta(t) + \frac{j}{2\pi t}\right] e^{j\omega_c t} + A_c\, m(t) * \left[\frac{1}{2}\delta(t) - \frac{j}{2\pi t}\right] e^{-j\omega_c t} \\ =& \frac{A_c}{2}\left[m(t) + j\hat{m}(t)\right] e^{j\omega_c t} + \frac{A_c}{2}\left[m(t) - j\hat{m}(t)\right] e^{-j\omega_c t} \\ =& A_c\, m(t)\, \cos(\omega_c t) - A_c\, \hat{m}(t)\, \sin(\omega_c t). \end{aligned}$$

Analytical Expression of the Modulated Signal - Lower Sideband

In this case, the above procedure could be repeated by changing the expression of the single-sideband filter to use the lower-sideband one. But it is easier take into account that the sum of the lower-sideband and upper-sideband signals gives rise to the double-amplitude double-sideband signal.

$$s_D(t) = 2\ A_c\ m(t)\ \cos(\omega_c t) = s_{usb}(t) + s_{lsb}(t)$$

as can be seen in Figure 2.24. Taking this into account

$$\begin{aligned} s_{lsb}(t) =& s_D(t) - s_{usb}(t) \\ =& A_c\ m(t)\ \cos(\omega_c t) + A_c\ \hat{m}(t)\ \sin(\omega_c t). \end{aligned}$$

The result is extended straightforward to consider a generic phase term ϕ_c in both the cosine and the sine function.

Power Spectral Density

Taking into account that the signals can be obtained from the filtering of a double-amplitude double-sideband signal, it is trivial to obtain the expressions of the power spectral density of a single-sideband signal from the expressions obtained for double sideband modulation, just taking into account the scaling factor introducing by the double amplitude of the carrier in this case (a factor 2 in the power spectral density):

- Upper sideband

$$S_{S_{usb}}(j\omega) = \begin{cases} A_c^2\ [S_M(j\omega - j\omega_c) + S_M(j\omega + j\omega_c)], & |\omega| > \omega_c \\ 0, & |\omega| < \omega_c \end{cases}$$

- Lower sideband

$$S_{S_{lsb}}(j\omega) = \begin{cases} 0, & |\omega| > \omega_c \\ A_c^2\ [S_M(j\omega - j\omega_c) + S_M(j\omega + j\omega_c)], & |\omega| < \omega_c \end{cases}$$

An illustrative example of the power spectral density of single sideband modulation is shown in Figure 2.27.

The power can be calculated by the integral of the power spectral density. From Figure 2.27, it is obvious that the integral is equal than the integral for $S_M(j\omega)$, up to a factor A_c^2. Therefore, for both variants the power is

$$P_S = \frac{1}{2\pi} \int_{-\infty}^{\infty} S_S(j\omega)\ d\omega = A_c^2 P_M.$$

Demodulation of SSB signals

For optimal demodulation of single-sideband signals, it is necessary to use a synchronous or coherent demodulator, such as the one shown in the diagram in Figure 2.20, where $\phi = \phi_c$. Again

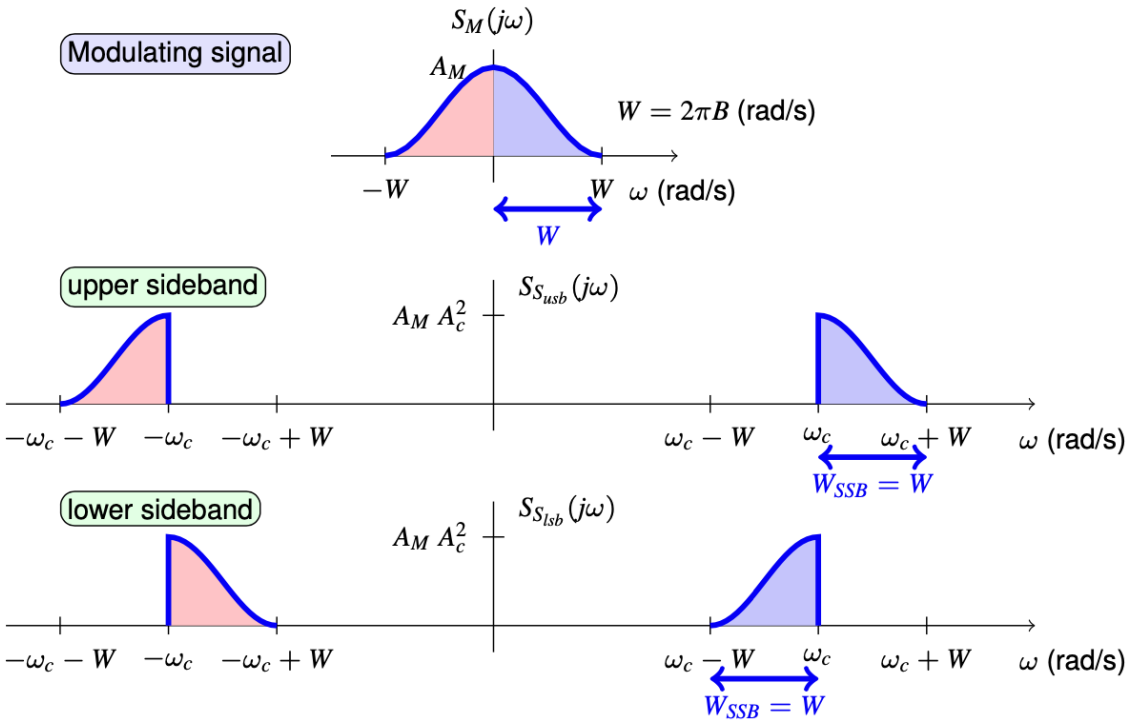


Figure 2.27: Power spectral density of signals modulated with single-sideband: upper-sideband, and lower-sideband amplitude modulation.

an ideal transmission is assumed without any distortion, in which case the received signal matches the modulated signal being transmitted.

$$r(t) = s(t) = A_c\, m(t)\ \cos(\omega_c t + \phi_c) \mp A_c\, \hat{m}(t)\ \sin(\omega_c t + \phi_c).$$

The modulated signal before filtering, $x(t)$, is

$$\begin{aligned} x(t) =& r(t) \times \cos(\omega_c t + \phi) \\ =& [A_c\, m(t)\ \cos(\omega_c t + \phi_c) \mp A_c\, \hat{m}(t)\ \sin(\omega_c t + \phi_c)] \times \cos(\omega_c t + \phi) \\ =& \frac{A_c}{2}\, m(t)\ \cos(\phi - \phi_c) \pm \frac{A_c}{2}\, \hat{m}(t)\ \sin(\phi - \phi_c) \\ &+ \frac{A_c}{2}\, m(t)\ \cos(2\omega_c t + \phi + \phi_c) \mp \frac{A_c}{2}\, \hat{m}(t)\ \sin(2\omega_c t + \phi + \phi_c). \end{aligned}$$

The filtering removes the high frequency components (terms in $2\omega_c$), so that the filtered demodulated signal is

$$d(t) = \frac{A_c}{2}\, m(t)\ \cos(\phi - \phi_c) \pm \frac{A_c}{2}\, \hat{m}(t)\ \sin(\phi - \phi_c).$$

Now, the effect of the phase error is, on the one hand, to reduce the amplitude of the received signal, which can even disappear, as it happened for a double sideband modulation. And on the other hand (second term), an unwanted signal, $\hat{m}(t)$, is added to the received signal, which can be interpreted as a distortion term that is proportional to the Hilbert transform of the modulating signal. Therefore, the use of a coherent receiver is even more necessary than in the case of DSB modulation.

The solution is the same as for the DSB, either to transmit a pilot tone or to use a PLL.

Figures 2.28 and 2.29 show a frequency interpretation of the demodulation process for the cases of upper sideband and lower sideband, respectively.

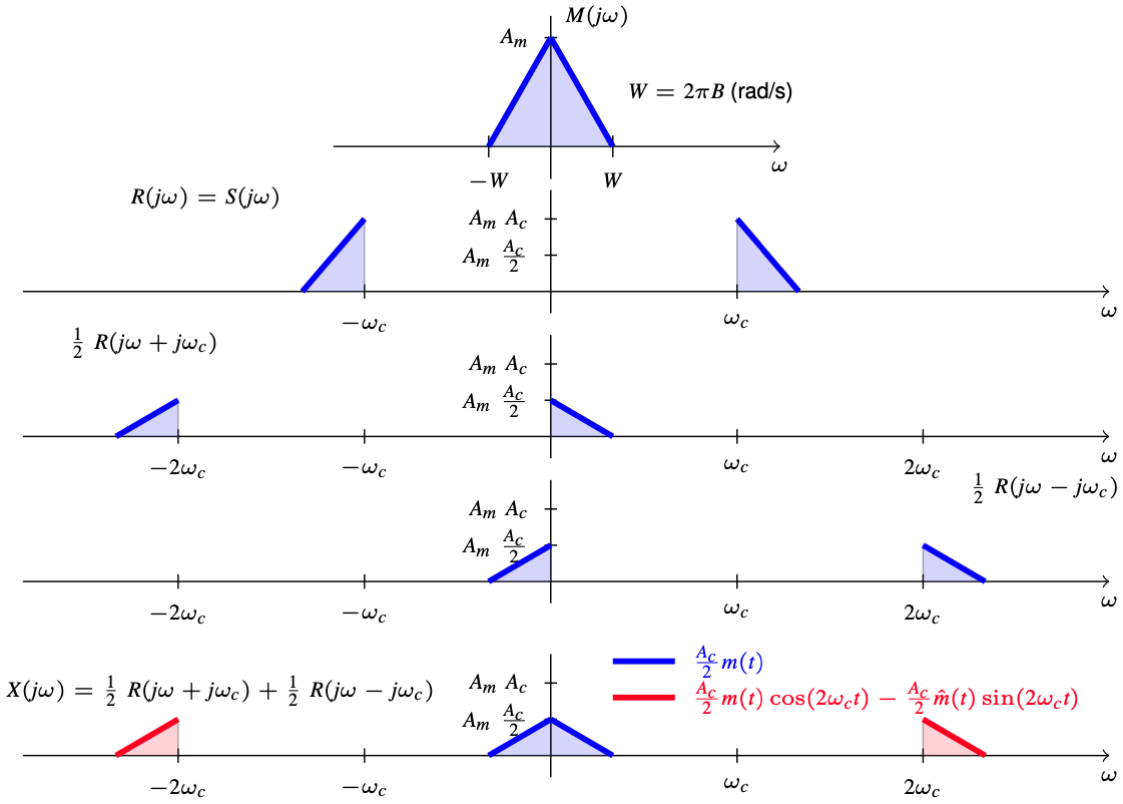


Figure 2.28: Frequency interpretation of the demodulation process of a single sideband signal (for upper sideband).

Due to the spectral efficiency of this modulation method, it is used, for example, for the transmission of voice channels over telephone cables. With this modulation the capacity, in number of channels, of the cable in question is doubled.

The main drawback is in its generation. Using the direct filtering technique, to avoid any distortion in the generated signal, ideal filters are required, which is not possible in practice. Using the technique based on the Hilbert filter, the drawback is that now two quadrature oscillators are needed, and also an exact implementation of a Hilbert transformer, which is also not possible. Therefore, in the signal generation process itself, a certain degree of distortion will usually be introduced.

2.2.4 Vestigial Sideband Modulation (VSB)

To relax the ideal filter requirements for SSB modulation, vestigial sideband modulation proposes to replace ideal filters by implementable filters: instead of frequency responses with an instantaneous transition from 0 to 1 (or from 1 to 0) at the carrier frequency, responses with progressive transition from 0 to 1 (or from 1 to 0) around the carrier frequency, with the transition in the interval $\pm\Delta_W$ rad/s around the carrier frequency, as it is shown in Figure 2.30. This type of filter allows a part of the band to be removed, called *vestige*, to remain in the modulated signal. In this way, the implementation of the filters is possible at the price of slightly increasing the bandwidth of the signal.

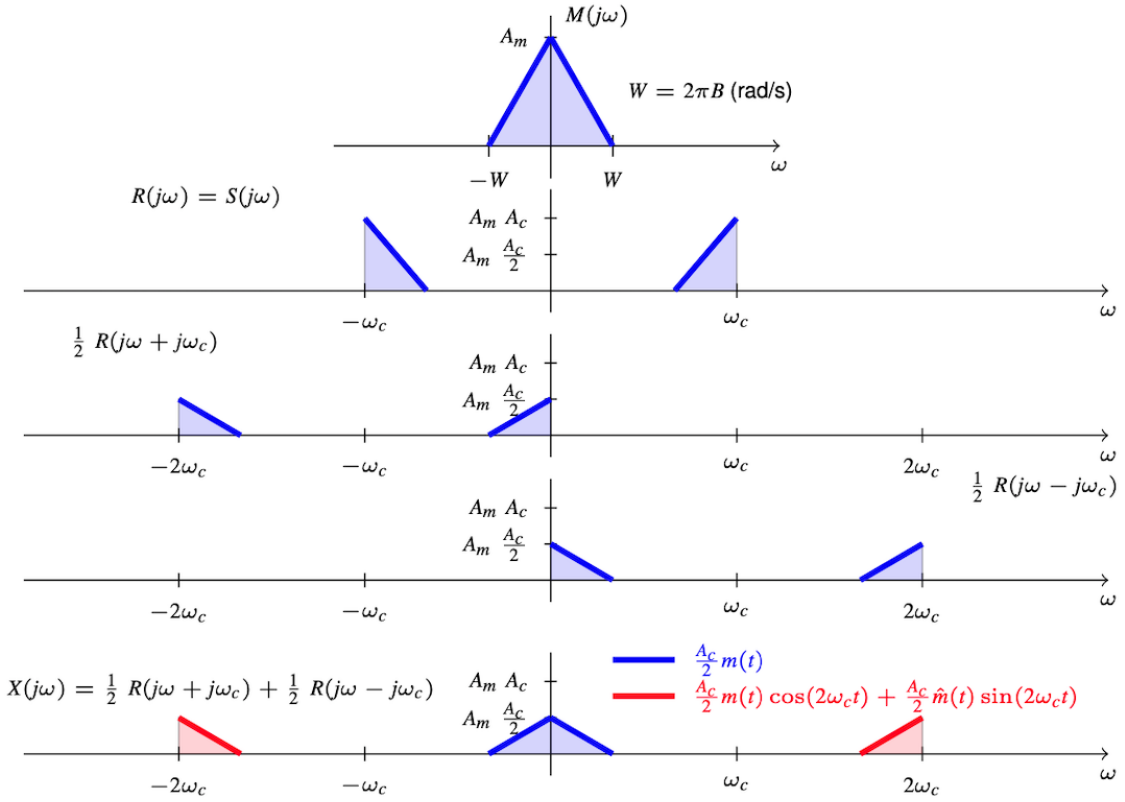


Figure 2.29: Frequency interpretation of the demodulation process of a single sideband signal (for lower sideband).

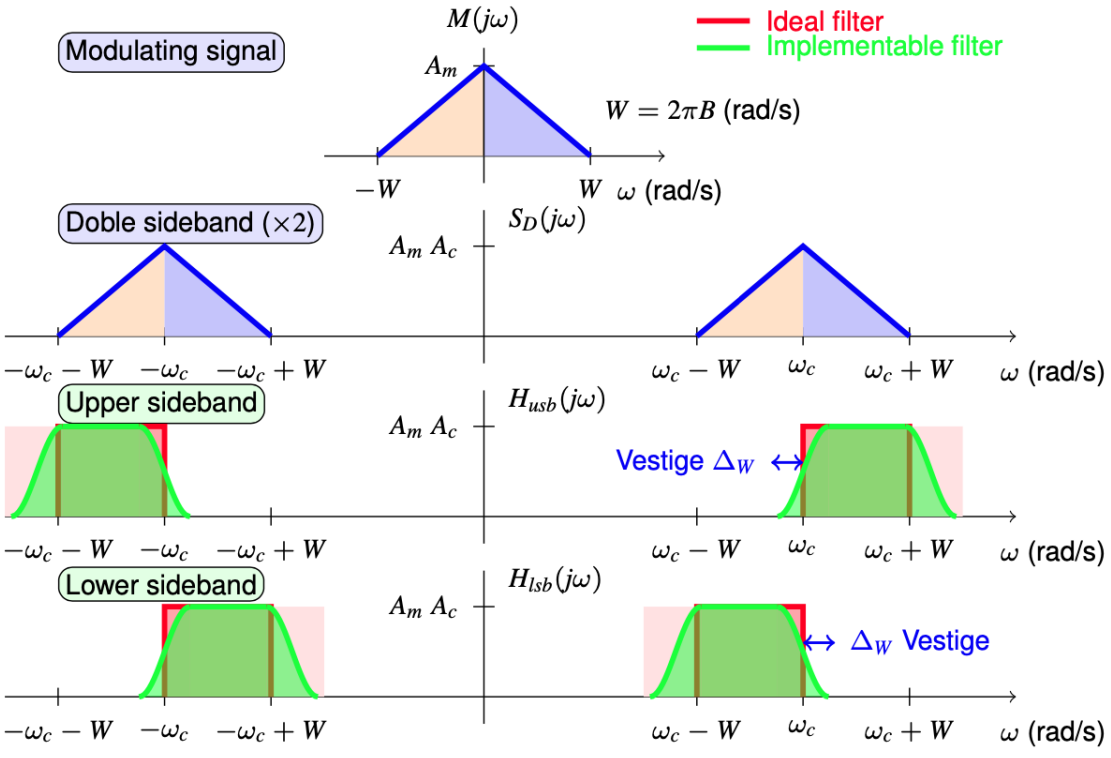


Figure 2.30: From SSB to VSB: from ideal filters to implementable filters.

To build this signal by direct filtering, similarly as in a SSB modulation first a double sideband signal (with double amplitude), $s_D(t)$, is generated, and then it is filtered, as shown in the Figure 2.31.

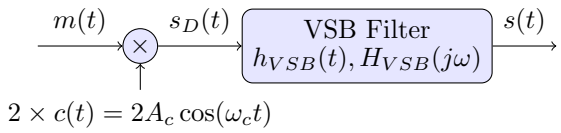


Figure 2.31: Generation of a vestigial sideband (VSB) signal using direct filtering.

The vestigial sideband modulated signal therefore has the analytical expression

$$s(t) = \left[\underbrace{m(t) \times 2A_c \cos(\omega_c t)}_{s_D(t)} \right] * h_{VSB}(t).$$

In the frequency domain, this expression corresponds to

$$S(j\omega) = A_c \ [M(j\omega - j\omega_c) + M(j\omega + \omega_c)] \ H_{VSB}(j\omega).$$

Let's analyze what happens in the receiver, if a synchronous receiver like the one in Figure 2.20 is used. It will be assumed that the signal has been transmitted without distortion, so the received signal will be equal to the transmitted VSB signal. In this case

$$R(j\omega) = S(j\omega) = A_c \ [M(j\omega - j\omega_c) + M(j\omega + \omega_c)] \ H_{VSB}(j\omega).$$

The demodulated (unfiltered) signal in the frequency domain is

$$x(t) = r(t) \ \cos(\omega_c t) \rightarrow X(j\omega) = \frac{1}{2} \ [R(j\omega - j\omega_c)) + R(j\omega + j\omega_c))].$$

$$\begin{aligned} X(j\omega) =& \frac{A_c}{2} \ [M(j\omega - j2\omega_c) + M(j\omega)] \ H_{VSB}(\omega - \omega_c) \\ &+ \frac{A_c}{2} \ [M(j\omega) + M(j\omega + j2\omega_c)] \ H_{VSB}(j\omega + j\omega_c). \end{aligned}$$

Finally, the filtered demodulated signal in the frequency domain is

$$D(j\omega) = \frac{A_c}{2} \ M(j\omega) \ [H_{VSB}(j\omega - j\omega_c) + H_{VSB}(j\omega + j\omega_c)].$$

It can be interpreted that this signal has been obtained by filtering the modulated signal with a filter with equivalent response

$$H_{EQ}(j\omega) = H_{VSB}(j\omega - j\omega_c) + H_{VSB}(j\omega + j\omega_c).$$

To avoid distortion, this joint response must have and ideal behavior in the bandpass of the signal, $|\omega| \leq W = 2\pi B$ rad/s; that is, its module must be constant, and its phase linear. From here we can obtain the condition that the vestigial sideband filter must satisfy

$$|H_{VSB}(j\omega - j\omega_c) + H_{VSB}(j\omega + j\omega)| = C \text{ in } |\omega| \leq 2\pi B \text{ rad/s}.$$

Ideally, the constant value would be $C = 1$. This condition is satisfied if the frequency response of the vestigial sideband filter has odd symmetry around ω_c in the frequency range $\omega_c - \Delta_W < \omega < \omega_c + \Delta_W$, where Δ_W is the bandwidth excess (vestige) in radians/s. In that case, the bandwidth of the modulated signal is

$$B_{VSB} = B + \Delta_B \text{ Hz, with } \Delta_B = \frac{\Delta_W}{2\pi} \text{ Hz,}$$

where Δ_B is the vestige or bandwidth excess in Hz (usually $\Delta_B << B$).

Figure 2.32 shows an example of filters for upper sideband and lower sideband. Note that the response of the filter out of the frequency range of the double sideband signal (above $\omega_c + W$ or below $\omega_c - W$)) is irrelevant, because the filtered signal has no spectral components there.

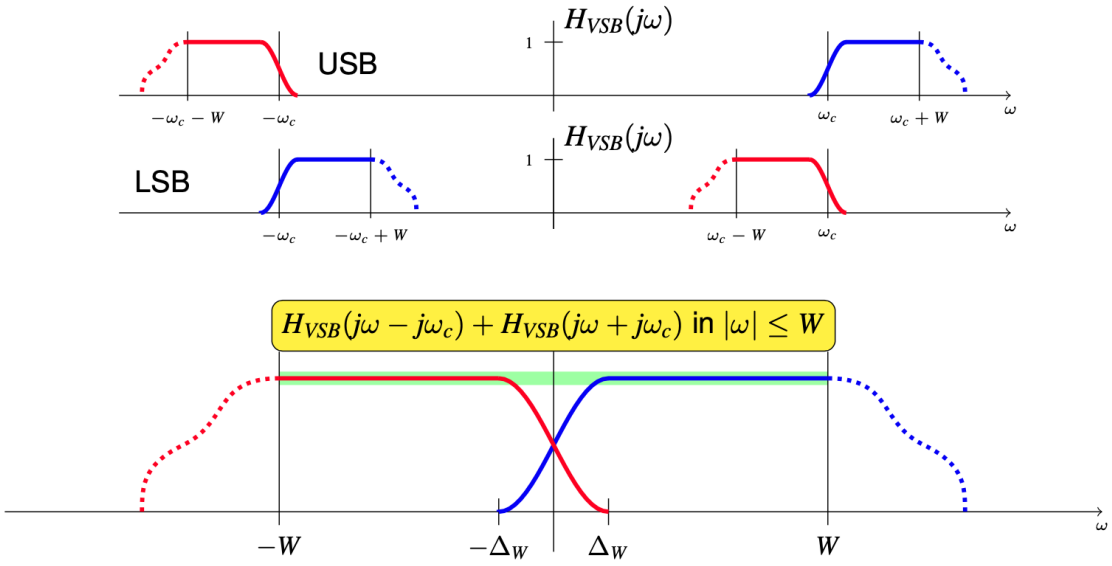


Figure 2.32: Examples of vestigial sideband filters for upper sideband and lower sideband.

Demodulation of VSB signals

If the filters satisfy the specified condition, it is possible to recover the signal without distortion using a synchronous receiver. Figure 2.33 shows the frequency interpretation of the modulation and demodulation process for an upper sideband VSB signal.

2.2.5 Summary of characteristics and comparison between the different amplitude modulations

Table 2.1 shows the main characteristics of the different amplitude modulations that have been studied. In particular, reference is made to the power of the modulated signal, and how much of it is related to the information transmitted, and to the consumption of bandwidth. If the power efficiency and the spectral efficiency are considered, the conclusions that can be drawn are:

- Power efficiency

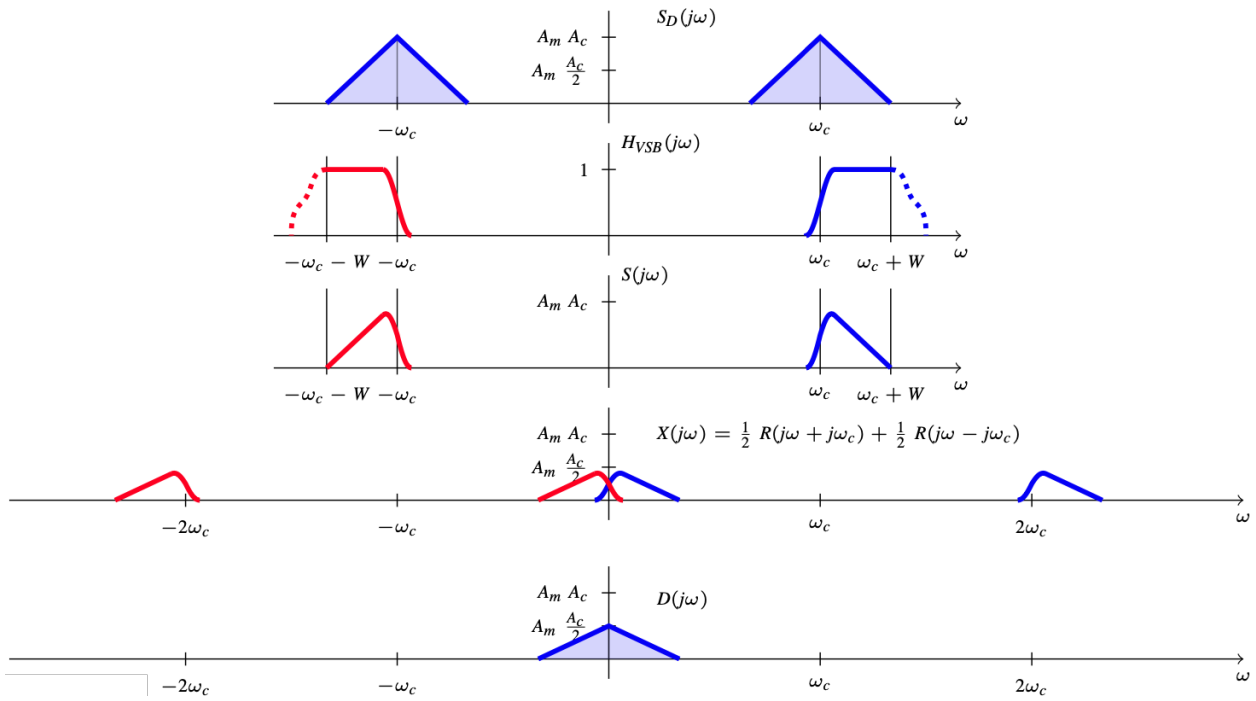


Figure 2.33: Frequency interpretation of the modulation and demodulation process for a vestigial sideband modulation (upper sideband).

Modulation	BW (Hz)	P_S	$P_S(m(t))$	$d(t)$	$P_d(m(t))$
Conv. AM	$2B$	$\frac{A_c^2}{2}\left[1 + P_{M_a}\right]$	$\frac{A_c^2}{2}P_{M_a}$	$\frac{A_c}{2}\left[1 + m_a(t)\right]$	$\frac{A_c^2}{4}P_{M_a}$
DSB	$2B$	$\frac{A_c^2}{2}P_M$	$\frac{A_c^2}{2}P_M$	$\frac{A_c}{2}m(t)$	$\frac{A_c^2}{4}P_M$
SSB	B	$A_c^2 P_M$	$A_c^2 P_M$	$\frac{A_c}{2}m(t)$	$\frac{A_c^2}{4}P_M$
VSB	$B + \Delta_B$	$A_c^2 P_M$	$A_c^2 P_M$	$\frac{A_c}{2}m(t)$	$\frac{A_c^2}{4}P_M$

BW (Hz): bandwidth of the modulated signal, in Hz
P_S: power of the modulated signal
$P_S(m(t))$: power of the modulated signal that is related to $m(t)$
$d(t)$: signal recovered with a coherent or synchronous receiver
$P_d(m(t))$: power of the demodulated signal that is related to $m(t)$

Table 2.1: Comparison of amplitude modulations.

  - If it is understood as modulations where all the power of the signal is related to the transmission of information (signal $m(t)$)
    * DSB, SSB, VSB

- Spectral efficiency
  - Minimum transmission bandwidth (same bandwidth as the modulating signal, B Hz)
    * SSB and VSB (in this case with a vestigial increment Δ_B)

Regarding the data corresponding to the demodulator that appear in the table, it has always been considered that a synchronous or coherent receiver has been used. As in the case of conventional AM modulation this receptor had not been studied, for completeness, it is included below.

Synchronous detection of a conventional AM modulation

In this section we study the demodulation of a conventional AM signal with a receiver like the one in Figure 2.20. As in the previous cases, it is assumed that the modulated signal is transmitted without distortion, so the received signal is equal to the modulated signal that has been transmitted, which in the case of conventional AM modulation is

$$r(t) = s(t) = A_c \left[1 + m_a(t)\right] \cos(\omega_c t + \phi_c)$$

The unfiltered demodulated signal $x(t)$ is then

$$\begin{aligned} x(t) =& r(t) \times \cos(\omega_c t + \phi) \\ =& A_c \left[1 + m_a(t)\right] \cos(\omega_c t + \phi_c) \times \cos(\omega_c t + \phi) \\ =& \frac{A_c}{2} \left[1 + m_a(t)\right] \cos(\phi_c - \phi) + \frac{A_c}{2} \left[1 + m_a(t)\right] \cos(2\omega_c t + \phi_c + \phi) \end{aligned}$$

Low-pass filtering removes the high-frequency terms (terms in $2\omega_c$), so the filtered demodulated signal is then

$$d(t) = \frac{A_c}{2} \left[1 + m_a(t)\right] \cos(\phi_c - \phi).$$

Therefore, the suitability of a synchronous or coherent demodulator is also revealed for this modulation, with $\phi = \phi_c$, in which case the demodulated signal is

$$d(t) = \frac{A_c}{2} \left[1 + m_a(t)\right].$$

2.3 Angle Modulations

In the previous section we have seen the amplitude modulation of a carrier, in which the amplitude of the carrier is modified as a function of the modulating signal, $m(t)$. Amplitude modulation methods are also called *linear modulations*, because the amplitude of the modulated signal has a linear relationship with $m(t)$.

Frequency (FM) and phase (PM) modulations are called *angle modulations*. In this case, the frequency or phase of the carrier signal is modified to be dependent on the modulating signal. These modulations are clearly non-linear, which implies a series of properties:

1. They are more difficult to implement (both modulator and demodulator).
2. Their analysis is generally more complex. In many cases, it is only possible to perform the analysis using approximations.

On the other hand, angle modulations expand the spectrum so that the bandwidth of the signal is several times greater than that of the modulating signal. In a strict sense, these signals have an infinite bandwidth, but normally we work with the so-called *effective bandwidth*, which is the one in which the amplitude of the spectrum is relevant.

The reason why angle modulations are used, despite these disadvantages, is that they have the advantage of high noise immunity: These modulations "trade" bandwidth for noise immunity. For this reason, FM modulation is used in high-fidelity music broadcast systems or in point-to-point communication systems where the transmission power is limited.

2.3.1 Representation of FM and PM signals

Both PM and FM modulations modify, depending on the message signal, the argument of a sinusoidal carrier. Therefore, it is possible to make a joint analysis of the two modulations and, as will be seen later, there is a clear relationship between them.

In general, an angle modulation can be represented mathematically as

$$s(t) = A_c \cos(\theta(t)),$$

where taking into account that it is a signal generated from a carrier of frequency f_c Hz ($\omega_c = 2\pi f_c$ rad/s), the angle $\theta(t)$ in general can be written as

$$\theta(t) = 2\pi f_c t + \phi(t) = \omega_c t + \phi(t).$$

This is a joint representation for PM and FM modulations. The argument $\theta(t)$ is the phase of the signal at time t, and the instantaneous frequency of the signal in Hz, $f_i(t)$, is given by its derivative

$$f_i(t) = \frac{1}{2\pi}\frac{d}{dt}\theta(t).$$

For the particular expression above

$$f_i(t) = f_c + \frac{1}{2\pi}\frac{d}{dt}\phi(t).$$

If $m(t)$ is the modulating signal (message to be transmitted), the relationship of the angle argument of $s(t)$ with the modulating signal is as follows for each of the two types of modulation:

- Phase Modulation (PM)

  $$\phi(t) = k_p\, m(t)$$

  k_p: phase deviation constant

- Frequency Modulation (FM)

  $$\Delta f_i(t) = f_i(t) - f_c = \frac{1}{2\pi}\,\frac{d}{dt}\phi(t) = k_f\, m(t)$$

  k_f: frequency deviation constant

Constants k_p and k_f are the phase and frequency *deviation constants*, respectively.

In a PM modulation, the phase term $\phi(t)$ is proportional to the modulating signal, and in an FM modulation, the diference between the instantaneous frequency and the carrier frequency is proportional to the modulating signal. In this case, given the definition of the instantaneous frequency, the derivative of $\phi(t)$ is proportional to $m(t)$ (or equivalently, $\phi(t)$ is proportional to the integral of $m(t)$).

There is a close relationship between both modulation methods, which can be easily seen by writing the expressions for $\phi(t)$ and $\frac{d}{dt}\phi(t)$ in PM and FM

$$\phi(t) = \begin{cases} k_p\, m(t), & \text{PM} \\ 2\pi\, k_f \int_{-\infty}^{t} m(\tau)\, d\tau, & \text{FM} \end{cases}$$

$$\frac{d}{dt}\phi(t) = \begin{cases} k_p\, \frac{d}{dt} m(t), & \text{PM} \\ 2\pi\, k_f\, m(t), & \text{FM} \end{cases}$$

It can be seen that the FM modulation of $m(t)$ is equivalent to the PM modulation of the integral of $m(t)$. Similarly, to modulate $m(t)$ with a PM modulation is equivalent to modulate the derivative of $m(t)$ with an FM modulation. Figure 2.34 represents the relationship between PM and FM modulations.

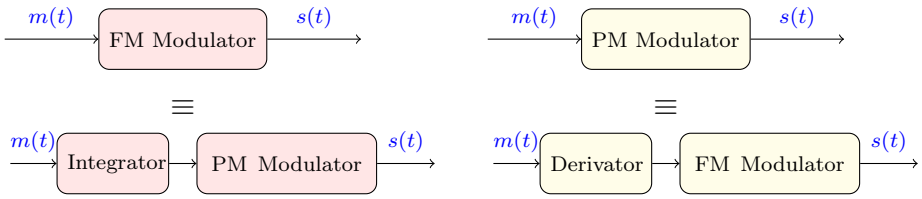


Figure 2.34: Relationship between FM and PM modulations.

This relationship between both types of modulation allows us to jointly analyze them and then highlight their differences.

The waveform of an angle modulation is that of a sinusoid of constant amplitude, whose angular information changes over time, which visually translates into a change in the distance between zero crossings of the signal. Figure 2.35 shows an example of a modulating signal, the carrier signal, and the resulting PM modulated signal.

The effect of the frequency or phase deviation constants is to weight the degree of deviation of the modulated signal with respect to the carrier signal of constant frequency and phase. Figures 2.36 and 2.37 show the modulated signal for a PM modulation for two different values of phase deviation constant, $k_p = 2\pi \times \frac{1}{4}$ and $k_p = 2\pi \times \frac{3}{4}$, respectively.

It can be seen how with respect to the carrier signal, when the modulating signal takes positive values, the modulated signal is shifted before the carrier (by increasing the phase term proportionally to $m(t)$), and it is shifted after the carrier when $m(t)$ takes negative values (by reducing the phase term in this case). The magnitude of the advance or delay at a given instant is proportional to the parameter k_p.

The waveforms are similar in a frequency modulation, and the effect of the frequency deviation constant k_f is similar to the effect of k_p in a PM modulation. Figures 2.38 and 2.39 show the

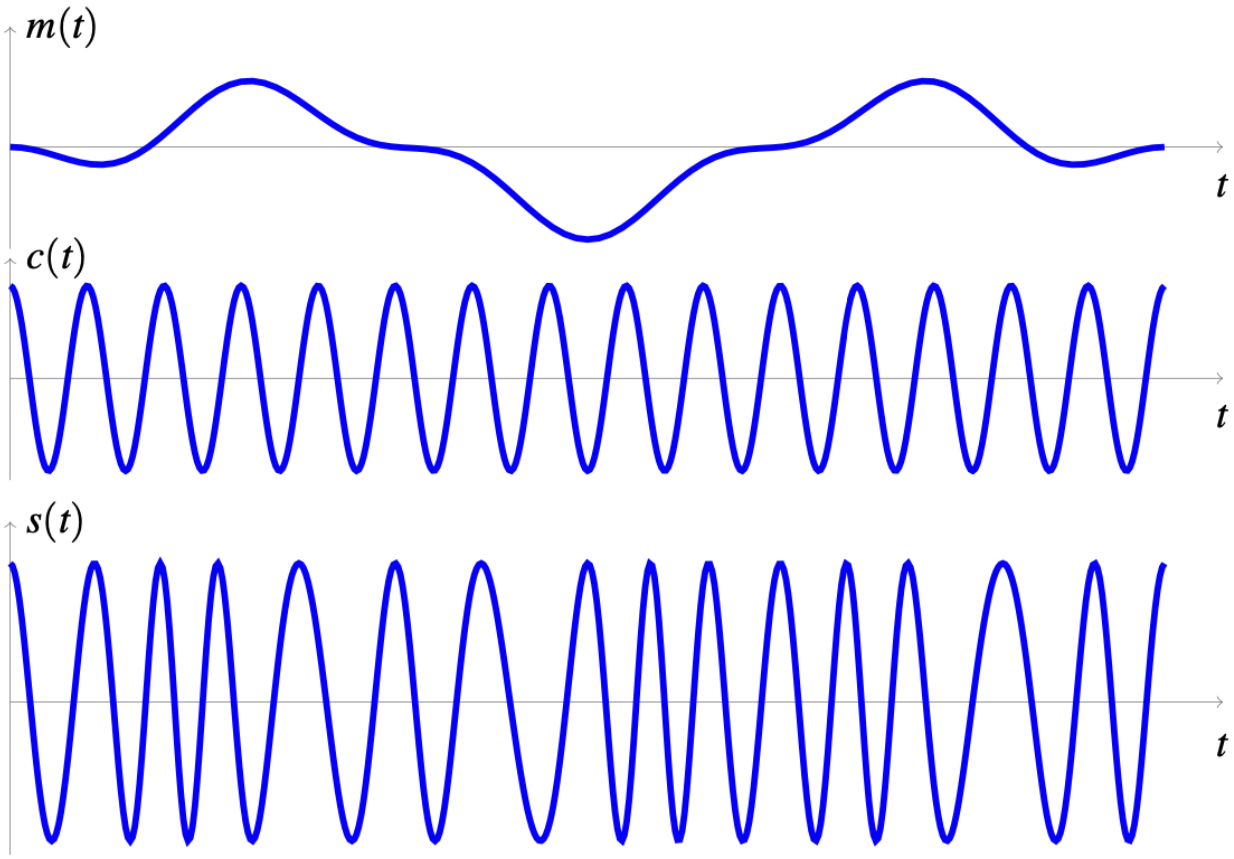


Figure 2.35: Waveform of an angle modulation (in this case PM) for an example of a modulating signal.

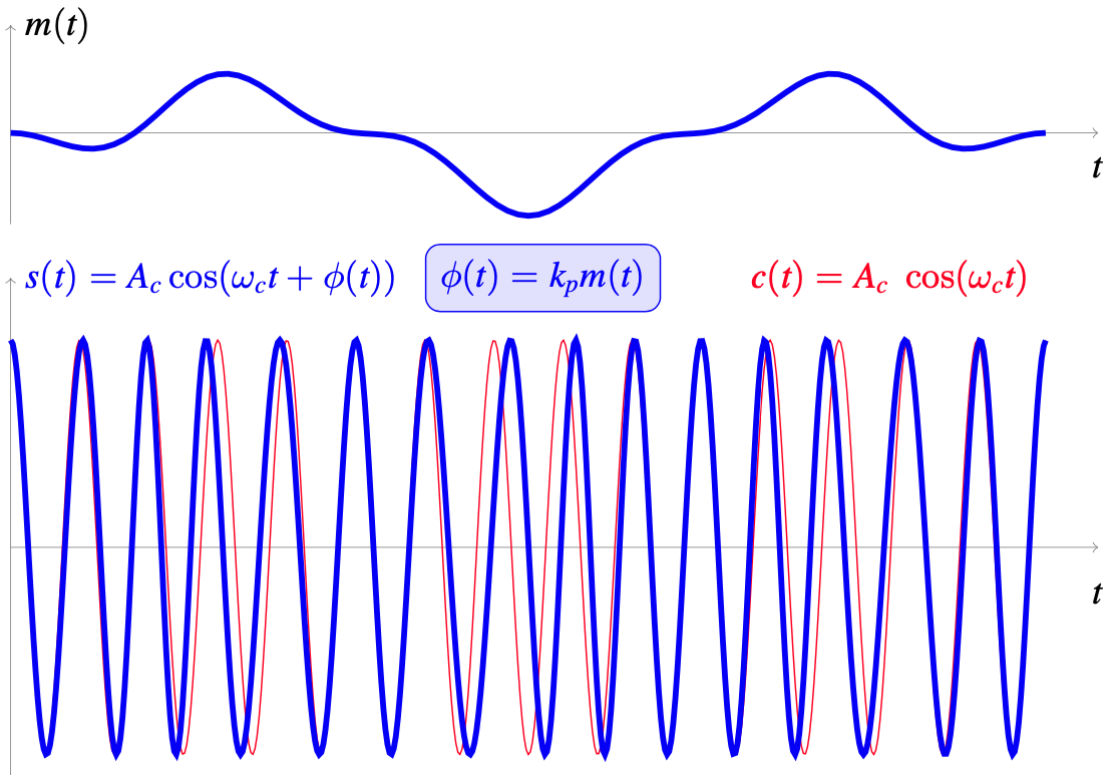


Figure 2.36: Modulated signal for a PM modulation with $k_p = 2\pi \times \frac{1}{4}$ for an example of a modulating signal.

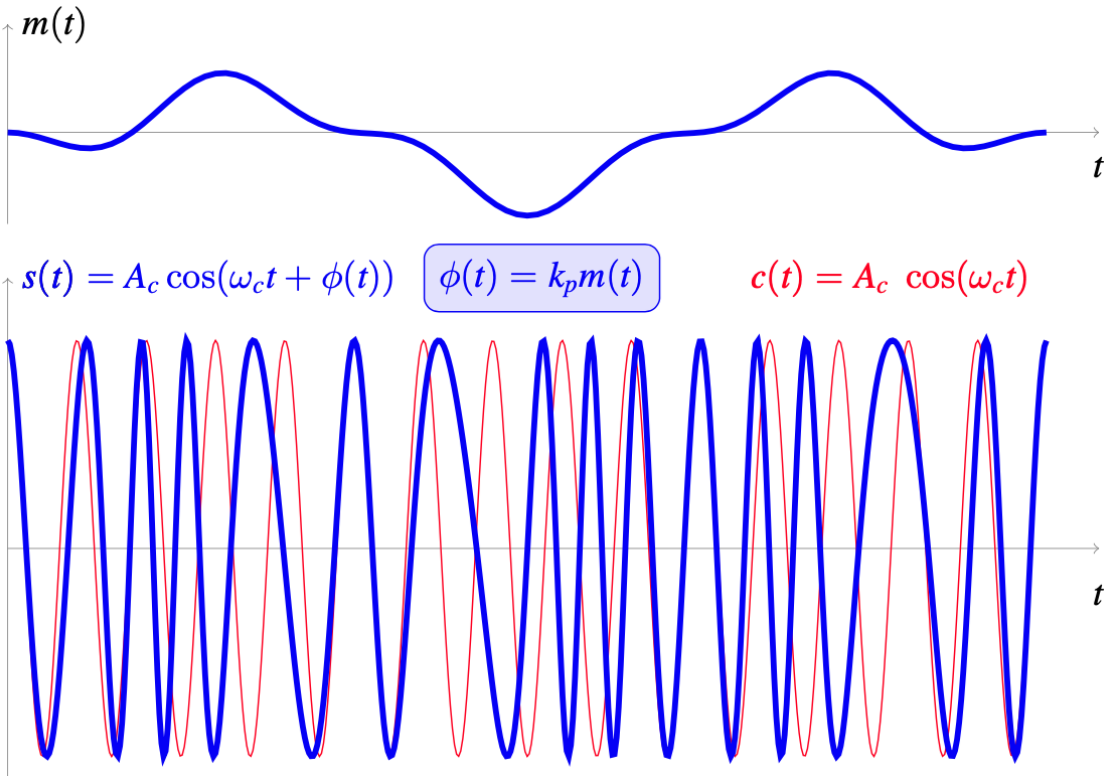


Figure 2.37: Modulated signal for a PM modulation with $k_p = 2\pi \times \frac{3}{4}$ for an example of a modulating signal.

modulated signal for an FM phase modulation for two different values of the frequency deviation constant, $k_f = 2\pi \times \frac{1}{4}$ and $k_f = 2\pi \times \frac{3}{4}$, respectively.

Now, the modulated signal is shifted before the carrier when the integral of the signal is positive, and it is shifted after the carrier when the integral is negative, since now the phase term $\phi(t)$ is proportional to the integral of the modulating signal. Again, the magnitude of the deviation with respect to the carrier at an instant is proportional to the value of the deviation constant, in this case the frequency deviation constant, k_f.

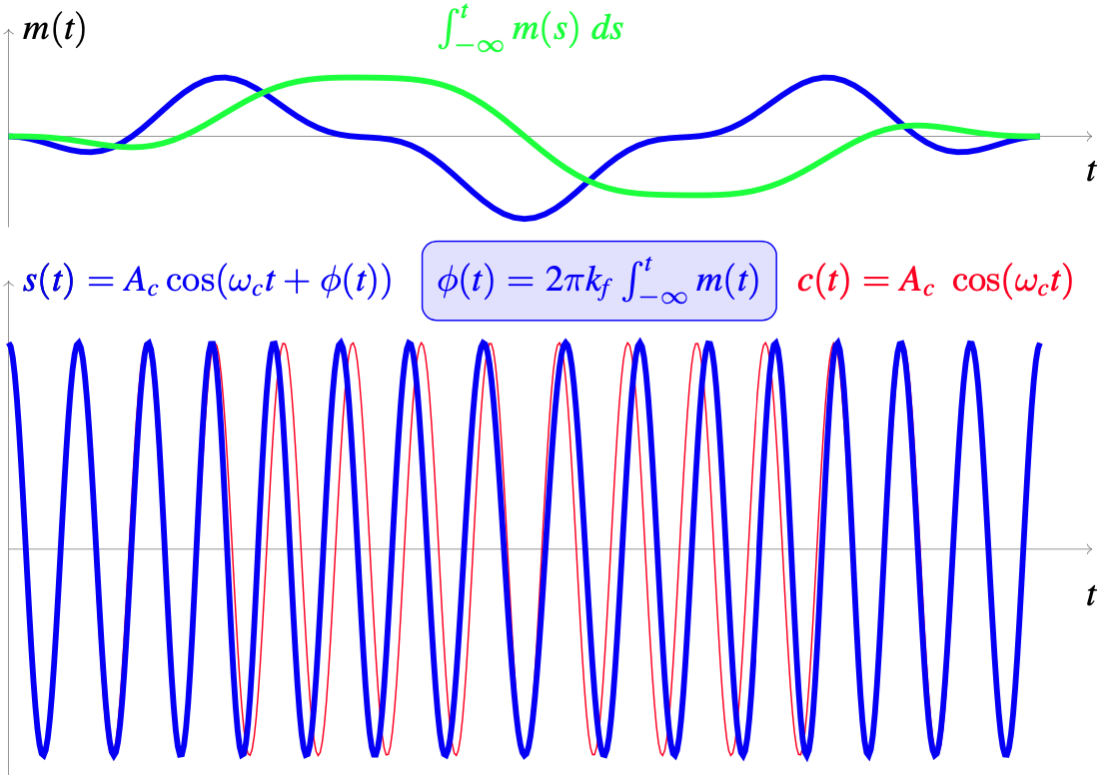


Figure 2.38: Modulated signal for an FM modulation with $k_f = 2\pi \times \frac{1}{4}$ for an example of a modulating signal.

2.3.2 Modulation indices

An important parameter of an angle modulation is the modulation index, since various aspects of modulation, such as bandwidth or noise immunity, depend on its value. The modulation indices are defined from the phase or frequency deviation constants, respectively. In a phase modulated signal, the maximum phase deviation of the signal is

$$\Delta\phi_{\max} = k_p \ \max(|m(t)|).$$

Similarly, in a frequency modulated signal, the maximum frequency deviation of the signal is

$$\Delta f_{\max} = k_f \ \max(|m(t)|).$$

From these maximum deviations, the modulation indices of a PM modulation and an FM modulation are defined, respectively, as

$$\beta_p = \Delta\phi_{\max} = k_p \ \max(|m(t)|) = k_p \, C_M$$

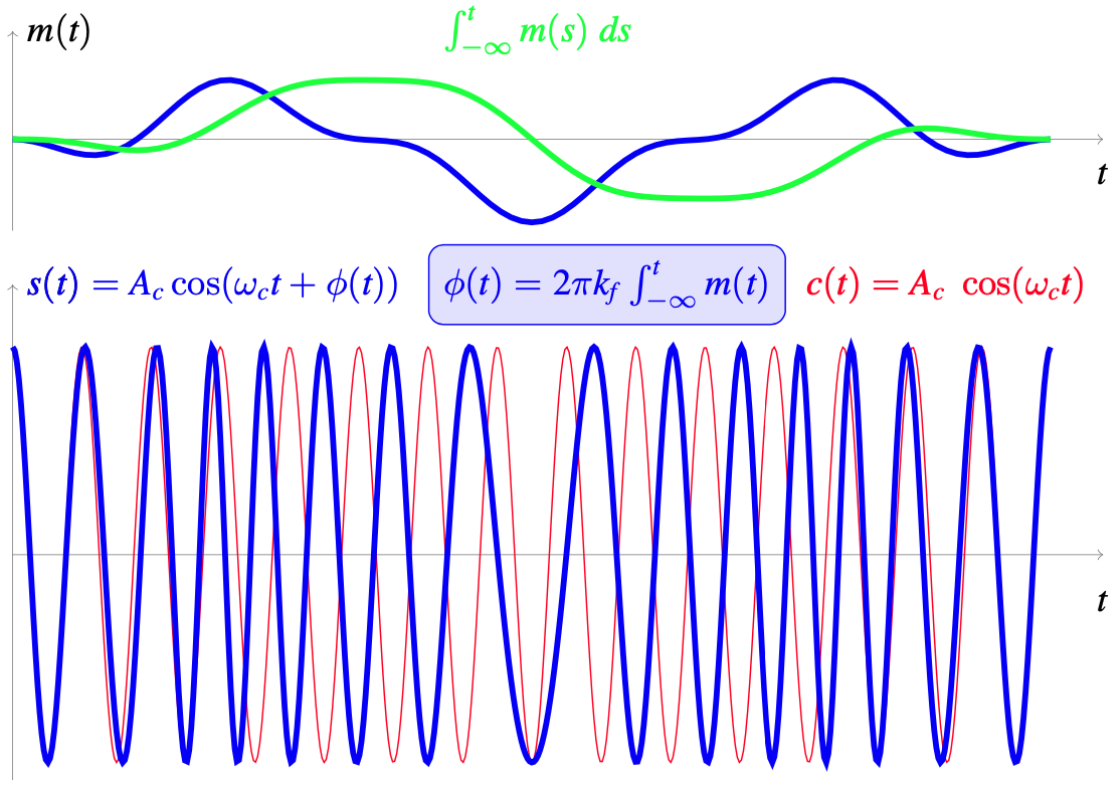


Figure 2.39: Modulated signal for an FM modulation with $k_f = 2\pi \times \frac{3}{4}$ for an example of a modulating signal.

and

$$\beta_f = \frac{\Delta f_{\max}}{B} = \frac{k_f \max(|m(t)|)}{B} = \frac{k_f\, C_M}{B},$$

where B is the bandwidth in Hz of the modulating signal $m(t)$, and C_M is the maximum value of its module, which defines its range: $-C_M \leq m(t) \leq +C_M$.

2.3.3 Spectral characteristics of an angle modulation

Due to the non-linearity of the angle modulations, in many cases the precise characterization of their spectrum cannot be treated in a strict mathematical way. Normally it is studied for simple modulating signals and certain approximations are made. Following are some particular cases: narrow band modulation, modulations by means of a sinusoidal signal and by means of a periodic signal, and finally the case of an arbitrary non-periodic modulating signal will be discussed.

Narrowband angle modulation

An angle modulation is narrowband if the constants k_p or k_f and the signal $m(t)$ are such that

$$\phi(t) << 1.$$

This implies small values of the deviation constants (and therefore, small value of the modulation index). In this case, since the modulated signal has the general expression

$$s(t) = A_c\ \cos(\omega_c t + \phi(t)),$$

and taking into account the trigonometric relation

$$\cos(A \pm B) = \cos(A)\cos(B) \mp \sin(A)\sin(B),$$

we can expand the expression of the modulated signal and make the following approximation

$$\begin{aligned} s(t) =& A_c\ \cos(\omega_c t)\ \cos\phi(t) - A_c\ \sin(\omega_c t)\ \sin\phi(t) \\ \approx& A_c\ \cos(\omega_c t) - A_c\ \phi(t)\ \sin(\omega_c t). \end{aligned}$$

Here, we have considered that for small values of $\phi(t)$

$$\cos(\phi(t)) \approx 1, \text{ and } \sin(\phi(t)) \approx \phi(t).$$

This equation is very similar to the expression of a conventional AM signal, where a negative sign and a sine appear instead of a cosine, and instead of the signal $m(t)$ we have the phase $\phi(t)$, which for a PM is proportional to $m(t)$ and for an FM it is proportional to the integral of $m(t)$. Remember that for a conventional AM

$$\text{Conventional AM: } s(t) = A_c\ \cos(\omega_c t) + A_c\ m_a(t)\ \cos(\omega_c t),$$

and that in angle modulations

$$\phi(t) = \begin{cases} k_p\ m(t) & \text{PM} \\ 2\pi\ k_f\ \int_{-\infty}^{t} m(\tau)\ d\tau & \text{FM} \end{cases}.$$

For this type of modulation, the spectrum of the signal is very similar to that of an AM signal, at least in terms of its bandwidth. It must be taken into account that the bandwidth of the integral of a signal is the same as that of the signal itself (the frequency response of a derivative is $j\omega$). Therefore, the spectrum will have the following components and characteristics:

- Two deltas, located at $\pm\omega_c$ (spectrum of the carrier)
- Two replicas of the spectrum of $\phi(t)$ located at $\omega = \pm\omega_c$
- With respect to the shape of the spectrum of $\phi(t)$
  - PM: proportional to the spectrum of $m(t)$

$$\phi(t) = k_p\ m(t) \leftrightarrow \Phi(j\omega) = k_p\ M(j\omega)$$

  - FM: proportional to the spectrum of the integral of $m(t)$

$$\phi(t) = 2\pi\ k_f\ \int_{-\infty}^{t} m(\tau)\ d\tau \leftrightarrow \Phi(j\omega) = 2\pi k_f\ \frac{M(j\omega)}{j\omega}$$

Thus, the bandwidth in Hz of these narrowband signals if the modulating signal has a bandwidth B Hz is

$$B_{NB} \approx 2B \text{ Hz}.$$

Modulation using a sinusoidal signal

In this section, the case in which the modulating signal is a sinusoidal signal, with amplitude a and frequency ω_m rad/s, will be analyzed. For convenience, in order to simultaneously analyze PM and FM modulations, the analytical expression of the modulating signal is

$$m(t) = \begin{cases} a\ \sin(\omega_m t) & \text{for a PM modulation} \\ a\ \cos(\omega_m t) & \text{for an FM modulation} \end{cases}$$

In this way, the analytical expression of the modulated signal, $s(t)$, coincides for the two variants. The modulation indices of a PM and FM modulation are

$$\beta_p = \Delta\phi_{\max} = k_p\ \max(|m(t)|) = k_p\ C_M = k_p\ a$$

$$\beta_f = \frac{\Delta f_{\max}}{B} = \frac{k_f\ \max(|m(t)|)}{B} = \frac{k_f\ C_M}{B} = k_f\ a\ \frac{2\pi}{\omega_m}$$

Therefore, the expressions of the phase term $\phi(t)$ are

- Expressions of $\phi(t)$ for PM

$$\phi(t) = k_p\ m(t) = k_p\ a\ \sin(\omega_m t) = \beta_p \sin(\omega_m t)$$

- Expressions of $\phi(t)$ for FM

$$\phi(t) = 2\pi\ k_f \int_{-\infty}^{t} m(\tau)\ d\tau = 2\pi\ k_f\ a\ \frac{1}{\omega_m}\ \sin(\omega_m t) = \beta_f \sin(\omega_m t)$$

Which means that the expression of the modulated signal is common for both types of modulation

$$s(t) = A_c \cos(\omega_c t + \phi(t)) = A_c\ \cos(\omega_c t + \beta\ \sin(\omega_m t)),$$

where β is the modulation index of the corresponding modulation, β_p or β_f. Since a cosine is the real part of a complex exponential, the modulated signal can be rewritten as

$$s(t) = Re\left(A_c\ e^{j\omega_c t}\ e^{j\beta\ \sin(\omega_m t)}\right).$$

The function $e^{j\beta\ \sin(\omega_m t)}$ is periodical, with frequency $f_m = \frac{\omega_m}{2\pi}$ Hz. This implies that it admits a Fourier series expansion, of the form

$$e^{j\beta\ \sin(\omega_m t)} = \sum_{n=-\infty}^{\infty} J_n(\beta)\ e^{j(n\ \omega_m)t}.$$

The n-th coefficient of the series expansion, denoted here as $J_n(\beta)$, is in this case the Bessel function of the first kind of order n and argument β. These coefficients are very well known. A series expansion of the Bessel function is

$$J_n(\beta) = \sum_{k=0}^{\infty} \frac{(-1)^k \left(\frac{\beta}{2}\right)^{n+2k}}{k!(k+n)!}.$$

For small values of β, the following approximation can be used

$$J_n(\beta) \approx \frac{\beta^n}{2^n n!}.$$

It can be seen that for small values of β only the first few terms are significant (in general it can be limited to $n = 1$ or $n = 2$).

The Bessel functions also satisfy the following symmetry properties

$$J_{-n}(\beta) = \begin{cases} J_n(\beta), & n \text{ even} \\ -J_n(\beta), & n \text{ odd} \end{cases}.$$

Figure 2.40 and Table 2.2 show the Bessel functions, as a function of β, for various values of n.

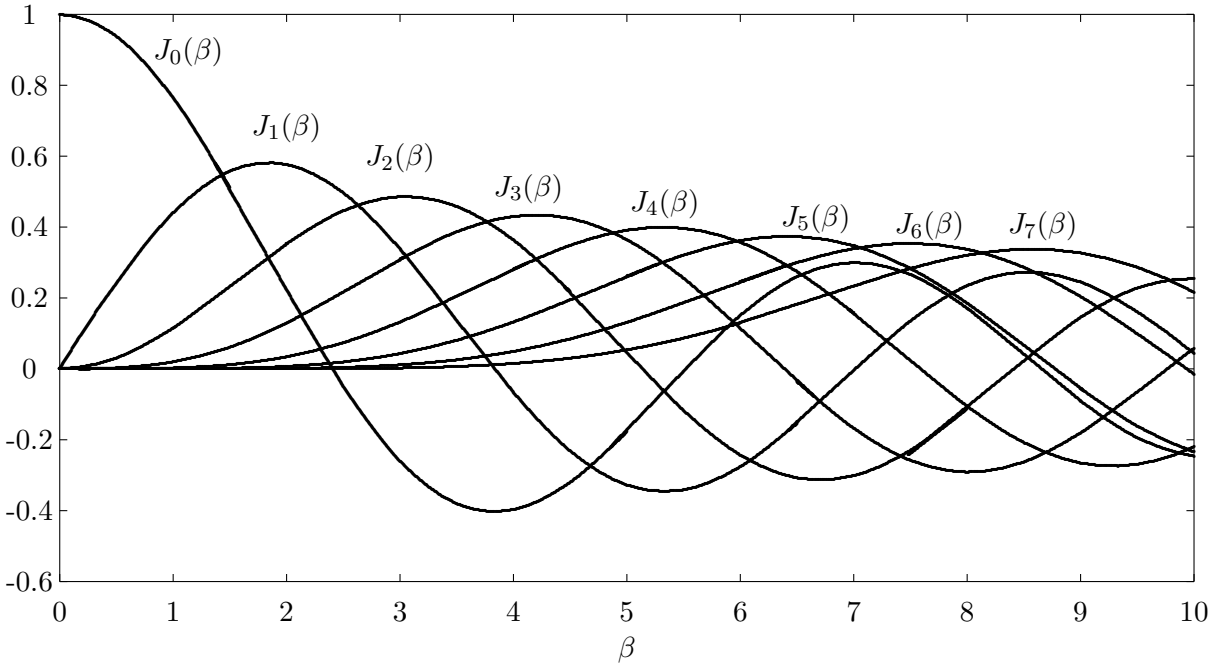


Figure 2.40: Bessel functions, $J_n(\beta)$, for different values of n.

Using the Fourier series expansion, an alternative expression of the modulated signal can be obtained

$$\begin{aligned} s(t) =& Re\left(A_c\, e^{j\omega_c t} \sum_{n=-\infty}^{\infty} J_n(\beta)\, e^{j(n\,\omega_m)t} \right) = Re\left(\sum_{n=-\infty}^{\infty} A_c\, J_n(\beta)\, \underbrace{e^{j\omega_c t}\, e^{j(n\,\omega_m)t}}_{e^{j(\omega_c + n\,\omega_m)t}} \right) \\ =& \sum_{n=-\infty}^{\infty} A_c\, J_n(\beta)\, \cos\left((\omega_c + n\,\omega_m)\, t\right) \end{aligned}$$

The modulated signal can be expressed as an infinite sum of sinusoids (cosines) with the following characteristics (bearing in mind that the index of the sum is n):

- Frequencies of the sinusoidals

$$\text{Frequencies (Hz)} : f_c + n\, f_m, \quad \text{para } n = 0, \pm 1, \pm 2, \cdots$$

$$\text{Angular Freq. (rad/s)} : \omega_c + n\, \omega_m, \quad \text{para } n = 0, \pm 1, \pm 2, \cdots$$

- Amplitudes of the sinusoidals $\omega_c + n\, \omega_m$

$$A_c\, J_n(\beta)$$

n	$\beta=0.1$	$\beta=0.2$	$\beta=0.5$	$\beta=1$	$\beta=2$	$\beta=5$	$\beta=8$	$\beta=10$
0	0.9975	0.9900	0.9385	0.7652	0.2239	−0.1776	0.1717	−0.2459
1	0.0499	0.0995	0.2423	0.4401	0.5767	−0.3276	0.2346	0.0435
2	0.0012	0.0050	0.0306	0.1149	0.3528	0.0466	−0.1130	0.2546
3		0.0002	0.0026	0.0196	0.1289	0.3648	−0.2911	0.0584
4			0.0002	0.0025	0.0340	0.3912	−0.1054	−0.2196
5				0.0002	0.0070	0.2611	0.1858	−0.2341
6					0.0012	0.1310	0.3376	−0.0145
7					0.0002	0.0534	0.3206	0.2167
8						0.0184	0.2235	0.3179
9						0.0055	0.1263	0.2919
10						0.0015	0.0608	0.2075
11						0.0004	0.0256	0.1231
12						0.0001	0.0096	0.0634
13							0.0033	0.0290
14							0.0010	0.0120
15							0.0003	0.0045
16							0.0001	0.0016

Table 2.2: Table with values of the Bessel functions $J_n(\beta)$.

This means that the bandwidth is theoretically infinite. However, the bandwidth of the modulated signal is not infinite, since the amplitude of the components for high values of n is very small. Therefore a finite *effective bandwidth* is defined. In general, the effective bandwidth is defined as that which contains at least 98% of the signal power. In this case, this effective bandwidth is

$$B_e = 2\,(\beta+1)\,f_m \text{ Hz}.$$

Let's see how the modulating signal affects the spectrum of the signal. In this case, assuming the same sinusoidal modulating signals of amplitude a and frequency f_m

$$B_e = 2(\beta+1)f_m = \begin{cases} 2(k_p a+1)f_m, & \text{PM} \\ 2\left(\frac{k_f a}{f_m}+1\right)f_m, & \text{FM} \end{cases},$$

or equivalently

$$B_e = 2(\beta+1)f_m = \begin{cases} 2(k_p a+1)f_m, & \text{PM} \\ 2\left(k_f a+f_m\right), & \text{FM} \end{cases}.$$

On the other hand, the total number of harmonics in the effective bandwidth B_e is

$$M_e = 2\lfloor\beta\rfloor+3 = \begin{cases} 2\lfloor k_p a\rfloor+3, & \text{PM} \\ 2\lfloor\frac{k_f a}{f_m}\rfloor+3, & \text{FM} \end{cases}.$$

These expressions show that increasing the amplitude of the signal, a, has practically the same effect on both modulations, it increases the bandwidth. On the other hand, increasing the frequency, f_m, also increases the bandwidth, but the effect is much greater in PM than in FM modulation. In PM the increase is multiplicative while it is additive in FM.

An increase in amplitude increases the number of harmonics in the signal bandwidth. However, increasing the frequency, for PM, does not change the number of harmonics, while for FM it

decreases it. This explains the relative insensitivity of the FM bandwidth with respect to the frequency of the signal. On the one hand, increasing f_m increases the space between the relevant harmonics, but on the other hand their number decreases.

An example of the shape of the spectrum will be seen below. In this case, it will be done for a modulation with modulation index $\beta = 5$. In this case, the modulated signal is a sum of sinusoids of amplitude $A_c\, J_n(\beta)$ and pulsations $\omega_c + n\, \omega_m$. The values of the Bessel functions for the value of β are

$$J_0(5) = -0.18,\ J_1(5) = -0.32,\ J_2(5) = 0.05,\ J_3(5) = 0.37,\ J_4(5) = 0.39,\ J_5(5) = 0.26, \cdots$$

Figure 2.41 shows the shape of the frequency response for positive frequencies. Note that the frequency response of a cosine is a delta with amplitude π times the amplitude of the cosine. The property of Bessel functions that $J_{-n}(\beta)$ is equal to $J_{+n}(\beta)$ or to that value changed sign, depending on whether n is even or odd, provides that peculiar symmetry of the frequency response with respect to the carrier frequency ω_c.

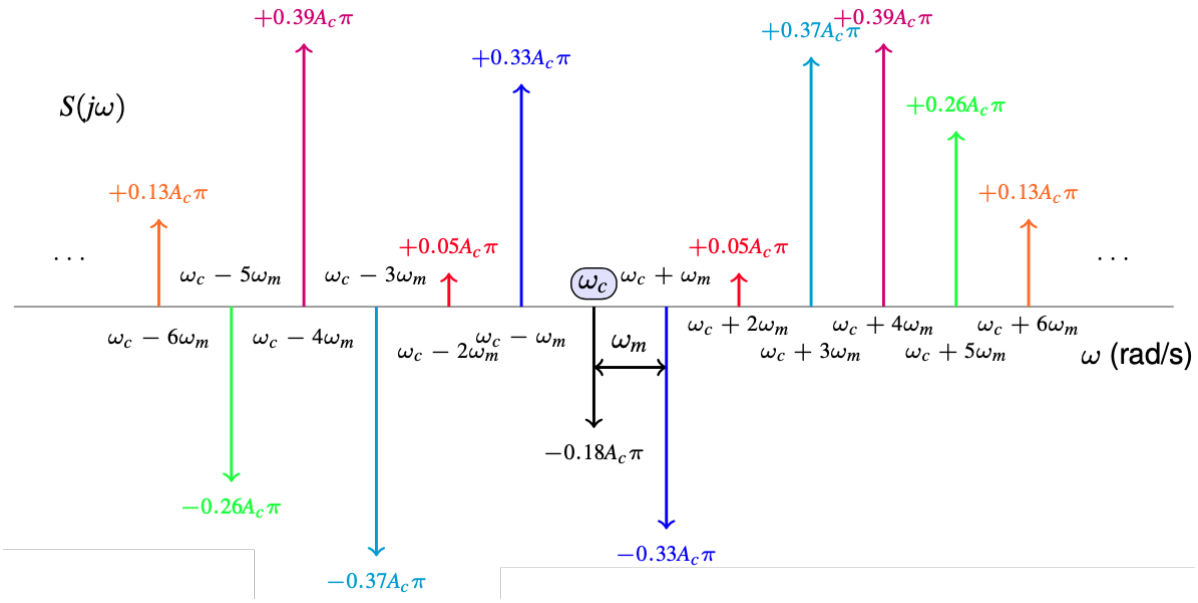


Figure 2.41: Frequency response of an angle modulation for a modulation index $\beta = 5$ with a sinusoidal modulator of frequency ω_m rad/s.

Modulation by a periodic signal

Similarly as in the case of a sinusoidal modulating signal, is the modulating signal is periodic the spectrum only has frequencies of the form $f_c + nf_m$. A periodic signal admits a Fourier series expansion, in such a way that it can be expressed as a sum of sinusoids with frequencies that are multiples of the one that defines the period. Therefore, the frequencies in the spectrum of the signal are

$$f_c \pm n\ f_m \ \text{ or } (\omega_c \pm n\ \omega_m)$$

The amplitudes of each frequency will be given by the sum of the contributions of all the harmonics.

Modulation by a non-periodic deterministic signal

The analysis of the spectum for a general modulating signal is very complex due to the non-linearity of angle modulations. In this case, there is only a heuristic rule that provides the approximate value of the bandwidth of the modulated signal. It is called the *Carson's rule*, which says that the bandwidth of the modulated signal when the modulating signal has a bandwidth of B Hz is, approximately

$$BW_{Carson} \approx 2\,(\beta + 1)\,B \text{ Hz}.$$

Taking into account that in wideband FM modulations the value of β is normally around 5 or even higher, this bandwidth is much greater than the bandwidth of amplitude modulations, which was B (SSB), $B + \Delta_B$ (VSB) or at most $2B$ Hz (for DSB and conventional AM).

2.3.4 Modulation of FM and PM signals

A discussion about the modulation and demodulation of angle modulations can be found in [Proakis and Salehi, 2002]. Any modulation or demodulation process, both for amplitude modulation and for angle modulations, involves the generation of frequencies that are not found in the original message signal. This means that both a modulator and a demodulator cannot be modeled by a linear and invariant system, since a linear invariant system does not produce new frequencies, frequencies not present in the input signal of the system, at the output of the system.

Angle modulators are, in general, non-linear and time-varying systems. One method of generating an FM signal directly is to design an oscillator whose voltage varies with an input voltage. These types of oscillators are called *voltage controlled oscillators* and are often denoted by the abbreviation VCO. A VCO can be implemented in several ways:

1. Using a *varactor diode*: A varactor is an element whose capacitance varies with the voltage applied to it. Therefore, if such a device is used in the design of an oscillator, its frequency varies as a function of the capacity and therefore of the applied voltage.

   $$C(t) = C_0 + k_0 m(t).$$

   When $m(t) = 0$, the frequency of the tuned circuit is

   $$f_c = \frac{1}{2\pi\sqrt{L_0 C_0}}.$$

   In general, for $m(t)$ not null we have

   $$\begin{aligned} f_i(t) =& \frac{1}{2\pi\sqrt{L_0(C_0 + k_0 m(t))}} \\ =& \frac{1}{2\pi\sqrt{L_0 C_0}} \frac{1}{\sqrt{1 + \frac{k_0}{C_0} m(t)}} \\ =& f_c \frac{1}{\sqrt{1 + \frac{k_0}{C_0} m(t)}}. \end{aligned}$$

   Assuming that

   $$\varepsilon = \frac{k_0}{C_0} m(t) << 1,$$

and using the approximations

$$\sqrt{1+\varepsilon} \approx 1 + \frac{\varepsilon}{2}, \quad \frac{1}{1+\varepsilon} \approx 1 - \varepsilon,$$

the following expression is obtained

$$f_i(t) \approx f_c \left(1 - \frac{k_0}{2C_0} m(t)\right).$$

2. Using *reactance tube*: A ballast tube is a device whose inductance varies with applied voltage. A similar analysis can be done for the varactor diode.

Another possibility for the generation of signals from an angle modulation is the so-called *indirect method*. In this case, the process is divided into two parts:

1. A narrow band angle modulation is generated. Due to the relationship with AM modulation this is easy to do.
2. The second step is to generate the broadband signal from the narrowband signal.

   If the narrowband signal is

$$s_{be}(t) = A_c \cos(2\pi f_c t + \phi(t)),$$

   the output of the frequency multiplier is

$$y(t) = A_c \cos(2\pi n f_c t + n\phi(t)).$$

   Finally, to set the desired carrier frequency, it is multiplied by a local oscillator

$$s(t) = A_c \cos(2\pi(n f_c - f_{OL})t + n\phi(t)).$$

2.3.5 Demodulation of FM and PM signals

Demodulation of an FM signal consists of finding the instantaneous frequency of the modulated signal, $s(t)$, and then subtracting the frequency of the carrier, since

$$m(t) = \frac{f_i(t) - f_c}{k_f}.$$

As for a PM signal, demodulation consists of finding the phase of the signal, since

$$m(t) = \frac{\phi(t)}{k_p}.$$

In general, an FM demodulator can be implemented using an FM to AM converter and then use an AM demodulator, as shown in Figure 2.42.

The FM-AM conversion can be done in many ways:

1. By derivation

$$|H(j\omega)| = \omega.$$

Figure 2.42: General scheme of an FM demodulator.

2. Frequency characteristic of a tuned circuit. The linear part of this response is used. This system is easy to implement, although the linear area may be small. When the linear area is too small, two circuits tuned to two frequencies can be used and combined with a so-called *balanced discriminator*.

These methods have the disadvantage that the bandwidth of the AM signal generated in the intermediate step is equal to the equivalent bandwidth, B_e, of the FM modulation, so the corresponding noise is the noise contained in that band that is generally greater than B.

2.4 Noise in analog communication systems

Previously, the power and bandwidth properties of the different analog modulations, amplitude and angle modulations, have been studied. In this section we analyze the effect of noise on these modulations. In all cases the following premises will be considered:

- The modulating signal $m(t)$ is bandlimited, with a bandwidth B Hz.
- An ideal transmission is assumed, a transmission over a Gaussian channel, where the transmitted signal $s(t)$ does not suffer any linear distortion and the only effect produced during transmission is the addition of thermal noise

  $$r(t) = s(t) + n(t).$$

  The power of the signal component that is received at the input of the receiver is therefore P_S, the power of the transmitted modulated signal. Regarding the thermal noise, the usual statistical model will be used.

  - Random process $n(t)$: stationary, ergodic, white, Gaussian, with power spectral density $S_n(j\omega) = \frac{N_0}{2}$.
- The receiver used for amplitude modulations will be a coherent receiver:
  - Filters will be introduced to limit the effect of noise before proceeding with demodulation. The filters will be fitted to the bandwidth of the transmitted signal, in such a way that the effect of noise is minimized without producing any distortion in the information signal $s(t)$.
  - The filters will be considered ideal, so as to obtain the maximum achievable performance.

The objective is to calculate the signal-to-noise ratio (S/N or SNR) of the demodulated signal for the different types of modulation. This figure of merit will be compared with the signal-to-noise ratio of a baseband transmission, when the signal is transmitted unmodulated. This reference ratio will be denoted as $\left(\frac{S}{N}\right)_b$.

2.4.1 Signal-to-noise ratio in a baseband transmission

In this section, the reference signal-to-noise ratio is obtained, the one obtained when the signal is transmitted without modulation, so that

$$s(t) = m(t) \rightarrow P_S = P_M.$$

The signal at the input of the receiver will therefore be

$$r(t) = s(t) + n(t) = m(t) + n(t)$$

In the receiver, the only processing that will be carried out will be filtering to minimize the effect of noise without distorting the information signal. To do this, an ideal low-pass filter of bandwidth B Hz will be used. Figure 2.43 shows the scheme of the receiver.

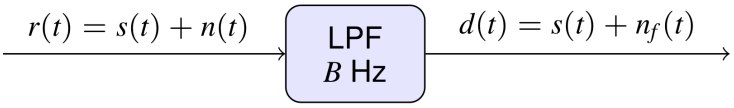


Figure 2.43: Diagram of a base band receiver for the reception of a signal with bandwidth B Hz.

The noise power at the output of the filter is obtained by integrating its power spectral density. Bearing in mind that at the output of a linear and invariant system the power spectral density is that of the input multiplied by the module squared of the frequency response of the filter, the power spectral density of the filtered noise $n_f(t)$ is

$$S_{n_f}(j\omega) = \begin{cases} \frac{N_0}{2} & \text{if } |\omega| \leq W = 2\pi B \text{ rad/s} \\ 0 & \text{if } |\omega| > W = 2\pi B \text{ rad/s} \end{cases}$$

Therefore, the power of the filtered noise is

$$P_{n_f} = \frac{1}{2\pi} \int_{-\infty}^{+\infty} S_{n_f}(j\omega)\, d\omega = \frac{1}{2\pi} \int_{-2\pi B}^{+2\pi B} \frac{N_0}{2}\, d\omega = N_o\, B \text{ Watt.}$$

This same result could be obtained considering that the noise power that passes through an ideal filter with bandwidth B Hz is $N_0 \times B$ Watt. In either case, the baseband signal-to-noise ratio is

$$\left(\frac{S}{N}\right)_b = \frac{P_S}{N_0\, B}.$$

This will be the reference value with which the signal-to-noise ratio obtained with the different modulation variants will be compared.

2.4.2 Effect of noise on amplitude modulations

In this section, the signal-to-noise ratio is determined at the output of the receiver that demodulates the amplitude modulated signals. The results obtained are compared with the result of the noise effect in a baseband transmission.

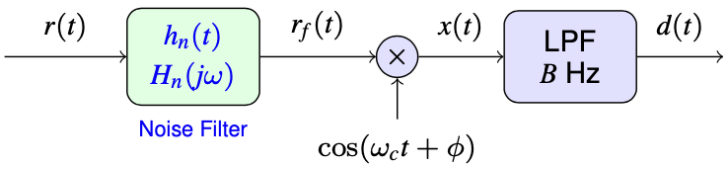


Figure 2.44: Block diagram of a coherent receiver used for amplitude modulations, including the noise filter.

Coherent and filtered receiver to minimize the effect of noise

For all amplitude modulations, the receiver shown in Figure 2.44 is going to be used, where $\phi = \phi_c$ as it is a coherent receiver. When necessary, for simplicity, we will consider $\phi_c = 0$.

The noise filter, with impulse reponse $h_n(t)$ and frequency response $H_n(j\omega)$, is placed before the synchronous demodulator to minimize the effect of noise. To do this, it is an ideal bandpass filter whose passband, and therefore the bandwidth, is the same as that of the modulated signal $s(t)$.

The received signal is modeled with the thermal additive noise model

$$r(t) = s(n) + n(t).$$

The signal component at the outpuf of the noise filter, taking into account that the filter is fitted to the bandwidth of the signal, is

$$r_f(t) = s(n) + n_f(t), \text{ con } n_f(t) = n(t) * h_n(t).$$

The demodulated signal is obtained as

$$x(t) = r_f(t) \times \cos(\omega_c t) = s(t)\ \cos(\omega_c t) + n_f(t)\ \cos(\omega_c t) = x_S(t) + x_n(t)$$

and the filtered demodulated signal is

$$d(t) = x(t) * h_{LPF-B}(t) = x_S(t) * h_{LPF-B}(t) + x_n(t) * h_{LPF-B}(t) = d_S(t) + d_n(t).$$

As you can see, at the output of the receiver there are two terms:

- A term due to the modulated signal $s(t)$, which is $d_S(t)$.
- A term due to thermal noise $n(t)$, which is $d_n(t)$.

The signal term is not affected by the noise filter. For amplitude modulations, it was calculated before, and the obtained result, along with the power of the demodulated signal that is related with the information signal $m(t)$, P_{d_S}, is shown in Table 2.3.

Regarding the noise term, its power depends on the noise filter that is used, which in turn depends on the modulation variant, and in particular on its bandwidth. We will denote the noise power as P_{d_n}, and it will be calculated later for each type of modulation.

Once the noise power at the demodulator output has been obtained, the signal-to-noise ratio after demodulation will be obtained as

$$\left(\frac{S}{N}\right)_d = \frac{P_{d_S}}{P_{d_n}},$$

and it will be compared with the baseband signal-to-noise ratio

$$\left(\frac{S}{N}\right)_b = \frac{P_S}{N_o\, B}.$$

Modulation	P_S	$d_S(t)$	P_{d_S}
Conventional AM	$\frac{A_c^2}{2}[1+P_{M_a}]$	$\frac{A_c}{2}[1+m_a(t)]$	$\frac{A_c^2}{4}P_{M_a}$
DSB	$\frac{A_c^2}{2}P_M$	$\frac{A_c}{2}m(t)$	$\frac{A_c^2}{4}P_M$
SSB	$A_c^2 P_M$	$\frac{A_c}{2}m(t)$	$\frac{A_c^2}{4}P_M$
VSB	$A_c^2 P_M$	$\frac{A_c}{2}m(t)$	$\frac{A_c^2}{4}P_M$

P_{d_S}: power in $d_S(t)$ related to $m(t)$
P_{M_a}: power of $m_a(t)$, $P_{M_a} = \frac{a^2}{C_M^2}P_M$

Table 2.3: Power of the modulated signal, demodulator output signal, and signal power for each type of amplitude modulation when using a synchronous demodulator.

Noise power at the demodulator output - General analysis

The power spectral density of the filtered noise $n_f(t)$ is

$$S_{n_f}(j\omega) = S_n(j\omega)\ |H_n(j\omega)|^2 = \frac{N_0}{2}|H_n(j\omega)|^2.$$

The power spectral density of the demodulated noise $x_n(t)$ is then

$$S_{x_n}(j\omega) = \frac{1}{4}S_{n_f}(j\omega - j\omega_c) + \frac{1}{4}S_{n_f}(j\omega + j\omega_c) = \frac{N_0}{8}\left[|H_n(j\omega - j\omega_c)|^2 + |H_n(j\omega + j\omega_c)|^2\right].$$

Finally, the power spectral density after low-pass filtering, for signal $d_n(t)$, is

$$S_{d_n}(j\omega) = S_{x_n}(j\omega)\ |H_{LPF-B}(j\omega)|^2 = \begin{cases} S_{x_n}(j\omega), & \text{si } |\omega| \le W = 2\pi B \\ 0, & \text{si } |\omega| > W = 2\pi B \end{cases}$$

The noise power at the ouput of the coherent receiver is now calculated by integrating this PSD

$$\begin{aligned} P_{d_n} &= \frac{1}{2\pi}\int_{-\infty}^{\infty} S_{d_n}(j\omega)\ d\omega = \frac{1}{2\pi}\int_{-2\pi B}^{2\pi B} S_{x_n}(j\omega)\ d\omega \\ &= \frac{N_0}{8}\left[\frac{1}{2\pi}\int_{-2\pi B}^{2\pi B}|H_n(j\omega - j\omega_c)|^2\ d\omega + \frac{1}{2\pi}\int_{-2\pi B}^{2\pi B}|H_n(j\omega + j\omega_c)|^2\ d\omega\right]. \end{aligned}$$

Noise power calculation - Conventional AM and DSB

Since the bandwidth of the conventional and double-sideband AM modulations is identical, the noise filter and therefore the noise power is also the same. For both modulations the noise filter is identical

$$H_n(j\omega) = \begin{cases} 1, & \text{if } \omega_c - W \le |\omega| \le \omega_c + W \\ 0, & \text{in other case} \end{cases},$$

where W is the bandwidth in rad/s ($W = 2\pi B$). Figure 2.45 represents the frequency response of this noise filter.

For this noise filter

$$\frac{1}{2\pi}\int_{-2\pi B}^{2\pi B}|H_n(j\omega - j\omega_c)|^2\ d\omega + \frac{1}{2\pi}\int_{-2\pi B}^{2\pi B}|H_n(j\omega + j\omega_c)|^2\ d\omega = \frac{1}{2\pi}\ 4W = 4B,$$

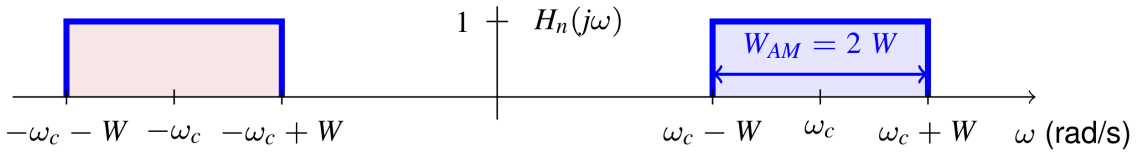


Figure 2.45: Frequency response of the noise filter used in conventional AM modulation and double sideband modulation.

and the noise power is therefore

$$P_{d_n} = \frac{1}{2} N_0 B.$$

Figure 2.46 shows the frequency interpretation of the process that noise undergoes in the receiver.

Noise Power Calculation - Single Sideband (SSB)

The case of an upper sideband SSB modulation will be considered. For lower sideband the result is the same. For this modulation the noise filter has the following frequency response

$$H_n(j\omega) = \begin{cases} 1, & \text{if } \omega_c \leq |\omega| \leq \omega_c + W \\ 0, & \text{in other case} \end{cases},$$

where W is the bandwidth in rad/s ($W = 2\pi B$). Figure 2.47 represents the module of the frequency response of the noise filter.

For this noise filter

$$\frac{1}{2\pi} \int_{-2\pi B}^{2\pi B} |H_n(j\omega - j\omega_c)|^2 \, d\omega + \frac{1}{2\pi} \int_{-2\pi B}^{2\pi B} |H_n(j\omega + j\omega_c)|^2 \, d\omega = \frac{1}{2\pi} \, 2W = 2B$$

and the noise power is therefore

$$P_{d_n} = \frac{1}{4} N_0 B.$$

Figure 2.48 shows the frequency interpretation of the process that noise undergoes in the receiver.

Noise Power Calculation - Vestigial Sideband (VSB)

The case of an upper sideband VSB modulation will be considered. For lower sideband the result is the same. For this modulation the noise filter has the following frequency response

$$H_n(j\omega) = \begin{cases} 1, & \text{if } \omega_c - \Delta_W \leq |\omega| \leq \omega_c + W \\ 0, & \text{in other case} \end{cases},$$

where W is the bandwidth in rad/s ($W = 2\pi B$), and Δ_W is the excess vestigial bandwidth in rad/s ($\Delta_W = 2\pi \ Delta_B$). Figure 2.47 represents the module of the frequency response of the noise filter used.

For this noise filter

$$\frac{1}{2\pi} \int_{-2\pi B}^{2\pi B} |H_n(j\omega - j\omega_c)|^2 \, d\omega + \frac{1}{2\pi} \int_{-2\pi B}^{2\pi B} |H_n(j\omega + j\omega_c)|^2 \, d\omega = \frac{1}{2\pi} \, 2(W + \Delta_W) = 2(B + \Delta_B)$$

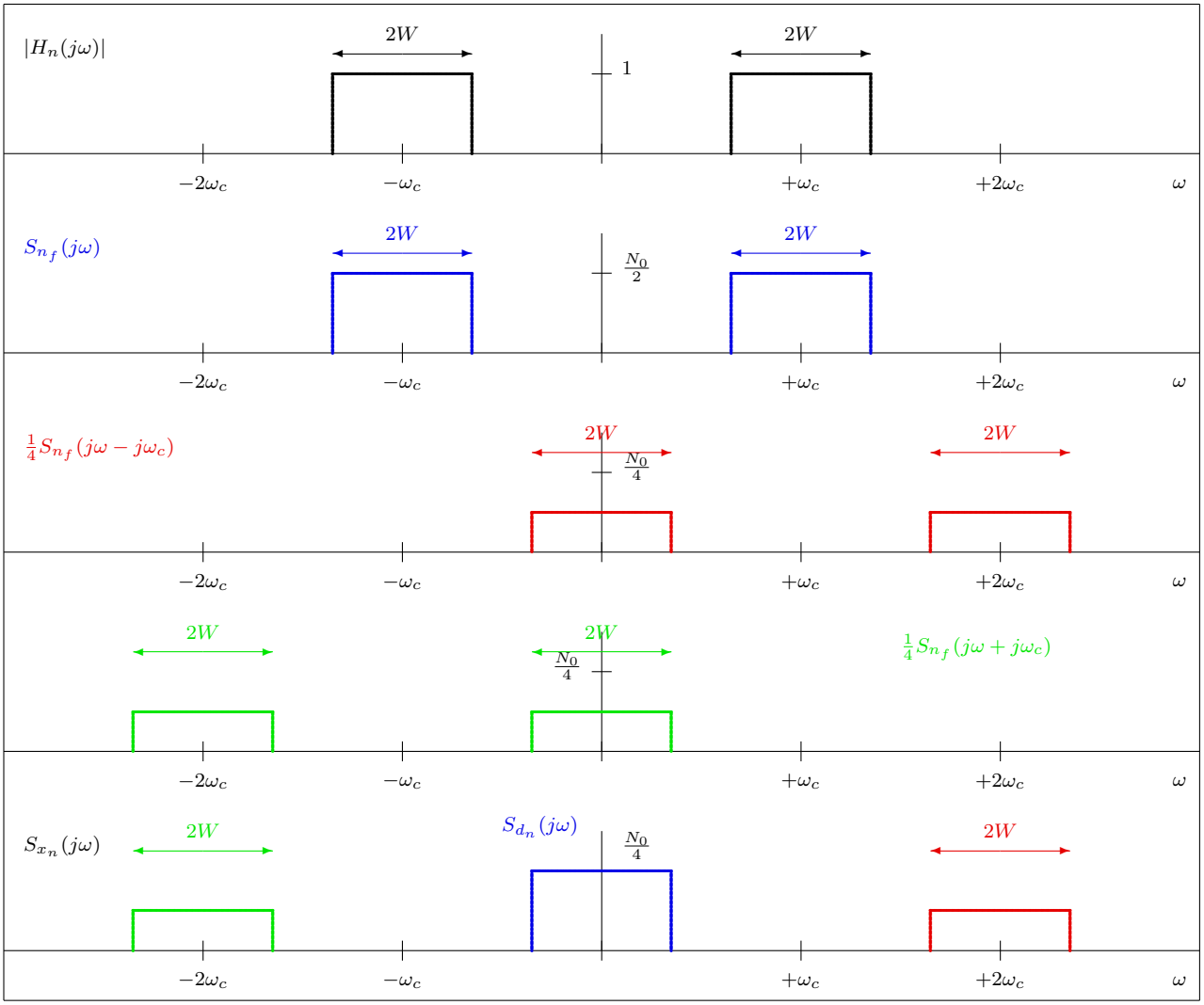


Figure 2.46: Frequency interpretation of the process that noise undergoes in the receiver for conventional AM and DSB modulations.

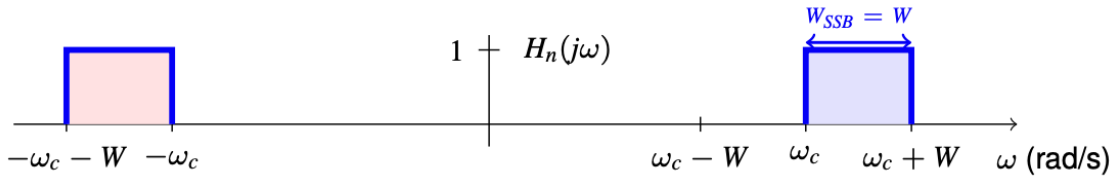


Figure 2.47: Frequency response of the noise filter used in single-sideband (upper-sideband) modulation.

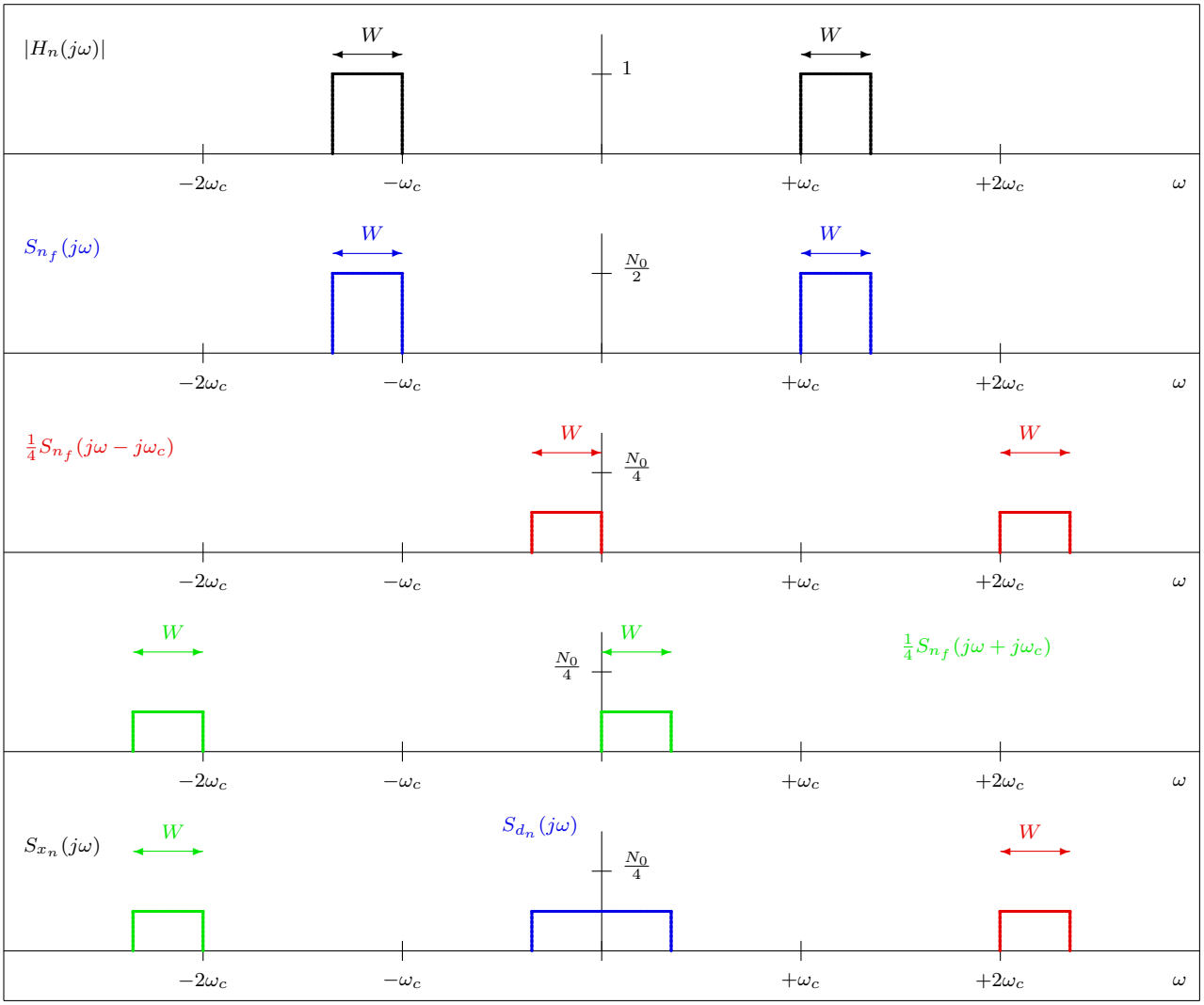


Figure 2.48: Frequency interpretation of the process that noise undergoes in the receiver for a single sideband modulation (upper sideband).

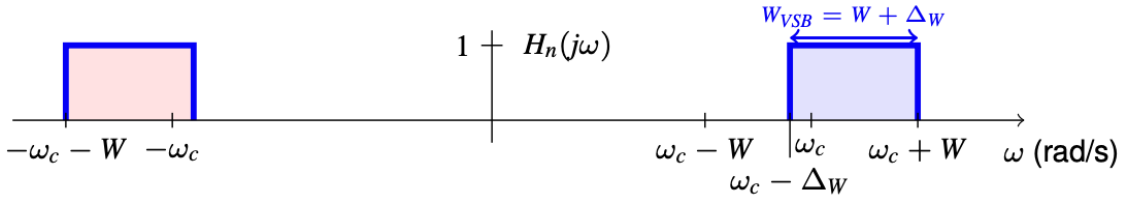


Figure 2.49: Frequency response of the noise filter used in a vestigial sideband (upper sideband) modulation.

and the noise power is therefore

$$P_{d_n} = \frac{1}{4} N_0 (B + \Delta_B).$$

Figure 2.50 shows the frequency interpretation of the process that noise undergoes in the receiver.

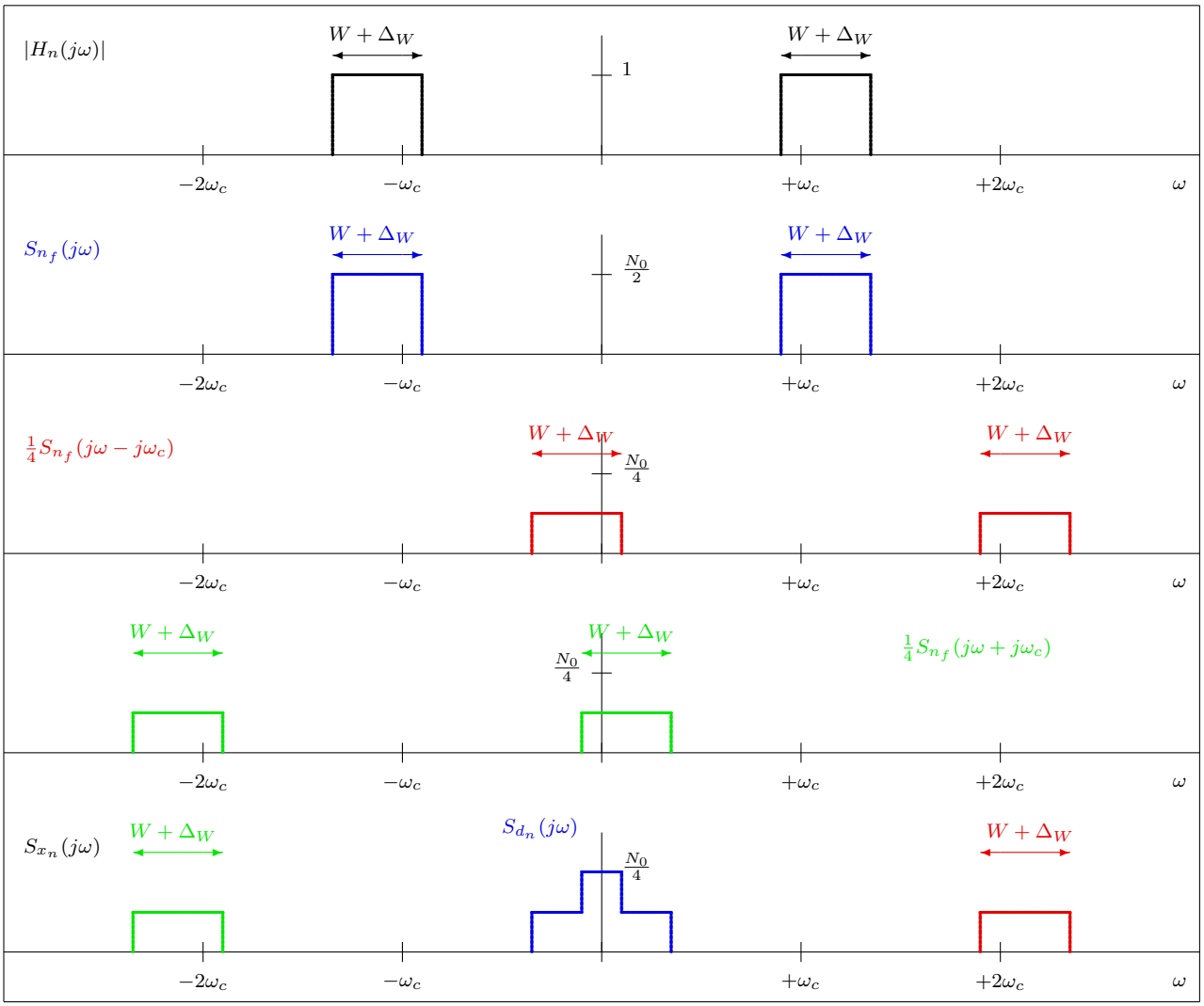


Figure 2.50: Frequency interpretation of the process that noise undergoes in the receiver for a vestigial sideband modulation (upper sideband).

Calculation of signal to noise ratios

Once the power of the signal and noise components at the demodulator output are known (summarized in Table 2.4), all that remains is to calculate the signal-to-noise ratio and to compare it with that obtained in a baseband transmission.

Double Sideband (DSB) modulation The signal to noise ratio for this type of modulation is

$$\left(\frac{S}{N}\right)_{DSB} = \frac{P_{d_S}}{P_{d_n}} = \frac{\frac{A_c^2}{4} P_M}{\frac{1}{2} N_0 B} = \frac{\frac{A_c^2}{2} P_M}{N_0 B} = \frac{P_S}{N_0 B} = \left(\frac{S}{N}\right)_b.$$

Modulation	P_S	$d_S(t)$	P_{d_S}	P_{d_n}
Conventional AM	$\frac{A_c^2}{2}[1+P_{M_a}]$	$\frac{A_c}{2}[1+m_a(t)]$	$\frac{A_c^2}{4}P_{M_a}$	$\frac{1}{2}N_0B$
DSB	$\frac{A_c^2}{2}P_M$	$\frac{A_c}{2}m(t)$	$\frac{A_c^2}{4}P_M$	$\frac{1}{2}N_0B$
SSB	$A_c^2P_M$	$\frac{A_c}{2}m(t)$	$\frac{A_c^2}{4}P_M$	$\frac{1}{4}N_0B$
VSB	$A_c^2P_M$	$\frac{A_c}{2}m(t)$	$\frac{A_c^2}{4}P_M$	$\frac{1}{4}N_0(B+\Delta_B)$

Table 2.4: Signal and noise powers at the demodulator output for amplitude modulations.

It can be seen that the DSB modulation haves exactly the same signal-to-noise ratio than a baseband transmission.

Single Sideband (SSB) modulation The signal to noise ratio for a SSB modulation is

$$\left(\frac{S}{N}\right)_{SSB}=\frac{P_{d_S}}{P_{d_n}}=\frac{\frac{A_c^2}{4}P_M}{\frac{1}{4}N_0B}=\frac{A_c^2P_M}{N_0B}=\frac{P_S}{N_0B}=\left(\frac{S}{N}\right)_b.$$

Again, the same signal-to-noise ratio as for a baseband transmission.

Conventional AM modulation The signal to noise ratio for a conventional AM modulation is

$$\begin{aligned}\left(\frac{S}{N}\right)_{AM}&=\frac{P_{d_S}}{P_{d_n}}=\frac{\frac{A_c^2}{4}P_{M_a}}{\frac{1}{2}N_0B}=\frac{\frac{A_c^2}{2}P_{M_a}}{N_0B}=\frac{P_{M_a}}{1+P_{M_a}}\frac{\frac{A_c^2}{2}[1+P_{M_a}]}{N_0B}\\&=\underbrace{\frac{P_{M_a}}{1+P_{M_a}}}_{\eta_{AM}}\frac{P_S}{N_0B}=\eta_{AM}\left(\frac{S}{N}\right)_b\end{aligned}$$

In this case, the signal-to-noise ratio is worse than in a baseband transmission. This is due to the transmission of the carrier, which does not contain information, and which makes this type of modulation inefficient in power. The efficiency factor $\eta_{AM}<1$ is in this case

$$\eta_{AM}=\frac{P_{M_a}}{1+P_{M_a}}=\frac{\frac{a^2}{C_M^2}P_M}{1+\frac{a^2}{C_M^2}P_M}=\frac{P_M}{\frac{C_M^2}{a^2}+P_M}.$$

The efficiency depends on the modulation index a, with lower efficiency for lower values of a.

Vestigial Sideband Modulation (VSB) The signal to noise ratio for a VSB modulation is

$$\begin{aligned}\left(\frac{S}{N}\right)_{VSB}&=\frac{P_{d_S}}{P_{d_n}}=\frac{\frac{A_c^2}{4}P_M}{\frac{1}{4}N_0(B+\Delta_B)}=\frac{A_c^2P_M}{N_0(B+\Delta_B)}=\frac{B}{B+\Delta_B}\frac{A_c^2P_M}{N_0B}\\&=\underbrace{\frac{B}{B+\Delta_B}}_{\eta_{BLV}}\frac{P_S}{N_0B}=\eta_{BLV}\left(\frac{S}{N}\right)_b.\end{aligned}$$

In this case the ratio is also worse than the signal-to-noise ratio transmitting in baseband. The efficiency factor $\eta_{BLV}<1$ is now

$$\eta_{BLV}=\frac{B}{B+\Delta_B}.$$

The efficiency therefore depends on the excess of bandwidth or vestige, Δ_B. If the vestige is small with respect to the bandwidth, $\Delta_B << B$, in that case $\eta_{BLV} \approx 1$, i.e., the signal-to-noise ratio is similar to the one in a baseband transmission.

2.4.3 Effect of noise on angle modulations

The analysis of the noise effect for angle modulations is relatively complicated due to the non-linear dependence of the modulated signal with respect to the modulating signal, since it is inside of the argument of the sinusoidal carrier.

In general, to summarize the main characteristics without going into a rigorous analytical development, the demodulated signal can be written as

$$d(t) = \begin{cases} k_p m(t) + Y_n(t), & \text{PM} \\ k_f m(t) + \frac{1}{2\pi}\frac{d}{dt}Y_n(t), & \text{FM} \end{cases}.$$

The noise term, $Y_n(t)$, leads to the following expressions for the power spectral density of the noise at the demodulator output

$$S_{n_d}(j\omega) = \begin{cases} \frac{N_0}{A_c^2}, & \text{PM} \\ \frac{N_0}{A_c^2}\omega^2, & \text{FM} \end{cases}.$$

This implies that the noise power at the output of the demodulator is

$$\begin{aligned} P_{n_d} =& \frac{1}{2\pi}\int_{-\infty}^{\infty} S_{n_d}(j\omega)\, d\omega \\ =& \begin{cases} \frac{1}{2\pi}\int_{-\infty}^{\infty} \frac{N_0}{A_c^2}\, d\omega, & \text{PM} \\ \frac{1}{2\pi}\int_{-\infty}^{\infty} \frac{N_0}{A_c^2}\omega^2\, d\omega, & \text{FM} \end{cases} \\ =& \begin{cases} \frac{2N_0 B}{A_c^2}, & \text{PM} \\ \frac{2N_0 B^3}{3A_c^2}\omega^2, & \text{FM} \end{cases}. \end{aligned}$$

And the signal to noise ratio

$$\left(\frac{S}{N}\right)_d = \frac{P_S}{P_{n_d}},$$

is

$$\left(\frac{S}{N}\right)_d = \begin{cases} \frac{k_p^2 A_c^2}{2}\frac{P_M}{N_0 B}, & \text{PM} \\ \frac{3k_f^2 A_c^2}{2B^2}\frac{P_M}{N_0 B}, & \text{FM} \end{cases}.$$

If the received signal power is denoted as $P_S = \frac{A_c^2}{2}$ and the modulation indices are considered

$$\left(\frac{S}{N}\right)_d = \begin{cases} P_R\left(\frac{\beta_p}{\max|m(t)|}\right)^2 \frac{P_M}{N_o B}, & \text{PM} \\ 3P_R\left(\frac{\beta_f}{\max|m(t)|}\right)^2 \frac{P_M}{N_o B}, & \text{FM} \end{cases}.$$

Taking into account that

$$\left(\frac{S}{N}\right)_b = \frac{P_S}{N_o B},$$

the signal-to-noise ratio can be written in terms of the signal-to-noise ratio transmitting at baseband as follows

$$\left(\frac{S}{N}\right)_d = \begin{cases} P_M \left(\dfrac{\beta_p}{\max|m(t)|}\right)^2 \left(\dfrac{S}{N}\right)_b, & \text{PM} \\ 3P_M \left(\dfrac{\beta_f}{\max|m(t)|}\right)^2 \left(\dfrac{S}{N}\right)_b, & \text{FM} \end{cases}.$$

It can be observed that in the angle modulations there is a fanance in relation to signal to noise proportional to the modulation index squared.

In summary, the expression for the signal-to-noise ratio for angle modulations could be written in a general way as

$$\left(\frac{S}{N}\right)_d = \alpha \left(\frac{P_M}{C_M^2}\right) \times \beta^2 \times \left(\frac{S}{N}\right)_b$$

In this general expression

- The factor α depends on the modulation: $\alpha_{PM} = 1$, $\alpha_{FM} = 3$
- The term $\left(\frac{P_M}{C_M^2}\right)$ is usually constant (its value depends on the type of modulating signals)

Threshold effect in angle modulations

This gain effect only occurs in practice if the baseband signal-to-noise ratio is greater than a threshold given by

$$\left(\frac{S}{N}\right)_{threshold} = 20\ (\beta + 1).$$

In practice, this implies that there is a threshold level for the received signal power, from which this gain is obtained in relation to signal-to-noise

$$P_{S_{threshold}} = (N_0 B) \times \left(\frac{S}{N}\right)_{threshold} \rightarrow A_{c,threshold} = \sqrt{2P_{S_{threshold}}}.$$

Chapter 3

Modulation and detection in Gaussian channels

The basic function of a digital communications system consists of sending and reliably retrieving a digital sequence of information sent over an analog channel. This is the problem addressed in this chapter.

The chapter begins by introducing the convenience of a geometrical (vector) representation of signals for the design and analysis of a digital communication system, and the basic characteristics of such a representation are presented. Next, the problem of transmission in white, Gaussian, additive noise channels, commonly referred to as a *Gaussian channel*, is studied. In this model it is assumed that the only distortion suffered by the communications signal during transmission is the sum of thermal noise, which is modeled by a stationary, ergodic, white, Gaussian, random process with power spectral density $N_0/2$ W/Hz. The functional elements of the transmitter and receiver are analyzed, addressing the optimal design of each element, and the methodology to evaluate the performance of a given system is presented.

3.1 Introduction

This chapter focuses on the analysis of digital communication systems. Therefore, it is convenient to recall the characteristics and the communication model of these systems.

3.1.1 Advantages of digital communication systems

Digital communications systems clearly prevaul over analog systems because they have multiple advantages over them. The most important are the following:

- The "*regeneration*" capability. This is undoubtedly the main advantage. Under certain circumstances it is possible to transmit without errors, or in general, with an arbitrarily low probability of error. This is possible because at a certain instant of time, the transmitted symbol is one among a finite set. This means that the signal can have a shape among a finite set, and if the distortion is not very severe, the most similar to the received shape is the transmitted one, and the signal can be "*regenerated*", as illustrated in Figure 3.1. As

it can be seen in the figure, in each interval one of two possible waveforms is transmitted, high level and low level, so if the distortion is moderate in the receiver, the signal can be recovered in each interval and recover exact transmitted signal.

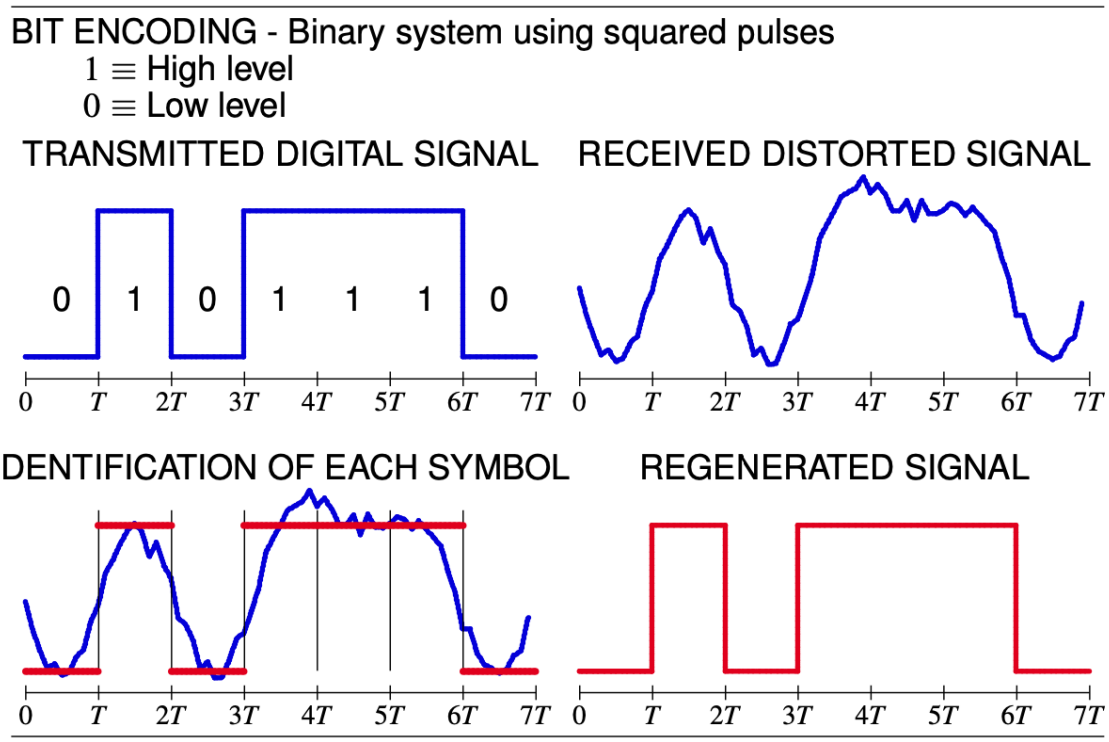


Figure 3.1: Illustration of the regeneration capacity in a digital communications system.

- Error detection and correction techniques exist. In the example of Figure 3.1 the distortion suffered by the transmitted signal is moderate, so the transmitted signal (binary sequence) is recovered without errors. In real cases errors occur, but there are techniques that allow to detect and correct most of these errors.
- Information can be encrypted (protected) relatively easily. It is enough to apply arithmetic operations on the binary information with a binary key.
- The distortion that is introduced by the channel can be compensated (channel equalization) in a much easier way than in analog systems. Knowing what possible values the signal can take in a certain interval makes it easier to estimate the effect of the channel on the transmitted signal in order to invert it.
- The format of the transmitted information is independent of the type of information (voice, data, TV, etc.). In any case, the information is stored in a sequence of binary symbols (bits). To design a digital system, the objective is to transmit bits efficiently, regardless of the type of information that they contain.
- New multiplexing or media access mechanisms, such as TDM/TDMA and CDM/CDMA (in addition to FDM/FDMA), can be used to simultaneously transmit several digital signals in a single medium.
- The circuits implementing a digital system are, in general
  - More reliable

- Of lower cost
- More flexible (programmable)

Obviously not all are advantages in the comparison between analog and digital systems. Although the advantages clearly predominate over the disadvantages, it is convenient to know these as well. The most important are:

- The need for synchronism at the receiver. In the example of Figure 3.1, the instants where the transmission of each bit begins and ends are identified, which allows to compare in each interval the received signal with a high or low level to perform the regeneration of the signal (or what is the same, the identification of the transmitted binary sequence). In a real system, the digital receiver must generate the synchronism signal that allows the identification of these instants. This is not necessary in analog systems.
- In general, an information signal for a source of the same type (e.g. a voice signal) requires a higher bandwidth for transmission in digital format than in analog format. This disadvantage can be compensated by using compression techniques for signals in digital format, which allows them to be transmitted with a lower bandwidth.
- Many information sources are analog in nature. In reality, this is a minor drawback, since as we have already seen in the Introduction chapter, the analog-to-digital conversion of information signals (aling with the corresponding digital-to-analog conversion at the receiver) allows the transmission of sources of analog nature through digital communications systems. Analog-to-digital (A/D) and digital-to-analog (D/A) conversion technology is now a mature and cost-effective technology.

3.1.2 Overview of a digital communications system

The main characteristic of a digital communications system is that the information to be transmitted is stored in a sequence of symbols belonging to a finite alphabet of symbols, as opposed to analog communications systems, where the information is contained in the shape of a continuous time waveform $m(t)$. This characteristic endows this type of systems with a series of advantages that have made them prevail over analog communications systems, as we have seen before.

Although in general the information can be encoded in different alphabets, this chapter will focus on the most frequent case, in which the information is contained in a binary sequence, so that the objective of the system will be the efficient transmission of bits between the source and the destination of the transmission.

Figures 3.2 and 3.3 show the block diagrams of a digital transmitter and receiver, respectively. These diagrams include the basic functional elements that appear in any system.

The first element of the transmitter, the source encoder, generates the binary information to be transmitted. Usually, it includes some compression to reduce the redundancy of the information source to low the necessary binary rate for the transmission.

The channel encoder is responsible for introducing redundancy in a controlled manner to allow the detection and correction of a certain number of bit errors. The simplest example of channel encoders are repetition codes. These codes repeat each bit to be transmitted several times. A

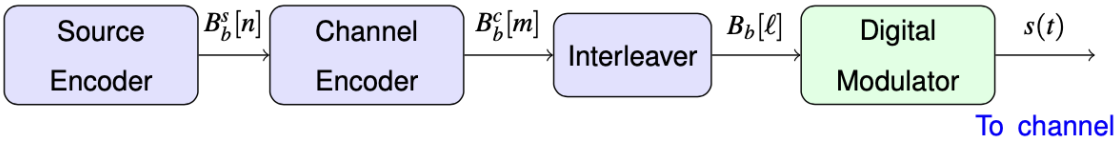


Figure 3.2: Functional block diagram of the transmitter of a digital communications system.

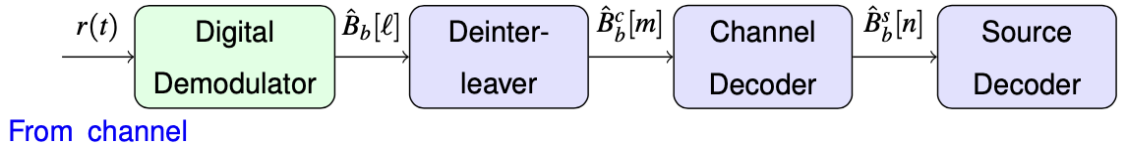


Figure 3.3: Functional block diagram of the receiver of a digital communications system.

repetition code of order 1 when it transmits a certain bit repeats it again, so that each bit is sent twice. A repetition code of order 2 repeats each bit twice, ultimately transmitting each bit three times, performing a coding process

$$0 \rightarrow 000$$

$$1 \rightarrow 111$$

In each block of three coded bits, the receiver can detect up to two errors. By adopting the majority decision strategy, the receiver will be able to correct an error on each block of three transmitted bits. If there are two or three errors, this majority decision scheme would result in the wrong decision on the block. Thus, each channel code has a certain detection and correction capability.

On many occasions, bit errors in the transmitted sequence in a system do not occur isolated, but rather in bursts of errors. The direct application of the channel coding techniques on bursts of errors is not useful, since they work well when there are a limited number of errors on each block of coded bits (1 error out of 3 encoded bits in the case of the previous example). The role of the interleaver is to convert bursts of bit errors into isolated errors. To do this, the bits are rearranged before transmission (interleaved) and returned to the original order at the receiver (deinterleaved), so that burst errors before deinterleaving become isolated errors at the deinterleaver output, as described in the example that is shown in Figure 3.4.

In the transmitter, from the source encoder, and after carrying out the channel coding and interleaving processes, a binary sequence $B_b[\ell]$ is generated that contains the information to be transmitted. In this chapter we will study the last element of the transmitter, the digital modulator. The main function of this digital modulator is the transmission of the bit sequence, $B_b[\ell]$, over an analog communications channel, for which it must convert the information sequence into an electromagnetic signal, $s(t)$, that is suitable to the analog medium used for transmission (a cable, optical fiber, the radio spectrum, etc.).

At the receiver, the digital demodulator will recover transmitted information from the received signal, $r(t)$. Ideally, the transmitted sequence should be recovered exactly, but the distortion suffered by the signal during its transmission will make this not possible in general. Therefore, the digital demodulator must provide an estimate of the transmitted sequence, $\hat{B}_b[\ell]$, trying to minimize the number of errors in the estimate.

In this chapter we will study the digital modulator of the transmitter and the digital demodulator

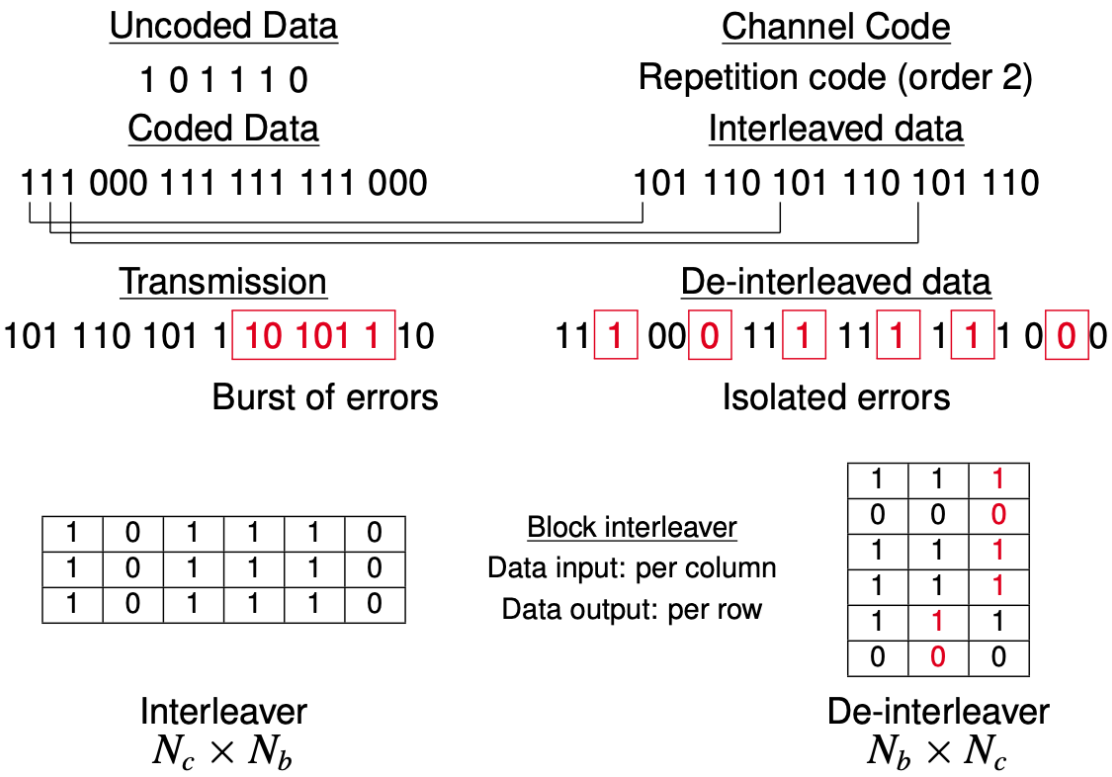


Figure 3.4: Example of a matrix interleaver, in which the bits are reordered by entering them in a matrix structure by columns and reading them out by rows.

of the receiver. In other words, the basic problems that are addressed are the transmission of a sequence of bits over an analog channel, and the recovery of that sequence from the signal that is received at the output of the channel, which suffers certain types of distortions during its transmission.

3.1.3 General design of a digital modulator and basic notation

As we have just seen, the function of the digital modulator is the transmission of bits, the binary sequence $B_b[\ell]$, at a bit rate $R_b = \frac{1}{T_b}$ bits/s over a transmission medium (channel) of an analog nature. To do this, it must convert the binary sequence into an electromagnetic signal $s(t)$.

In general, the transmission of the binary sequence is not carried out bit-by-bit, but will be carried out by blocks of m bits. The bit sequence $B_b[\ell]$ shall be splitted into blocks of m bits, and each of these blocks shall be sent. Each block of m bits is called a *symbol*, so the alphabet of symbols has $M = 2^m$ possible values: the M possible combinations of m bits. The elements of this alphabet are denoted as b_i, with $i \in \{0, 1, \cdots, M-1\}$, that is

$$B[n] \in \{b_0, b_1, \cdots, b_{M-1}\}.$$

Thus, the first step in the digital modulator is the bits-to-symbols conversion of the binary sequence $B_b[\ell]$ to a the symbol sequence $B[n]$. An important aspect in this conversion is the rate conversion. In a digital communication systems two transmission rates are identified: the binary rate, R_b bits/s, and the symbol rate, R_s symbols/s or bauds. The relationship between these two transmission rates is evident. Since each symbol contains m bits, the relation is

$$R_b = m \times R_s.$$

Associated with each rate is a time duration, the bit time duration, T_b and the symbol time duration, T, respectively. The relationship of these times with the transmission rates is also evident

$$T_b = \frac{1}{R_b},\ T = \frac{1}{R_s},\ T = m \times T_b.$$

This chapter will present the simplest form of information transmission, with the objective of introducing the basic modulation and demodulation techniques. The sequence of digital information has to be converted into an electrical signal. This conversion is piecewise. One symbol is transmitted every T seconds, so that a time interval is associated with the transmission of each symbol. The transmission of the first symbol of the sequence, $B[0]$, starts at $t = 0$, and lasts until $t = T$. Then the second symbol of the sequence, $B[1]$, will be transmitted between instants $t = T$ and $t = 2T$, and so on. In general, the symbol $B[n]$ will have associated with it the interval $nT \leq t < (n+1)T$. Thus, the shape of the signal in each symbol interval will be associated with the symbol transmitted in that interval.

As explained above, the symbol sequence can take one among M possible values for each discrete instant n.

$$B[n] \in \{b_0, b_1, \cdots, b_{M-1}\}.$$

The simplest way to perform the symbol-to-signal conversion $s(t)$ is to define a set of M signals of duration T seconds

$$\{s_0(t), s_1(t), \cdots, s_{M-1}(t)\},\ \text{with support in } 0 \leq t < T$$

and making a symbol-to-waveform association:

$$b_i \leftrightarrow s_i(t).$$

When at instant n the sequence of symbols takes on a certain value, for example $B[n] = b_j$, the shape of the signal $s(t)$ in the interval associated with this symbol, $nT \leq t < (n+1)T$, will be the waveform associated to b_j, which is $s_j(t)$. Obviously, the waveform $s_j(t)$ must be shifted to the corresponding symbol interval

$$s(t) = s_j(t - nT),\ \text{in } nT \leq t < (n+1)T \text{ if } B[n] = b_j.$$

This procedure is illustrated below with a simple example. In this case, the bits will be transmitted in blocks of size $m = 2$ bits, so the system has an alphabet of $M = 2^2 = 4$ symbols. For example, the following association can be made

$$b_0 \equiv 00,\ b_1 \equiv 01,\ b_2 \equiv 10,\ b_3 \equiv 11,$$

although a different one could have been made. A set of $M = 4$ signals must be chosen, and a signal must be associated to each symbol. In this example, the four signals of Figure 3.5 have been chosen. The signal-to-symbol association is implicit in the subindices ($s_i(t)$ is associated with b_i).

The initial part of the binary sequence is in this example

$$B_b[\ell] = 011110001101\cdots$$

This sequence is splitted into blocks of $m = 2$ bits

$$B_b[\ell] : 01 \mid 11 \mid 10 \mid 00 \mid 11 \mid 01 \mid \cdots,$$

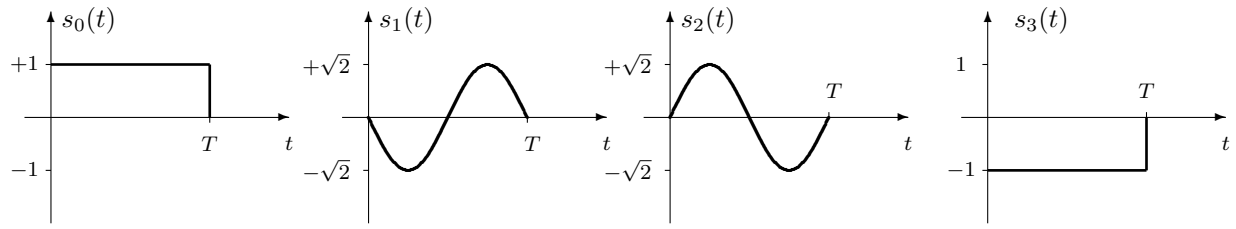


Figure 3.5: Set of 4 signals chosen for the example.

and the bits-to-symbols conversion is performed, identifying each 2-bit block according to the previously specified binary assignment:

$$B[n] = b_1 \mid b_3 \mid b_2 \mid b_0 \mid b_3 \mid b_1 \mid \cdots$$

Now the transmitted signal is generated piecewise, by symbol intervals:

- Since the first symbol is b_1, the associated waveform, $s_1(t)$, is placed in the first symbol interval ($0 \leq t < T$).
- Since the second symbol is b_3, in the second symbol interval ($T \leq t < 2T$) the associated waveform, $s_3(t)$, is placed. The signal is delayed by T seconds to be shifted into that interval.
- Since the third symbol is b_2, the third symbol interval ($2T \leq t < 3T$) contains the associated waveform, $s_2(t)$, but delayed by $2T$ seconds to be shifted into that interval.
- In general, if $B[n] = b_i$, then $s(t) = s_i(t - nT)$ in the the n-th symbol interval, $nT \leq t < T$.

Following this basic procedure, the modulated signal $s(t)$ is generated by intervals as follows

$$s(t) = \{s_1(t) \mid s_3(t-T) \mid s_2(t-2T) \mid s_0(t-3T) \mid s_3(t-4T) \mid s_1(t-5T) \mid \cdots\}$$

Figure 3.6 shows the signal resulting from applying this piecewise, by symbol interval, generation process.

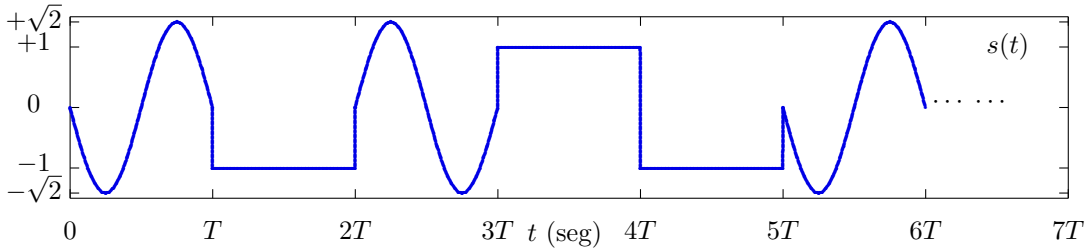


Figure 3.6: Signal generated by the example binary sequence.

In the light of this example, it can be clearly seen that the design of the digital modulator consists in the choice of M signals, which must be associated with each one of the M possible values of the symbol sequence. In this example the choice has been arbitrary. Is the example choice a good choice? What criteria must be taken into account when choosing the set of M signals? These questions will be answered throughout this chapter.

3.1.4 Transmission through a communications channel

Once the modulated signal $s(t)$ containing the digital information is generated, it is transmitted through a physical medium or communications channel, such as a cable, optical fiber, the radio spectrum or any other medium that allows the transmission of a signal. The signal is distorted during transmission so that the received signal, $r(t)$, does not match the transmitted signal: $r(t) \neq s(t)$.

There are several types of distortion that affect a signal during its transmission over a physical medium. In this subject, only the two most important ones will be considered: linear distortion and noise. The usual way of modeling linear distortion is by means of a linear and invariant system model, characterized by a certain impulse response $h(t)$ and its corresponding frequency response $H(j\omega)$, responses related through the Fourier transform. The noise is modeled as an additive process whose statistical characteristics are associated in most cases with the usual model of thermal noise: a stationary, ergodic, white, Gaussian random process with zero mean and power spectral density $N_0/2$ W/Hz, where the constant N_0 is obtained by the product of the Boltzmann constant and the temperature expressed in degrees Kelvin. Therefore, the communications channel model that will be used in this chapter will be the one defined by the relationship

$$r(t) = s(t) * h(t) + n(t)$$

which is shown in Figure 3.7.

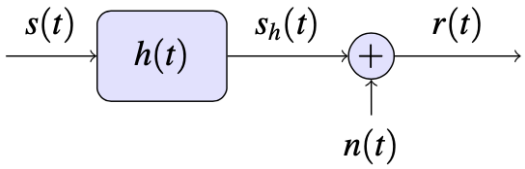


Figure 3.7: Communications channel model.

3.1.5 Generic design of a digital demodulator and basic notation

The function of the digital demodulator is the recovery of the bit sequence $B_b[\ell]$ from the signal received through the channel, $r(t)$. The signal is distorted during its transmission through the communications channel, so the received signal will be different from the transmitted signal

$$r(t) \neq s(t).$$

The basic procedure to recover the signal is presented now. The received signal is processed piecewise, specifically by symbol intervals. In the symbol interval associated to the discrete time index n, interval $nT \leq t < (n+1)T$, the received signal is observed and compared with the M waveforms of the system. The signal with which it is most "*similar*" is chosen, and if this is $s_j(t)$, the estimate for the symbol associated with that interval is the symbol associated to that waveform $\hat{B}[n] = b_j$.

Returning to the previous example, after the transmission of the signal, it suffers a certain distortion, giving rise to the signal shown in Figure 3.8.

After segmenting the signal into symbol intervals, a search is made for each symbol interval to find which of the M possible signals in the system (signals in Figure 3.5 in this example) is most

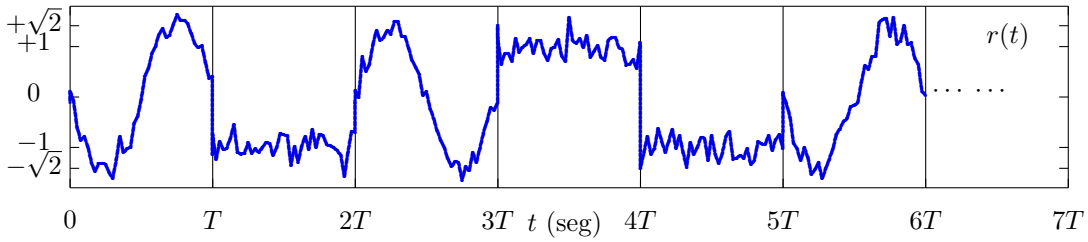


Figure 3.8: Signal received at the receiver input in the example.

similar to the received signal in the interval. In this case, a simple visual inspection allows the identification of the signal corresponding to each interval

- $n = 0$, interval $0 \leq t < T$ - "*Most similar*" signal: $s_1(t) \rightarrow \hat{B}[0] = b_1$
- $n = 1$, interval $T \leq t < 2T$ -"*Most similar*" signal: $s_3(t) \rightarrow \hat{B}[1] = b_3$
- In general, for n, interval $nT \leq (n+1)T$ - "*Most similar*" signal: $s_k(t) \rightarrow \hat{B}[n] = b_k$

Therefore, following this procedure: $\hat{B}[2] = b_2$, $\hat{B}[3] = b_0$, $\hat{B}[4] = b_3$, $\hat{B}[5] = b_1$, and the estimated symbol sequence is

$$\hat{B}[n] = b_1 \mid b_3 \mid b_2 \mid b_0 \mid b_3 \mid b_1 \mid \cdots$$

which after the symbols-to-bits conversion produces the following binary sequence

$$\hat{B}_b[\ell] : 01 \mid 11 \mid 10 \mid 00 \mid 11 \mid 01 \mid \cdots$$

In the example the distortion of the signal is moderate, so that by means of a visual inspection the original sequence is recovered without problems. But the real system must somehow measure the "*similarity*" or "*difference*" between signals. It is necessary to define a measure that quantifies the difference of the signal that is received in the n-th symbol interval with the waveform $s_i(t)$, which we could denote as

$$\text{diference}(n, i).$$

In this way, the criterion for making a decision would be to decide the symbol associated with the signal with the smallest difference (greatest similarity) with respect to the signal received in the interval

$$\hat{B}[n] = b_i \text{ if diference}(n, i) < \text{diference}(n, j) \text{ for all } j \neq i.$$

For example, a possible measure may sum the difference in modulus between the observed signal at the symbol interval and each of the possible reference signals (the sum over each time instant becomes in an integral)

$$\text{diference}_A(n, i) = \int_{nT}^{(n+1)T} |s(t) - s_i(t - nT)| \, dt.$$

Another possible measure is energy of the difference signal, which is given by

$$\text{diference}_B(n, i) = \int_{nT}^{(n+1)T} |s(t) - s_i(t - nT)|^2 \, dt.$$

Intuitively both measures somehow quantify the difference between the signals, giving lower values the more they are similar (they would give 0 only if the signals are identical). And we could think

of many more measures that could be used to quantify the similarity or difference between signals. What would be the best measure to minimize the probability of being wrong? The design of the digital demodulator consists in finding the best way to quantify this resemblance so that the estimated sequence is the one that minimizes the number of erroneous estimates.

3.1.6 Factors to consider in the selection of the M waveforms

We have just seen what basic function must be performed by the digital modulator and demodulator in a digital communications system, what procedure will be used to perform this function, and what its design consists of. Summarizing:

- Digital modulator
  - Function: Conversion of the binary information sequence $B_b[\ell]$ into an electromagnetic signal $s(t)$.
  - Procedure: Bits are grouped into blocks of m bits (symbols), each symbol is associated to a waveform out of M possible ones, which is transmitted into the interval associated with this symbol.
  - Design: Choice of an appropriate set of M signals to carry each of the M possible symbols (blocks of m bits).
- Digital demodulator
  - Function: To estimate the sequence of transmitted bits, $\hat{B}_b[\ell]$, from the received signal $r(t)$.
  - Procedure: Decision for each symbol interval, of which of the M signals was transmitted in each symbol interval in view of the signal received in said interval, comparing the shape of the signal in the interval with the M possible signals and choosing one of them.
  - Design: Determine the optimal measure to make the comparison and the choice, so as to minimize the errors made taking into account the distortion suffered by the signal in its transmission through the channel.

Several factors must be considered in the system design. In particular, when designing the digital modulator and demodulator, three main factors must be taken into account:

1. Performance: measured by means of the error probabilities associated to the estimates at the receiver, i.e., the symbol error rate and the bit error rate (P_e, BER)
   - If at the receiver, for each symbol interval, the most "*similar*" signal is decided, obviously the probability of error will depend on the "*similarity*" between signals. Therefore, it is convenient to select a set of M signals as different as possible.
   - Another problem is to determine the measure of "*similarity*", or alternative of "*difference*", that leads to minimizing the error probability taking into account the type of distortion suffered by the signal in its transmission.
2. Energy/power of the transmitted signal

- The popwer of the transmitted signal is limited in practice, since the amplifiers used in the transmitter have a power that cannot be exceeded for various kinds of reasons. Since the transmitted signal is composed of "*pieces*" at each symbol interval, and at each interval one of the M waveforms $s_i(t)$ appears, in practice a reasonable measure of the average power or energy of the modulated signal is the average of the energy of each of the signals that compose it, which is called the average or mean energy per symbol (E_s). In practice, this energy is limited. This average energy per symbol is obtained by weighting the energy of each symbol (more properly, of the signal associated with each symbol) taking into account the probability of each symbol:
  - Probability of each symbol: $p_B(b_i) = P(B[n] = b_i)$
  - Energy of symbol $b_i \equiv$ energy of signal $s_i(t)$
  - Average energy per symbol: average of the energy of the M symbols

$$E_s = \sum_{i=0}^{M-1} p_B(b_i)\, \mathcal{E}\{s_i(t)\}, \quad \text{with } \mathcal{E}\{s_i(t)\} = \int_{-\infty}^{\infty} |s_i(t)|^2\, dt.$$

3. Characteristics of the communication channel ($h(t) \overset{\mathcal{FT}}{\leftrightarrow} H(j\omega)$)

- The system must be designed to minimize the distortion suffered by the signal during transmission. As seen in Section 3.1.4, the effect of the channel is usually modeled through the relationship

$$r(t) = s(t) * h(t) + n(t).$$

  The transmitted signal fundamentally suffers two effects: a linear distortion given by the channel response (described by the impulse response $h(t)$ or the frequency response $H(j\omega)$) and the additive noise (usually thermal noise). Noise will always be present, and a filter at the input of the receiver is necessary to minimize its effect. The choice of signals $s_i(t)$, which determine the characteristics of the transmitted signal, is relevant to minimize the linear distortion. Taking into account the channel response, there will be signals that suffer greater distortion, and signals that have less distortion. The ideal situation would be zero linear distortion, which theoretically could be achieved if an appropriate set of M signals is chosen. That appropriate set would satisfy

$$s_i(t) * h(t) = s_i(t) \text{ for } i \in \{0, 1, \cdots, M-1\},$$

  or equivalently, it is easier to interpret this condition in the frequency domain

$$S_i(j\omega) \times H(j\omega) = S_i(j\omega) \text{ for } i \in \{0, 1, \cdots, M-1\}.$$

  If this condition is satisfied for the M signals, the only distortion that the transmitted signal will suffer will be the addition of noise

$$r(t) = s(t) + n(t),$$

  which is called the Additive White and Gaussian Noise (AWGN) model, or Gaussian channel model for short.

Selecting M signals jointly attending to these three factors working with signals in the time domain is complicated, since some of the factors are even in opposition. For example, it is easy to increase the difference between signals with signals of higher amplitude, but these have higher energy. Finding M signals of limited energy, as different as possible, and at the same time having

a suitable Fourier transform taking into accout the channel response, is a complicated problem. In addition, there is also the problem of determining the measure of "*difference*" that leads to minimum error probability. This problem can be made simpler by working with the geometric representation of the signals in a vector space. This representation will be presented in the next section.

3.2 Geometric representation of signals

When designing the modulator, it must associate a waveform to each symbol of the alphabet of $B[n]$, which means that each b_i is associated with a signal $s_i(t)$. If the design goal is to have the lowest probability of error, intuitively it is logical to think that these signals should be as different as possible. But it is not so easy to find a measure of difference between signals, when these are defined in the time domain, easy to handle analytically and that is appropriate to minimize the error probability.

This is the reason for using a geometric representation of the signals. Let's assume that each signal can be represented within a vector space: Each received signal can be represented by a vector, as shown in Figure 3.9 (dashed line). The set of $M = 4$ signals $\{s_0(t), s_1(t), s_2(t), s_3(t)\}$, can also be represented by vectors (continuous lines). The signal $s_i(t)$ is represented by the vector $\mathbf{a}_i$ (association based on the subindices). On the one hand, if a measure of distance between vectors is used, it is easy to determine how far or close the vectors associated with two signals are. And that measure of distance can be a good measure of similarity between signals. On the other hand, the noise will also be represented as a vector, which will be added to the signal vector to produce the vector for the received signal (in this case the vector $\mathbf{q}$ in the figure). With a measure of distance between vectors it is easy to choose the signal that is closer to the received vector (in this example, the signal $s_0(t)$, since the vector $\mathbf{q}$ is closest to the vector $\mathbf{a}_0$ than to the other three vectors).

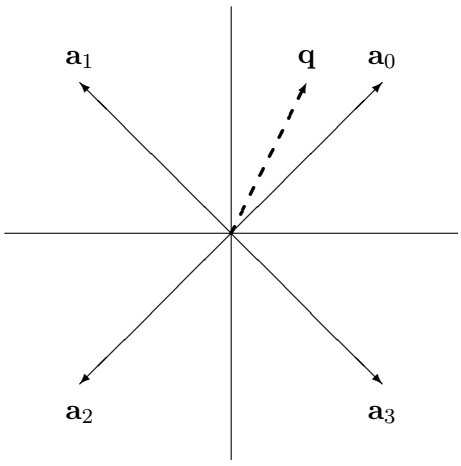


Figure 3.9: Geometric representation of signals (example of a 4-ary system in a two-dimensional space).

In this section, a vectorial representation of the signals will be presented, which is appropriate for the design and analysis of digital communications systems. Firstly, the main characteristics of vector spaces and Hilbert spaces will be reviewed, how to obtain an orthonormal basis for

the vectorial representation of a set of M signals, and how to calculate energies and difference measurements on the vector representations.

3.2.1 Vector spaces

Signals admit a representation as vectors of a vector space. As we are going to see, this representation allows us to apply all the analysis and synthesis tools of this type of space to work with the signals.

Let us first analyze whether a signal can be considered as a vector inside a vector space. A vector space $\mathbb{V}$ is a set of elements, which we call vectors, that have the following properties:

1. There is an internal composition law, which is called sum and is represented by the symbol $+$ that, applied to two vectors $(\mathbf{x}, \mathbf{y}, \in \mathbb{V})$ of the form $\mathbf{x}+\mathbf{y}$ results in another vector of the space $(\mathbf{x}+\mathbf{y} \ \in \mathbb{V})$. This law must satisfy the following properties:
   (a) Commutative: $\forall\, \mathbf{x}, \mathbf{y} \ \in \mathbb{V}; \mathbf{x}+\mathbf{y} = \mathbf{y}+\mathbf{x}$.
   (b) Associative: $\forall\, \mathbf{x}, \mathbf{y}, \mathbf{z} \ \in \mathbb{V}; \mathbf{x}+(\mathbf{y}+\mathbf{z}) = (\mathbf{x}+\mathbf{y})+\mathbf{z}$.
   (c) Existence of a neutral element (zero): $\exists\, \mathbf{0} \in \mathbb{V} : \forall\, \mathbf{x} \in V; \mathbf{x}+\mathbf{0} = \mathbf{0}+\mathbf{x} = \mathbf{x}$.
   (d) Existence of an inverse element: $\forall\, \mathbf{x} \in \mathbb{V} \exists (-\mathbf{x}) : \mathbf{x}+(-\mathbf{x}) = \mathbf{0}$.
2. There is a law of external composition that we call product with a set $\mathbb{C}$ of elements called scalars (which must have the Field structure) that, applied to a scalar α ($\alpha \in \mathbb{C}$) and a vector $\mathbf{x}$ ($\mathbf{x} \in \mathbb{V}$) of the form $\alpha\mathbf{x}$ results in another vector ($\alpha\mathbf{x} \in \mathbb{V}$). This law must satisfy the following properties:
   (a) Associative: $\forall\, \alpha, \beta \in \mathbb{C}; \forall\, \mathbf{x} \in \mathbb{V}; \alpha(\beta\mathbf{x}) = (\alpha\beta)\mathbf{x}$.
   (b) Existence of a neutral element (one): $\exists \mathbf{1} \in \mathbb{C} : \forall\, \mathbf{x} \in \mathbb{V}; \mathbf{1}\mathbf{x} = \mathbf{x}$.
   (c) Distributive with respect to sum: $\forall\, \alpha \in \mathbb{C}; \forall\, \mathbf{x}, \mathbf{y} \in \mathbb{V}; \alpha(\mathbf{x}+\mathbf{y}) = \alpha\mathbf{x}+\alpha\mathbf{y}$.
   (d) Distributive with respect to the product by a scalar:

$$\forall\, \alpha, \beta \in \mathbb{C}; \forall\, \mathbf{x} \in \mathbb{V}; (\alpha+\beta)\mathbf{x} = \alpha\mathbf{x}+\beta\mathbf{x}.$$

If the general case of a complex signal is considered (both in continuous time and in discrete time), this signal fulfills all the conditions of a vector space. The addition operation of the vector space is the point-to-point addition of the signal, which is commutative, associative, has a neutral element ($x(t) = 0$) and an inverse element, the inverted signal itself (multiplied by -1).

If the scalars are the field of complex numbers, the law of external composition is the multiplication of the signal by a complex, and it is easy to verify that all the properties are satisfied.

The case of the real signals and the real numbers also fulfill it, since in fact they are a particular case of the complex signals and complex numbers.

However, this generic vector space structure is too simple to be useful. A more elaborate structure is found in Hilbert vector spaces or simply Hilbert spaces.

3.2.2 Hilbert spaces for signals of finite energy

A Hilbert space is basically an inner product vector space[1]. The inner product, also known as scalar product or dot product, is an application of pairs of vectors in the field of scalars (complex or real in our case),

$$f : (\mathbb{V}, \mathbb{V}) \rightarrow \mathbb{C},$$

which is denoted as $\langle \mathbf{x}, \mathbf{y} \rangle$, and which satisfies the following properties:

1. $\langle \mathbf{x}, \mathbf{y} \rangle = \langle \mathbf{y}, \mathbf{x} \rangle^*$
2. $\langle (\alpha \mathbf{x} + \beta \mathbf{y}), \mathbf{z} \rangle = \alpha \langle \mathbf{x}, \mathbf{z} \rangle + \beta \langle \mathbf{y}, \mathbf{z} \rangle$
3. $\langle \mathbf{x}, \mathbf{x} \rangle \geq 0$
4. $\langle \mathbf{x}, \mathbf{x} \rangle = 0 \Leftrightarrow \mathbf{x} = \mathbf{0}$ (neutral zero vector)

The inner product induces a norm for the vector space, defined as

$$||\mathbf{x}|| = \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle},$$

and from the norm a measure of distance between vectors

$$d(\mathbf{x}, \mathbf{y}) = ||\mathbf{x} - \mathbf{y}||.$$

The angle between two vectors is measured as

$$\theta = \cos^{-1} \left(\frac{\mathrm{Re}\{\langle \mathbf{x}, \mathbf{y} \rangle\}}{||\mathbf{x}|| \, ||\mathbf{y}||} \right).$$

For signals and, in general, for any generic vector space, there is no a single definition of the inner product: any function that meets the previously established requirements can be chosen as a inner product. Each definition gives rise to a different Hilbert space, with a different norm and distance measure. Next, the structure of two Hilbert spaces for energy signals is presented, which are the most appropriate for their application in communication systems.

1. L_2: Hilbert space for energy signals in continuous time.
2. ℓ_2: Hilbert space for energy signals in discrete time.

Hilbert space for continuous time energy signals

The space L_2 is defined by the following inner product

$$\langle \mathbf{x}, \mathbf{y} \rangle = \int_{-\infty}^{\infty} x(t) \; y^*(t) \; dt.$$

[1]Strictly speaking, it is a scalar product vector space that satisfies the completeness property. Completeness is fulfilled when every Cauchy sequence is convergent in the metric induced by the scalar product. If it does not have this property, the vector space is called a pre-Hilbert space.

The norm induced by this inner product is

$$||\mathbf{x}|| = \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle} = \sqrt{\int_{-\infty}^{\infty} |x(t)]|^2 \, dt} = \sqrt{\mathcal{E}\{x(t)\}},$$

i.e., it is the square root of the energy of the signal. This is interesting because the norm of a vector defines the distance of the vector representation from the origin of coordinates, so the energy of a signal can be easily evaluated as the squared distance of its vector representation from the origin of coordinates.

The distance between the vector representation of two signals is given by

$$d(\mathbf{x}, \mathbf{y}) = ||\mathbf{x} - \mathbf{y}|| = \sqrt{\int_{-\infty}^{\infty} |x(t) - y(t)|^2 \, dt} = \sqrt{\mathcal{E}\{x(t) - y(t)\}}.$$

In this case, we have the square root of the energy of the difference signal, which intuitively seems like a reasonable quantitative measure of the difference between two signals (later we will see that in certain cases it is the optimal measure to decide with the minimum error probability in a digital demodulator).

Hilbert space for discrete-time energy signals

The space ℓ_2 is defined by the following inner product

$$\langle \mathbf{x}, \mathbf{y} \rangle = \sum_{n=-\infty}^{\infty} x[n] \, y^*[n].$$

The norm induced by this inner product is

$$||\mathbf{x}|| = \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle} = \sqrt{\sum_{n=-\infty}^{\infty} |x[n]|^2} = \sqrt{\mathcal{E}\{x[n]\}},$$

and the distance is the well known Euclidean distance

$$d(\mathbf{x}, \mathbf{y}) = ||\mathbf{x} - \mathbf{y}|| = \sqrt{\sum_{n=-\infty}^{\infty} |x[n] - y[n]|^2} = \sqrt{\mathcal{E}\{x[n] - y[n]\}},$$

that is also related with the energy of the difference signal.

The basic definitions are equivalent in both spaces, with the only difference being the continuous or discrete nature of the time index (t or n, respectively).

Some aspects of interest

In any case, the inner product provides a measure of the resemblance or similarity between two signals. The inner product of two signals with a similar waveform will be "large", while that of two signals with very different waveforms will be "small".

When the inner product of two signals is equal to zero, those signals are said to be *orthogonal*. This means that the corresponding vector representations have a 90 degrees angle between them.

There is an important relationship between the vector norms and the magnitude of the inner product. The Cauchy-Schwarz inequality makes use of this relationship and states that

$$|\langle \mathbf{x}, \mathbf{y} \rangle| \leq ||\mathbf{x}||\ ||\mathbf{y}||.$$

The equality only holds if the vector $\mathbf{y}$ is a scaled version of the vector $\mathbf{x}$, which in the case of signals means that $y(t) = \alpha\ x(t)$ or $y[n] = \alpha\ x[n]$. This inequality takes the following form for the L_2 and ℓ_2 spaces

$$\left| \int_{-\infty}^{\infty} x(t)\ y^*(t)\ dt \right| \leq \sqrt{\int_{-\infty}^{\infty} |x(t)|^2\ dt} \sqrt{\int_{-\infty}^{\infty} |y(t)|^2\ dt}$$

and

$$\left| \sum_{n=-\infty}^{\infty} x[n]\ y^*[n] \right| \leq \sqrt{\sum_{n=-\infty}^{\infty} |x[n]|^2} \sqrt{\sum_{n=-\infty}^{\infty} |y[n]|^2},$$

respectively.

3.2.3 Representation of vectors in a basis

The inner product also makes it possible to easily find the representation of a signal in a basis of the vector space.

A basis for a Hilbert space $\mathbb{H}$ of dimension D is a subset of D elements $\{\phi_n\} \in \mathbb{H}$, which determine a set of D unique coefficients $\{c_n(\mathbf{x})\}$, $n \in \{0, 1, \cdots, D-1\}$, for any element $\mathbf{x}$ in the space, such that the element can be uniquely represented as a linear expansion over the elements of the basis

$$\mathbf{x} = \sum_{n=0}^{D-1} c_n(\mathbf{x})\ \phi_n.$$

The D coefficients that uniquely define the vector $\mathbf{x}$, $\{c_n(\mathbf{x})\}_{n=0}^{D-1}$, are usually called coordinates of the vector in the basis.

A basis is *orthogonal* when the elements of the basis are orthogonal to each other, which means that their inner product is zero

$$\langle \phi_n, \phi_m \rangle = 0,\ \forall\ n \neq m.$$

A basis is *orthonormal* when it is an orthogonal basis, that is, it is true that the elements of the basis are orthogonal, and it is also true that each of them has unit norm. Taking into account the definition of norm through the inner product, this is equivalent to saying that the inner product of each element of the base with itself is one

$$||\phi_n|| = 1 \rightarrow \langle \phi_n, \phi_n \rangle = 1.$$

In Hilbert spaces, one of the advantages of working with an orthonormal basis is that the coefficients of the expansion in terms of the basis (coordinates) are obtained by means of the inner product of the vector with the different elements that form the orthonormal basis

$$c_n(\mathbf{x}) = \langle \mathbf{x}, \phi_n \rangle.$$

In the case of the L_2 space, an example of a trivial orthonormal basis is the set of Dirac delta functions $\{\delta(t-\tau),\ \tau \in I\!R\}$. If $\phi_i = \delta(t-\tau_i)$,

$$\langle \phi_i, \phi_j \rangle = \int_{-\infty}^{\infty} \delta(t-\tau_i)\delta(t-\tau_j)\ dt = \begin{cases} 1, & \tau_i = \tau_j \\ 0, & \tau_i \neq \tau_j \end{cases}.$$

The representation of the signal in terms of this basis is

$$\mathbf{x} \equiv x(t) = \int_{-\infty}^{\infty} x(\tau)\delta(t-\tau)\ d\tau.$$

The coefficients of the expansion are obtained as

$$c_i(\mathbf{x}) = \langle \mathbf{x}, \delta(t-\tau_i) \rangle = \int_{-\infty}^{\infty} x(t)\ \delta(t-\tau_i)\ dt = x(\tau_i).$$

The same is true for the ℓ_2 space, where the set $\{\delta[n-k], k \in \mathbb{Z}\}$, forms an orthonormal basis. In this case it can be denoted as $\phi_k = \delta[n-k]$

$$\langle \phi_k, \phi_i \rangle = \sum_{n=-\infty}^{\infty} \delta[n-k]\delta[n-i] = \delta[k-i] = \begin{cases} 1, & k = i \\ 0, & k \neq i \end{cases}.$$

The representation of the signal in terms of this basis is

$$\mathbf{x} \equiv x[n] = \sum_{k=-\infty}^{\infty} x_k \delta[n-k].$$

The coefficients of the expansion are obtained as

$$c_k(\mathbf{x}) = \langle \mathbf{x}, \phi_k \rangle = \sum_{n=-\infty}^{\infty} x[n]\ \delta[n-k] = x[k].$$

These are valid bases, but non-practical because of the infinite dimension. In the next section, we will see how to obtain a finite dimension basis to represent a set of M signals.

3.2.4 Gram-Schmidt orthogonalization procedure

The two orthonormal bases presented above are not practical since they involve infinitely many coefficients, which means knowing the complete signal in practice. It is possible to obtain a finite-dimension orthonormal basis that allows to represent a finite set of elements of the vector space (in the particular case that affects the design and analysis of systems of communications, a finite-dimension basis that allows representing M signals). A suitable basis for this type of representation can be obtained by means of the so-called *Gram-Schmidt orthogonalization* method. This procedure allows to obtain an orthonormal basis of $N \leq M$ elements to represent a set of M signals. In this way it is possible to represent each signal as an N-dimensional vector, in the space defined by that basis.

In a communications system there is a set of M signals $\{s_i(t), i = 0, \cdots, M-1\}$, of duration T seconds and defined in the interval $0 \leq t < T$, which are used to transmit information over

the communication channel. From this set, another set of N orthonormal signals $\{\phi_j(t), j = 0, \cdots, N-1\}$, with $N \leq M$, can be obtained. They allow to represent each signal $s_i(t)$ by the following linear combination

$$s_i(t) = \sum_{j=0}^{N-1} a_{i,j}\ \phi_j(t).$$

According to the definition of an orthonormal basis, any signal $s_i(t)$ can be expressed as a linear combination of the elements of the basis where the coefficients of the expansion, or coordinates of the vector representation for the signal $s_i(t)$, $a_{i,j}$, can be obtained by the dot product

$$a_{i,j} = \langle s_i(t), \phi_j(t) \rangle = \int_{-\infty}^{\infty} s_i(t)\ \phi_j^*(t)\ dt.$$

This results in a vector representation (N dimensional) of each signal via the coordinates vector

$$s_i(t) \rightarrow \mathbf{a}_i = \begin{bmatrix} a_{i,0} \\ a_{i,1} \\ \vdots \\ a_{i,N-1} \end{bmatrix}.$$

Although at first glance it may seem that this change in representation does not represent a great advance, it must be taken into account that in this way N is the dimension of the signal space, and each signal is represented by N coordinates, which is more convenient than handling the expression of the signal in the continuous time domain. In addition, it also simplifies the calculation of the energy of the signals $s_i(t)$ and their difference.

It is important to remember that since it is an orthonormal basis, the inner product of two different elements of the basis is zero, and the inner product of an element of the basis with itself is one, i.e.,

$$\langle \phi_i(t), \phi_j(t) \rangle = \int_{-\infty}^{\infty} \phi_i(t)\ \phi_j^*(t)\ dt = \delta[i-j].$$

The inner product between two signals $s_i(t)$ and $s_k(t)$ is calculated as

$$\begin{aligned}
\langle s_i(t), s_k(t) \rangle &= \int_{-\infty}^{\infty} s_i(t)\ s_k^*(t)\ dt \\
&= \int_{-\infty}^{\infty} \left(\sum_{j=0}^{N-1} a_{ij}\ \phi_j(t) \right) \left(\sum_{\ell=0}^{N-1} a_{k\ell}^*\ \phi_\ell^*(t) \right) dt \\
&= \int_{-\infty}^{\infty} \left(\sum_{j=0}^{N-1} \sum_{\ell=0}^{N-1} a_{ij}\ a_{k\ell}^*\ \phi_j(t)\ \phi_\ell^*(t) \right) dt \\
&= \sum_{j=0}^{N-1} a_{ij}\ a_{kj}^* \int_{-\infty}^{\infty} \phi_j^2(t)\ dt + \sum_{j=0}^{N-1} \sum_{\substack{\ell=0 \\ \ell \neq j}}^{N-1} a_{ij} a_{k\ell}^* \int_{-\infty}^{\infty} \phi_j(t) \phi_\ell(t)^*\ dt,
\end{aligned}$$

which applying the orthonormality property of the basis reduces to

$$\langle \mathbf{s}_i, \mathbf{s}_k \rangle = \langle s_i(t), s_k(t) \rangle = \sum_{j=0}^{N-1} a_{ij}\ a_{kj}^*.$$

The energy of a signal, which in the time domain involves an integral

$$\mathcal{E}_i = \mathcal{E}\left\{s_i(t)\right\} = \int_{-\infty}^{\infty} |s_i(t)|^2 \ dt,$$

can now be calculated from the vectorl representation of the signal in a simpler way

$$\mathcal{E}_i = \mathcal{E}\left\{s_i(t)\right\} = \langle \mathbf{s}_i, \mathbf{s}_i \rangle = ||s_i(t)||^2 = \sum_{j=0}^{N-1} |a_{ij}|^2.$$

The calculation from integrals of the signals in the time domain has been replaced by a sum of N squared coordinates, much easier to perform.

This comment is also valid for the calculation of the difference between signals. An intuitively reasonable measure (later we will also see that in many cases it is the optimal measure) is the energy of the difference signal, which in the time domain again requires an integral

$$\int_{-\infty}^{\infty} |s_i(t) - s_k(t)|^2 \ dt.$$

With the definition of the inner product for the $\mathcal{L}_2$ space, the square root of this energy corresponds to the norm of the difference vector, i.e., the distance between the vector representation of the signals. In short, this intuitive measure of difference can be obtained from the norm of the difference vector

$$d(s_i(t), s_k(t)) = ||s_i(t) - s_k(t)|| = \sqrt{\int_{-\infty}^{\infty} |s_i(t) - s_k(t)|^2 \ dt}.$$

Taking into account the linear expansion for both signals

$$d(s_i(t), s_k(t)) = \sqrt{\langle \mathbf{s}_i - \mathbf{s}_k, \mathbf{s}_i - \mathbf{s}_k \rangle} = \sqrt{\sum_{j=0}^{N-1} |a_{ij} - a_{k,j}|^2}.$$

Once again, integrals over signals have been replaced by sums over N coordinates. Therefore, the vector representation of signals by coordinates in an orthonormal basis is useful.

To obtain an orthonormal basis that allows representing M signals, the Gram-Schmidt orthogonalization process begins selecting a non-zero energy signal, $s_0(t)$. The first element of the basis is obtained simply by normalizing the first signal

$$\phi_0(t) = \frac{s_0(t)}{\sqrt{\mathcal{E}_0}}.$$

This ensures that the energy of $\phi_0(t)$, $\mathcal{E}\{\phi_0(t)\} = 1$, i.e., we have the first element of an orthonormal basis.

To obtain the second element of the basis, first the projection of $s_1(t)$ on $\phi_0(t)$ is obtained. The coordinate of this signal on the first element of the basis is

$$a_{1,0} = \int_{-\infty}^{\infty} s_1(t) \ \phi_0^*(t) \ dt.$$

Then the projection over $\phi_0(t)$ is subtracted from $s_1(t)$

$$d_1(t) = s_1(t) - a_{1,0}\phi_0(t).$$

Now $d_1(t)$ is orthogonal to $\phi_0(t)$, and it must be normalized. If $\mathcal{E}_1$ denotes the energy of $d_1(t)$

$$\mathcal{E}_1 = \mathcal{E}\{d_1(t)\} = \int_{-\infty}^{\infty} |d_1(t)|^2 \, dt,$$

the second element of the orthonormal basis is

$$\phi_1(t) = \frac{d_1(t)}{\sqrt{\mathcal{E}_1}},$$

which also has unit energy.

In general, obtaining the element $k+1$ of the basis, $\phi_k(t)$ is obtained as

$$\phi_k(t) = \frac{d_k(t)}{\sqrt{\mathcal{E}_k}},$$

where

$$d_k(t) = s_k(t) - \sum_{j=0}^{k-1} a_{k,j} \, \phi_j(t),$$

$$\mathcal{E}_k = \mathcal{E}\{d_k(t)\} = \int_{-\infty}^{\infty} |d_k(t)|^2 \, dt,$$

and

$$a_{k,j} = \int_{-\infty}^{\infty} s_k(t) \, \phi_j^*(t) \, dt.$$

Below is an example of the calculation of a basis that allows to represent 4 signals.

Example

The Gram-Schmidt procedure is applied to the set of signals in Figure 3.10.

To obtain the first element of the base we calculate the energy of $s_0(t)$. It is easy to check that $\mathcal{E}_0 = 2$. Therefore

$$\phi_0(t) = \frac{s_0(t)}{\sqrt{2}}.$$

Then the projection of $s_1(t)$ over $\phi_0(t)$ is calculated.

$$a_{1,0} = \int_{-\infty}^{\infty} s_1(t) \, \phi_0^*(t) \, dt = 0.$$

So in this case $d_1(t) = s_1(t)$, which is orthogonal to $\phi_0(t)$, and must be normalized. Its energy is

$$\mathcal{E}_1 = 2,$$

and therefore

$$\phi_1(t) = \frac{s_1(t)}{\sqrt{2}}.$$

To calculate $\phi_2(t)$, the projections of $s_2(t)$ over $\phi_0(t)$ and $\phi_1(t)$ are obtained. The coordinates are

$$a_{2,0} = \int_{-\infty}^{\infty} s_2(t) \, \phi_0^*(t) \, dt = 0.$$

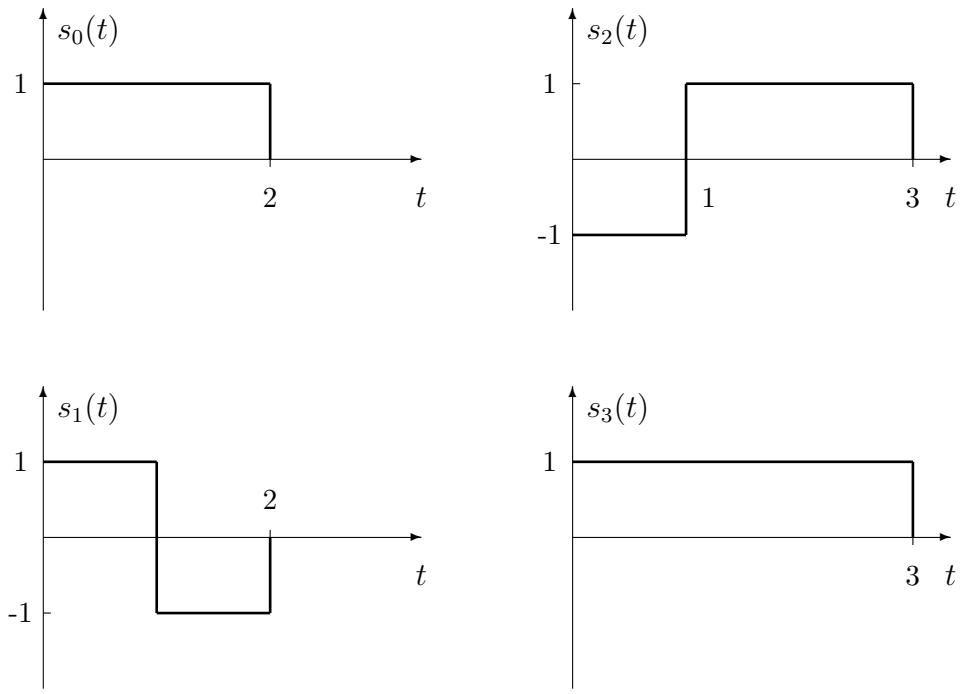


Figure 3.10: Set of $M = 4$ signals.

and

$$a_{2,1} = \int_{-\infty}^{\infty} s_2(t)\ \phi_1^*(t)\ dt = -\sqrt{2}.$$

Therefore

$$\begin{aligned} d_2(t) &= s_2(t) - a_{20}\ \phi_0(t) - a_{21}\ \phi_1(t) \\ &= s_2(t) + \sqrt{2}\phi_1(t). \end{aligned}$$

The energy of $d_2(t)$ is

$$\mathcal{E}_2 = 1,$$

and

$$\phi_2(t) = d_2(t) = s_2(t) + \sqrt{2}\phi_1(t).$$

Now the last signal, $s_3(t)$, is processed. The projections of this signal over the 3 elements of the basis are obtained. The coordinates are

$$a_{3,0} = \int_{-\infty}^{\infty} s_3(t)\ \phi_0^*(t)\ dt = \sqrt{2}.$$

$$a_{3,1} = \int_{-\infty}^{\infty} s_3(t)\ \phi_1^*(t)\ dt = 0.$$

$$a_{3,2} = \int_{-\infty}^{\infty} s_3(t)\ \phi_2^*(t)\ dt = 1.$$

Now

$$d_3(t) = s_3(t) - a_{3,0}\ \phi_0(t) - a_{3,1}\ \phi_1(t) - a_{3,2}\ \phi_2(t) = 0.$$

This means that no additional element is necessary in the basis to be able to represent $s_3(t)$, since this signal can be obtained as a linear combination of the 3 current elements. So, in this case, the dimension of this space of signals is 3, and the 3 basis functions are shown in Figure 3.11.

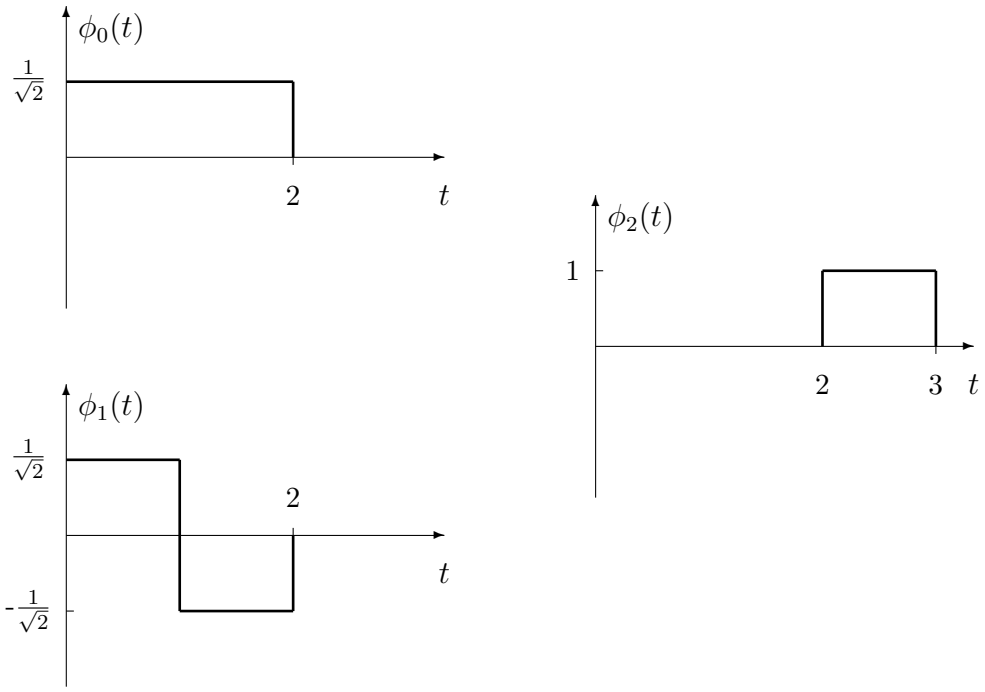


Figure 3.11: Set of orthonormal signals.

Once the N signals of the orthonormal basis have been calculated

$$\{\phi_j(t), j = 0, \cdots, N-1\}$$

each signal can be obtained as a linear combination of the N elements of the basis

$$s_i(t) = \sum_{j=0}^{N-1} a_{ij}\phi_j(t), \quad i = 0, \cdots, M-1.$$

Therefore, each signal can be represented by a vector

$$\mathbf{a}_i = \begin{bmatrix} a_{i,0} \\ a_{i,1} \\ \vdots \\ a_{i,N-1} \end{bmatrix}.$$

or equivalently, as a point in an N-dimensional signal space with coordinates $(a_{i,0}, a_{i,1}, \cdots, a_{i,N-1})$.

On the other hand, the energy of $s_i(t)$, $\mathcal{E}_i$

$$\mathcal{E}_i = \int_{-\infty}^{\infty} |s_i(t)|^2 \, dt,$$

can be easily calculated from these coordinates

$$\mathcal{E}_i = \langle \mathbf{s}_i, \mathbf{s}_i \rangle = \sum_{j=0}^{N-1} |a_{i,j}|^2.$$

Example

The coordinates of the 4 signals of the previous example (which have been calculated during the development of the orthogonalization procedure) are

$$\left.\begin{array}{l} a_{0,0} = \sqrt{2} \\ a_{0,1} = 0 \\ a_{0,2} = 0 \end{array}\right\} \rightarrow \mathbf{a}_0 = [\sqrt{2}, 0, 0]^T.$$

$$\left.\begin{array}{l} a_{1,0} = 0 \\ a_{1,1} = \sqrt{2} \\ a_{1,2} = 0 \end{array}\right\} \rightarrow \mathbf{a}_1 = [0, \sqrt{2}, 0]^T.$$

$$\left.\begin{array}{l} a_{2,0} = 0 \\ a_{2,1} = -\sqrt{2} \\ a_{2,2} = 1 \end{array}\right\} \rightarrow \mathbf{a}_2 = [0, -\sqrt{2}, 1]^T.$$

$$\left.\begin{array}{l} a_{3,0} = \sqrt{2} \\ a_{3,1} = 0 \\ a_{3,2} = 1 \end{array}\right\} \rightarrow \mathbf{a}_3 = [\sqrt{2}, 0, 1]^T.$$

These signals can be represented as points in 3D space, such as and as shown in Fig. 3.12

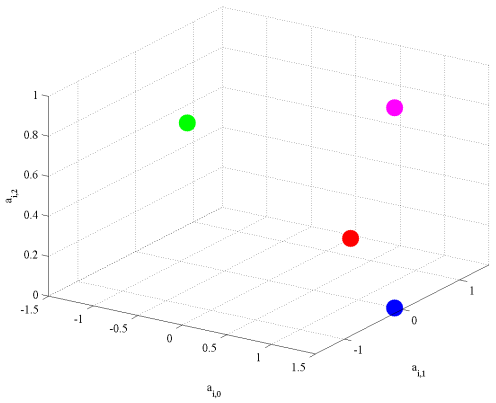


Figure 3.12: Signals represented in the orthonormal basis.

It is easy to check that

$$\mathcal{E}_0 = 2,\ \mathcal{E}_1 = 2,\ \mathcal{E}_2 = 3,\ \mathcal{E}_3 = 3.$$

The distances between signals can also be calculated in a simple way. In this case they are as follows

$$d(\mathbf{s}_0, \mathbf{s}_1) = 2,\ d(\mathbf{s}_0, \mathbf{s}_2) = \sqrt{5},\ d(\mathbf{s}_0, \mathbf{s}_3) = 1$$
$$d(\mathbf{s}_1, \mathbf{s}_2) = \sqrt{9},\ d(\mathbf{s}_1, \mathbf{s}_3) = \sqrt{5},\ d(\mathbf{s}_2, \mathbf{s}_3) = 2$$

It is important to note that the basis for representing a set of M signals is not unique. If the Gram-Schmidt procedure is applied but changing the order of the signals, the obtained absis and the corresponding coordinates are different. However, the energies and distances between signals are maintained. This means that the new representation is just a rotated version of any other valid basis.

In some cases, as in this simple example, it is possible to find a set of signals a suitable basis by visual inspection. For example, for these signals the following signals could be used as basis:

$$\phi_0(t) = \begin{cases} 1, & \text{si } 0 \leq t < 1 \\ 0, & \text{en otro caso} \end{cases}.$$

$$\phi_1(t) = \begin{cases} 1, & \text{si } 1 \leq t < 2 \\ 0, & \text{en otro caso} \end{cases}.$$

$$\phi_2(t) = \begin{cases} 1, & \text{si } 2 \leq t < 3 \\ 0, & \text{en otro caso} \end{cases}.$$

They form a valid orthonormal basis: they are orthogonal (they are non-overlapping, therefore the integran in the inner product is zero) and all of them has unit energy.

In this case, it is trivial to see that the coordinates vectors of the 4 signals in this new basis are

$$\mathbf{a}'_0 = \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix} \quad \mathbf{a}'_1 = \begin{bmatrix} 1 \\ -1 \\ 0 \end{bmatrix} \quad \mathbf{a}'_2 = \begin{bmatrix} -1 \\ 1 \\ 1 \end{bmatrix} \quad \mathbf{a}'_3 = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$$

It is easy to check how energies (which are related to the distance of each point to the origin) and distances between vector representations are maintained using this new representation.

Therefore, the corresponding vector representation assumes a rotation of the representation using the other orthonormal basis, as can be seen in Figure 3.13, showing the vector representation of the 4 signals in the new orthonormal basis.

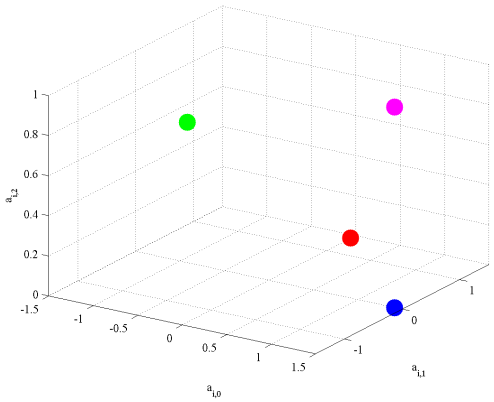


Figure 3.13: Signals represented in the alternative orthonormal basis.

3.3 Digital communication model

The vector representation of the signals in an orthonormal N-dimensional basis allows the use of a digital communication model based on this representation. This model notably simplifies the design and analysis of the system. This model has 4 functional elements (in addition to the communications channel), as shown in Figure 3.14.

In this model, both the digital modulator and the digital demodulator are divided into two modules:

- Digital modualtor:
  - Encoder

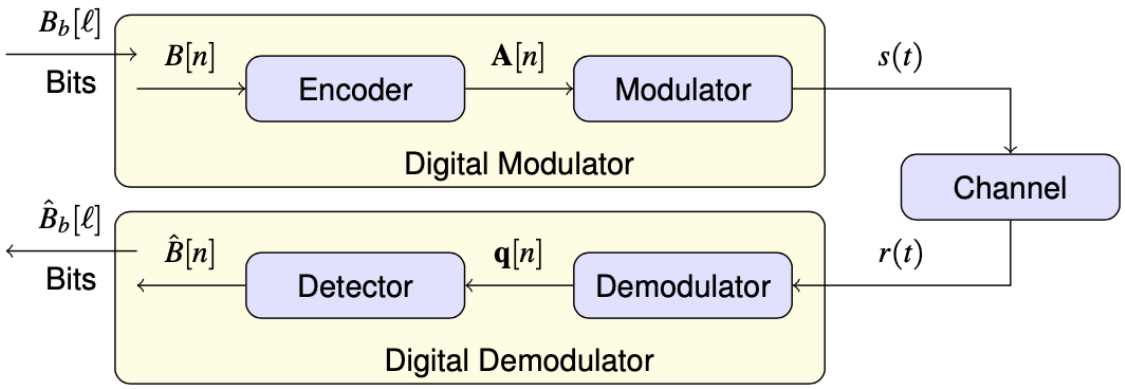


Figure 3.14: Basic model of a digital communications system.

  - Modulator

- Digital demodulator:
  - Demodulator
  - Detector

It must be remembered that the task of the digital modulator and demodulator was to convert bits (grouped into symbols) into signals and to do the inverse conversion, respectively. With this division into modules, a vectorial representation of the intermediate signals has been introduced in both cases, $\mathbf{A}[n]$ in the transmitter and $\mathbf{q}[n]$ in the receiver.These intermediate vector representations greatly facilitate the design and analysis of the system.

In the transmitter, the assignment of a signal $s_i(t)$ to each value of the alphabet of symbols b_i is now done in two steps:

- An assignment is made of the vector representation of $s_i(t)$, $\mathbf{a}_i$, to b_i (encoder).
- The vector representation of the signal, $\mathbf{a}_i$, is converted to signal $s_i(t)$ (modulator). To do this, an orthonormal basis of dimension N must be defined, since

$$s_i(t) = \sum_{j=0}^{N-1} a_{i,j} \times \phi_j(t).$$

This division considerably simplifies the problem of choosing the M signals to take into account the 3 factors that were specified in Section 3.1.6: performance (similarity/difference between signals), energy and channel features. The first two factors are uncoupled from the third one. The energy and the similarity between signals can be evaluated from the vector representation of the signals, regardless of the chosen orthonormal basis. In this way, the design of the encoder, which consists in choosing the vector representation of the M signals,

$$\{\mathbf{a}_0, \mathbf{a}_1, \cdots, \mathbf{a}_{M-1}\},$$

which will usually be called constellation. These vector are chosen taking into account two factors: performance (measure of difference between the signals) and energy.

Regarding the characteristics of the channel, it is necessary to ensure that the signals have a frequency response appropriate to the frequency response of the channel. The form of the frequency response of the signal $s_i(t)$, given that

$$s_i(t) = \sum_{j=0}^{N-1} a_{i,j} \times \phi_j(t) \leftrightarrow S_i(j\omega) = \sum_{j=0}^{N-1} a_{i,j} \times \Phi_j(j\omega),$$

depends on the choice of the orthonormal basis. Choosing an orthonormal basis or another has no impact on the energy and difference measure (distance) between signals, since these parameters are completely defined by the vector representation of the signals. The shape of the signals, both in the time and frequency domains depend only on the basis. In this way, the design of the modulator, which consists of choosing an appropriate orthonormal basis, is carried out taking into account only the third factor: the signals are adequate to the channel characteristics.

At the receiver, the division into two modules also simplifies its functionality. First, the demodulator obtains the vector representation of the signal that is received in each symbol interval, $\mathbf{q}[n]$ for the interval $nT \leq t < T$. From this representation, it is much easier to find the optimal rule to decide which of the M signals was transmitted in that interval (the rule that minimizes the error probability) than working with the continuous-time representation of the signals.

The function and main characteristics of each of the 4 functional elements of the system are briefly summarized below:

- Encoder
  - Defines the vector representation of the signal associated with each symbol (constellation)
    - Discrete time index n: vector $\mathbf{A}[n]$ representing $s(t)$ in $nT \leq t < (n+1)T$
  - Design criteria (to select the constellation)
    - Energy
    - Performance: distance ("*disimilarity*") between signals
- Modulator
  - Defines the orthonormal basis of the signal space
  - Design criteria (to select the N functions $\phi_j(t)$, for $i = 0, 1, \cdots, N-1$)
    - Adaptation to the characteristics of the communication channel
- Demodulator
  - Converts the received signal, by symbol intervals, into vectors in the signal space defined by the basis $\{\phi_j(t)\}_{j=0}^{N-1}$
    - Discrete time index n: vector $\mathbf{q}[n]$ representing $r(t)$ in $nT \leq t < (n+1)T$
- Detector
  - Compares the "*disimilarity*" between the received signal and the M possible signals $s_i(t)$ to decide symbols
    - Distance are measure over vector representations
    - Distances are obtained between:
      - Vector of the received signal in the symbol interval: $\mathbf{q}[n]$
      - Vectors of the M possible symbols: $\mathbf{a}_i$, for $i \in \{0, 1, \cdots, M-1\}$

3.3.1 Example to illustrate the advantage of vector representation in the design of a system

We will start by considering a binary communication system in which each symbol corresponds to 1 bit ($m = 1$, $M = 2^m = 2$) and the two bits are transmitted with the same probability. As signals for the transmission of each symbol, there are four candidate sets of signals $\{s_0(t), s_1(t)\}$. These sets are plotted in Figure 3.15.

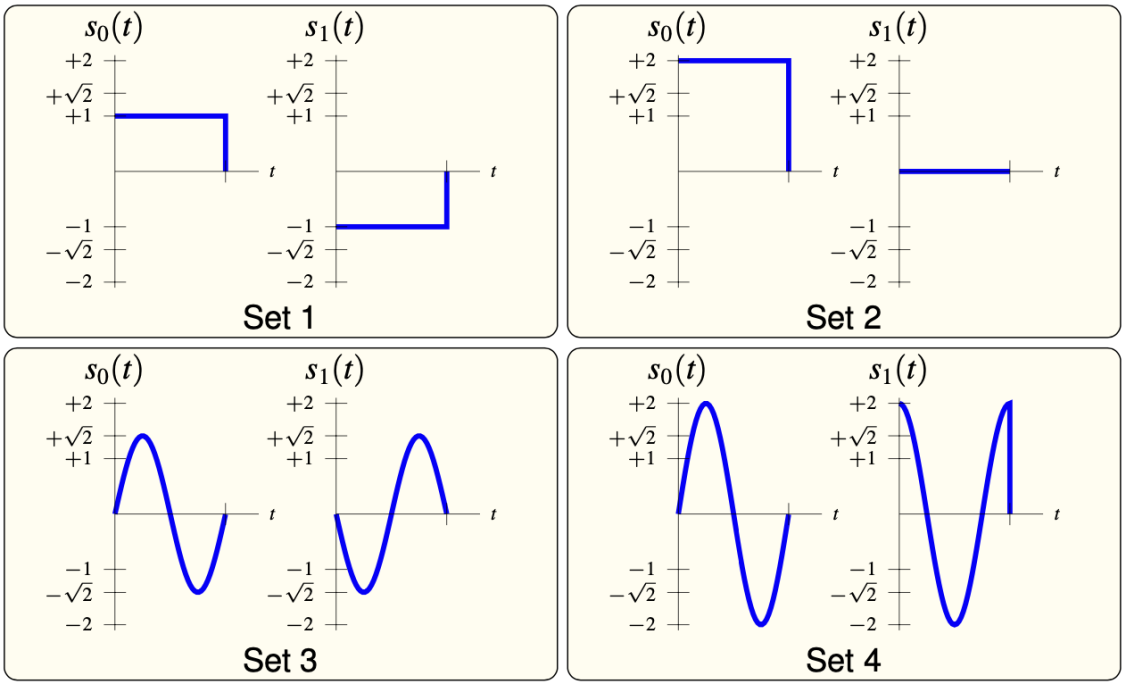


Figure 3.15: Sets of condidate signals for a binary system

To decide which set of signals is the most appropriate, it must be remembered that the objective is to recover the symbolwith minimum error probability, taking into account the characteristics of the channel. White and Gaussian noise is added. It seems logical to think that the greater the "disimilarity" or "difference" between $s_0(t)$ and $s_1(t)$, the easier to discern between them in the receiver despite the added noise. To measure this disimilarity, a measure of distance between signals is needed.

The distance measure defined in the Hilbert space of finite energy signals, L_2, will be used. In this space, the signals $s_i(t)$ become vectors $\mathbf{a}_i$ and the distance is defined as

$$d(s_i(t), s_j(t)) = ||\mathbf{a}_i - \mathbf{a}_j|| = \sqrt{\int_{-\infty}^{\infty} |s_i(t) - s_j(t)|^2 dt}.$$

Applying this measure to the two signals of the first set

$$d(s_0(t), s_1(t)) = \sqrt{\int_0^T |1 - (-1)|^2 dt} = 2\sqrt{T}.$$

For the second set of signals we have

$$d(s_0(t), s_1(t)) = \sqrt{\int_0^T |2 - 0|^2 dt} = 2\sqrt{T}.$$

For the third set of signals

$$\begin{aligned} d(s_0(t), s_1(t)) =& \sqrt{\int_0^T \left| \sqrt{2}\sin\left(\frac{2\pi t}{T}\right) - \left(-\sqrt{2}\sin\left(\frac{2\pi t}{T}\right)\right)\right|^2 dt} \\ =& \sqrt{\int_0^T 8\sin^2\left(\frac{2\pi t}{T}\right) dt} \\ =& \sqrt{4\left[t - \frac{T}{2\pi}\sin\left(\frac{2\pi t}{T}\right)\cos\left(\frac{2\pi t}{T}\right)\right]_0^T} \\ =& 2\sqrt{T}. \end{aligned}$$

And for the last set of signals

$$\begin{aligned} d(s_0(t), s_1(t)) =& \sqrt{\int_0^T \left| 2\sin\left(\frac{2\pi t}{T}\right) - \left(2\cos\left(\frac{2\pi t}{T}\right)\right)\right|^2 dt} \\ =& \sqrt{\int_0^T 4 - 8\sin\left(\frac{2\pi t}{T}\right)\cos\left(\frac{2\pi t}{T}\right) dt} \\ =& 2\sqrt{T}, \end{aligned}$$

because

$$\int_0^T 8\sin\left(\frac{2\pi t}{T}\right)\cos\left(\frac{2\pi t}{T}\right) dt = \left[\frac{2T}{\pi}\sin^2\left(\frac{2\pi t}{T}\right)\right]_0^T = 0.$$

Thus, the four sets have the same distance and, in principle, the signals would behave in the same way when faced with the disturbance introduced by the channel. But we must also ask ourselves if it takes the same effort for the transmitter to send them through the channel. And the way to measure the effort is the *average energy per symbol* which is defined as

$$\begin{aligned} E_s =& E\left[\mathcal{E}\{s(t)\}\right] \\ =& E\left[\int_{-\infty}^{\infty} |s(t)|^2 dt\right] \\ =& \sum_{i=0}^{M-1} P\left(s(t) = s_i(t)\right) \times \mathcal{E}\{s_i(t)\} \\ =& \sum_{i=0}^{M-1} p_A(\mathbf{a}_i) \int_{-\infty}^{\infty} |s_i(t)|^2 dt. \end{aligned}$$

If the symbols are equiprobable, $p_A(\mathbf{a}_i) = \frac{1}{M}$, so

$$E_s = \sum_{i=0}^{M-1} \frac{1}{M} \int_{-\infty}^{\infty} |s_i(t)|^2 dt.$$

Applying this definition to the first set of signals

$$E_s = \frac{1}{2}\int_0^T |1|^2 dt + \frac{1}{2}\int_0^T |-1|^2 dt = \frac{1}{2}T + \frac{1}{2}T = T.$$

If we calculate this energy for the rest of the sets we obtain, respectively, $E_s = 2T$, $E_s = T$ and $E_s = 2T$. It can therefore be said that the best sets are the first and the third, since they have the same distance with the lowest energy.

When handling sets of signals with different distance between signals, it is necessary to use some kind of normalization that allows us to establish their comparison. The most common of these normalizations consists of expressing the distance between signals as a function of the average energy per symbol. For the first set of symbols, this normalized distance is

$$d(s_0(t), s_1(t)) = 2\sqrt{T} = 2\sqrt{E_s}.$$

For the rest of the sets, respectively

$$d(s_0(t), s_1(t)) = 2\sqrt{T} = \sqrt{2E_s},$$

$$d(s_0(t), s_1(t)) = 2\sqrt{T} = 2\sqrt{E_s},$$

and

$$d(s_0(t), s_1(t)) = 2\sqrt{T} = \sqrt{2E_s}.$$

Clearly, sets of signals 1 and 3 provide a greater separation distance between signals than sets 2 and 4 for the same value of E_s.

Another question to consider is whether it is possible to find a way to represent the signals that allows us to see in a simpler way the benefits that we are going to obtain with a certain set. The answer is found in the same vector space structure that has allowed us to measure the distance between signals. if an orthonormal basis allows to represent a certain set of signals, it is possible to work directly with the coordinates in that basis, thus avoiding any type of calculation on the waveforms.

Expressed formally, a representation is sought for each of the signals in the set $\{s_i(t) : i = 0, \cdots, M-1\}$ of the form

$$s_i(t) = \sum_{j=0}^{N-1} a_{i,j}\phi_j(t),$$

where $\{\phi_j(t) : i = 0, \cdots, N-1\}$ are the orthonormal basis elements, and $\{a_{i,j} : j = 0, \cdots, N-1\}$ are the coordinates of the signal $s_i(t)$ in that basis, which are grouped in the vector $\mathbf{a}_i$. Thus, the symbols $\mathbf{a}_i$ correspond to the coordinates of $s_i(t)$.

If the basis for the first set of signals is obtained, it can be verified that

$$s_0(t) = \sqrt{T}\phi_0(t), \qquad s_1(t) = -\sqrt{T}\phi_0(t),$$

where $\phi_0(t) = s_0(t)/\sqrt{T}$, which is plotted in Figure 3.16.

The basis has a single element, $\phi_0(t)$, for a very simple reason: since $s_1(t) = -s_0(t)$, a change in the sign of the coordinate for $s_0(t)$ is enough to generate $s_1(t)$. The amplitude of $\phi_0(t)$ is taken from the orthonormality condition, normalization of the energy ($\int_{-\infty}^{\infty} \phi_0^2(t)\, dt = 1$). Now it is easy to calculate the distance between the signals and their energy

$$d(s_0(t), s_1(t)) = d(\mathbf{a}_0, \mathbf{a}_1) = \sqrt{(a_{0,0} - a_{1,0})^2} = \sqrt{(\sqrt{T} - (-\sqrt{T}))^2} = 2\sqrt{T}.$$

The energy per symbol is

$$E_s = E[\mathcal{E}\{s(t)\}] = E[\mathcal{E}\{\mathbf{a}_i\}] = p_A(\mathbf{a}_0)a_{0,0}^2 + p_A(\mathbf{a}_1)a_{1,0}^2 = \frac{1}{2}T + \frac{1}{2}T = T.$$

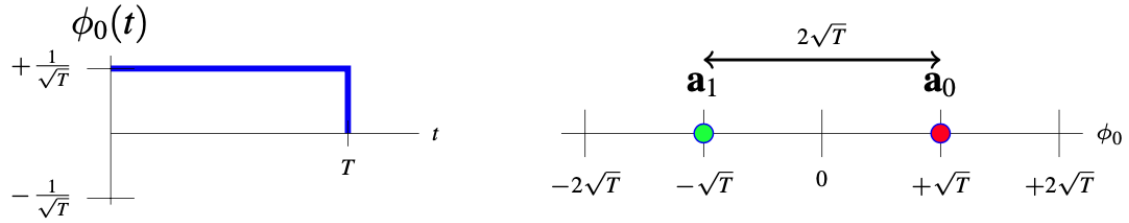


Figure 3.16: Orthogonalization for the first set of signals.

The result, of course, is the same that was previously obtained by calculating on the waveforms.

For the second set of signals, the basis and coordinates are shown in Figure 3.17.

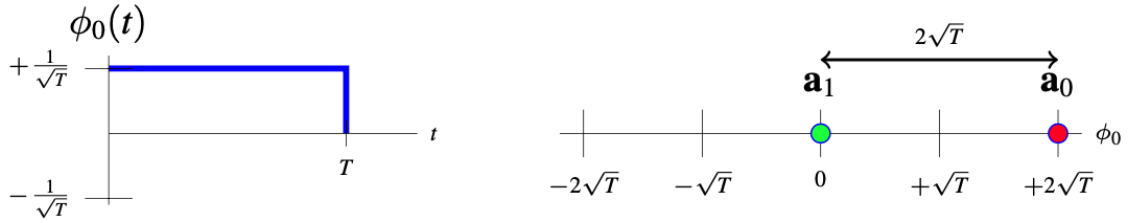


Figure 3.17: Orthogonalization for the second set of signals.

If we compare it with the result for the first set, we can see that the unique element of the basis is the same signal. The distance between signals to be maintained, so that in the face of noise it would behave the same as the first set. However, the average energy of the constellation increases

$$E_s = E[\mathcal{E}\{\mathbf{a}_i\}] = p_A(\mathbf{a}_0)a_{0,0}^2 + p_A(\mathbf{a}_1)a_{1,0}^2 = \frac{1}{2}4T = 2T.$$

The orthogonalization of the third set is shown in Figure 3.18. This constellation is the same one that was obtained for the first set, but now the single element of the basis is different: a sinusoid with a cycle in T seconds, instead of a rectangle of duration T seconds.

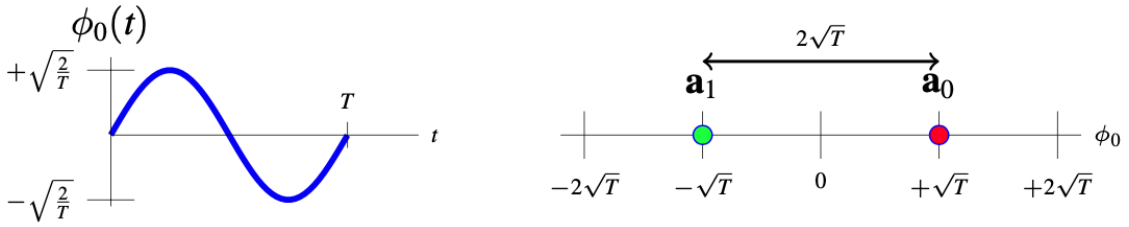


Figure 3.18: Orthogonalization for the third set of signals.

Finally, the orthogonalization of the fourth set is shown in Figure 3.19. Unlike the previous sets, two elements appear in the basis of the vector space. If previously the symbols of the constellation $\mathbf{a}_0$ and $\mathbf{a}_1$ could be interpreted as points on a line, now they are interpreted as points on a plane whose axes represent the coordinates on $\phi_0(t)$ and $\phi_1(t)$. The symbols $\mathbf{a}_0$ and $\mathbf{a}_1$ are now expressed as 2-dimensional vectors whose values are

$$\mathbf{a}_0 = \begin{bmatrix} a_{0,0} \\ a_{0,1} \end{bmatrix} = \begin{bmatrix} \sqrt{2T} \\ 0 \end{bmatrix} \qquad \mathbf{a}_1 = \begin{bmatrix} a_{1,0} \\ a_{1,1} \end{bmatrix} = \begin{bmatrix} 0 \\ \sqrt{2T} \end{bmatrix}.$$

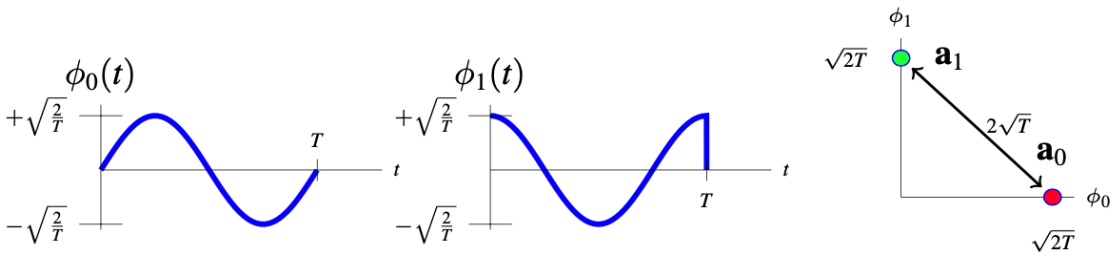


Figure 3.19: Orthogonalization for the fourth set of signals.

Now the distance is calculated as the distance on the plane, that is

$$\begin{aligned} d(s_0(t), s_1(t)) =& d(\mathbf{a}_0, \mathbf{a}_1) = ||\mathbf{a}_0 - \mathbf{a}_1|| \\ =& \sqrt{(a_{0,0} - a_{1,0})^2 + (a_{0,1} - a_{1,1})^2} \\ =& \sqrt{(\sqrt{2T})^2 + (-\sqrt{2T})^2} = 2\sqrt{T}. \end{aligned}$$

And the average energy per symbol is

$$\begin{aligned} E_s =& E[\mathcal{E}\{\mathbf{a}_i\}] = E[||\mathbf{a}_i||^2] \\ =& p_A(\mathbf{a}_0)||\mathbf{a}_0||^2 + p_A(\mathbf{a}_1)||\mathbf{a}_1||^2 \\ =& p_A(\mathbf{a}_0)(a_{0,0}^2 + a_{0,1}^2) + p_A(\mathbf{a}_1)(a_{1,0}^2 + a_{1,1}^2) \\ =& \frac{1}{2}2T + \frac{1}{2}2T = 2T. \end{aligned}$$

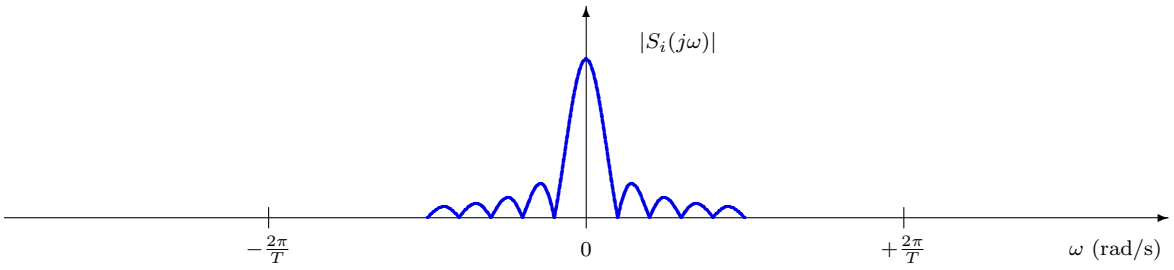
If the first set and the third set are compared, it is observed that the basis is different but the constellation is the same. If the constellation is the same, the distances and the average energy of the constellation, or average energy per symbol, are identical. From this fact we can draw a conclusion: the choice of the basis, $\{\phi_j(t) \mid j = 0, \cdots, N-1\}$, does not affect the performance or the energy. In fact, performance and energy are defined by the constellation.

For this example, the best compromise between performance and energy is obtained with sets 1 and 3, since they share a constellation, so the difference (distance) between signals and the energy of each signal is the same. When will it be more appropriate to use set 1, and when set 3? What does the choice of one or the other orthonormal basis depend on? To answer this question, it is necessary to consider what changes when choosing between one or another orthonormal basis. What is modified is the shape of the signals, in the time domain (seen in Figure 3.15) and in the frequency domain. The frequency response for the four sets, which depends on the Fourier transform of the basis functions, is shown in Figure 3.20.

It can be seen that the signals of sets 1 and 2 have their frequency response centered at 0 Hz, so they are signals appropriate to transmit over channels with "*good response*" at low frequencies. The signals of sets 3 and 4 have their frequency response centered on $\omega = \frac{2\pi}{T}$ rad/s, so they are appropriate signals to transmit over channels with "*good response*" around the frequency $\frac{2\pi}{T}$ radians/s. In other words, the choice of the orthonormal basis depends on the characteristics of the channel.

This conclusion is not specific to these 4 signal sets, but it is rather a general conclusion for the transmitter design. In a situation where the goal is to design the transmitter from scratch, the constellation and the basis are designed almost independently:

- Set 1 and Set 2



- Set 3 and Set 4

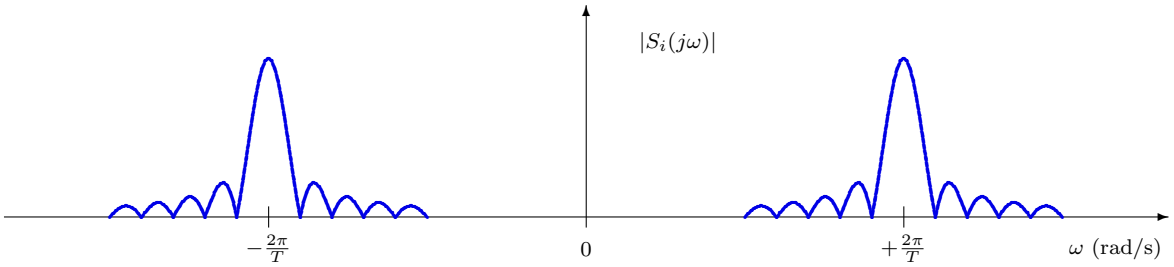


Figure 3.20: Frequency response of the signals of the four sets proposed in the example.

- The constellation according to performance and energy criteria.
- The elements of the basis attending to the behavior of the communications channel (trying that the physical channel behaves like the Gaussian channel).

The geometric interpretation of the signals is derived from the vector space structure of signals, and more specifically from the definition of the inner product in this vector space. With this geometric interpretation, the energy of a particular symbol is the squared norm of the vector $E_i = ||\mathbf{a}_i||^2$.

The validity of everything raised and discussed above is not restricted to this set of signals, but can be applied to the general case of a constellation of M symbols $\{\mathbf{a}_i \mid i = 0, \cdots, M-1\}$ and an orthonormal basis of N signals $\{\phi_j(t) \mid j = 0, \cdots, N-1\}$ that generate the set of signals $\{s_i(t) \mid i = 0, \cdots, M-1\}$. In general:

- The constellation is designed looking for a trade-off between getting a maximum separation between the symbols, and at the same time having an average energy per symbol, or energy of the constellation, within the limits of the system.
  - The choice of the basis is irrelevant in terms of performance and energy.
- The basis is chosen to be suitable for the channel, taking into account its response.
  - The choice of the constellation has no relevance on the adequacy of the signals to the channel.

This allows the design of the constellation to be decoupled from the design of the orthonormal basis that defines the signal space (corresponding to the encoder design and the modulator design, respectively).

Next, each of the 4 functional elements of the communication model illustrated in Figure 3.14 will be analyzed individually. The analysis will start with the receiver and will end with the transmitter.

3.4 Demodulator

The demodulator is the first functional element of the receiver (see Fig. 3.14). This element converts the received signal $r(t)$ into a discrete sequence $\mathbf{q}[n]$ of vectors of dimension N, as illustrated in Figure 3.21.

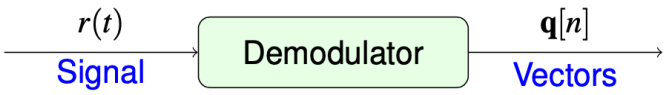


Figure 3.21: Demodulator of a digital communications system.

The sequence $\mathbf{q}[n]$ contains the vector representation of the received signal $r(t)$ into the vector space defined by the orthonormal basis of dimension N

$$\{\phi_0(t), \phi_1(t), \cdots, \phi_{N-1}(t)\}.$$

It is therefore a sequence of N-dimensional vectors

$$\mathbf{q}[n] = \begin{bmatrix} q_0[n] \\ q_1[n] \\ \vdots \\ q_{N-1}[n] \end{bmatrix}$$

Specifically, the received signal is processed by symbol intervals, and at the discrete instant n, the sequence $\mathbf{q}[n]$ contains the vector representation of the piece of the received signal in the symbol interval associated with that instant, that is, in $nT \leq t < (n+1)T$:

$$\mathbf{q}[n] \equiv \text{projection of } r(t) \text{ in } nT \leq t < (n+1)T \text{ through } \{\phi_j(t)\}_{j=0}^{N-1}$$

The projection of a signal over an element of the basis, $\phi_k(t)$, is obtained by using the inner product. Since the signals that form the basis have by definition their support in the interval $0 \leq t < T$, to make the inner product of the fragment of the signal $r(t)$ in the symbol interval of index n, in $nT \leq t < (n+1)T$, is equivalent to the inner product of $r(t)$ with the signal $\phi_k(t)$ delayed nT seconds. Therefore, the k-th coordinate of vector $\mathbf{q}[n]$ can be obtained as

$$q_k[n] = \langle r(t), \phi_k(t - nT) \rangle = \int_{-\infty}^{\infty} r(t)\ \phi_j^*(t - nT)\ dt = \int_{nT}^{(n+1)T} r(t)\ \phi_k^*(t - nT)\ dt.$$

The conjugate operator is only relevant for complex signals. In this subject only real signals will be considered, although for completeness in the notation the possibility of working with complex signals will be included[2]. Therefore, for the implementation of a demodulator, some kind of structure capable of obtaining N inner products is needed. Here we will study two equivalent structures to carry out this task:

[2]There are no complex electromagnetic signals, but in some cases the complex notation is used for the simultaneous handling of two real signals, one contained in the real part and the other in the imaginary part of the analytical signal.

1. Based on correlators.

2. Based on matched filters.

3.4.1 Demodulation by correlation

The idea of the demodulator by correlation is to use the structure that comes from the direct implementation of the inner product operation

$$q_k = \langle r(t), \phi_k(t - nT) \rangle = \int_{nT}^{(n+1)T} r(t)\phi_k^*(t - nT)dt.$$

Therefore, the demodulator has the structure shown in Figure 3.22.

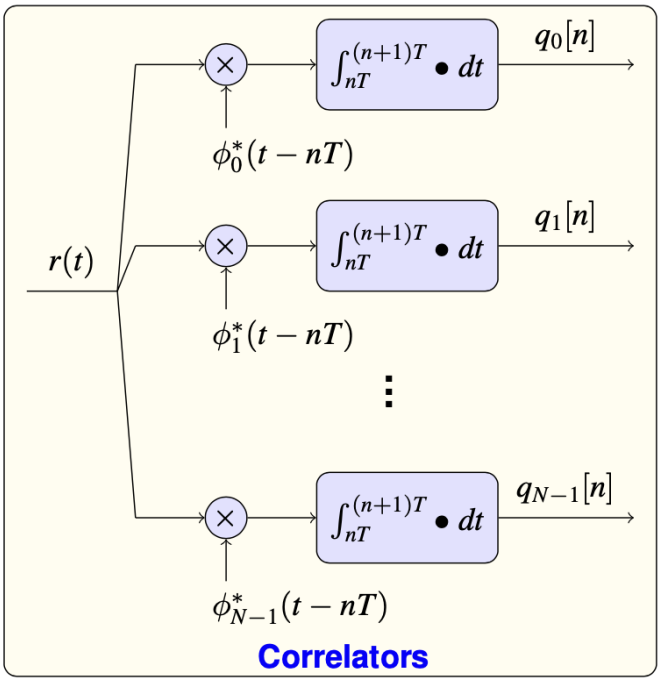


Figure 3.22: Structure of the demodulator by correlation.

The vector notation that we are using may give the impression of an excessively complex demodulator structure for its implementation in hardware. However, when we particularize the structure in practical systems, the structure is generally not very complex. For example, for the four sets of signals in the example of Section 3.3.1, the structure is very simple. For the first two sets, where the basis function takes a constant value over the entire symbol interval, the demodulator reduces to a scaling of the received signal and an integration over the symbol interval, as shown in Figure 3.23 a). For the third set, the demodulator multiplies the received signal by a sinusoidal signal (in hardware, an oscillator) and integrates the product into the symbol interval (Figure 3.23 b)). For the fourth set of signals, the received signal must be multiplied by two sinusoidal components in quadrature (a sine and a cosine, which can be obtained in hardware with an oscillator and a 90 degrees phase shifter) and two integrators (Figure 3.23 c)).

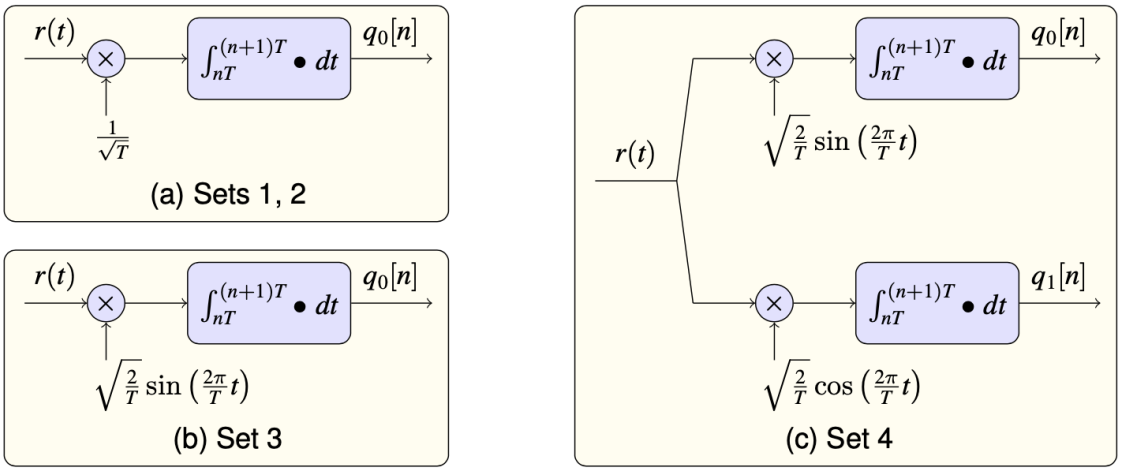


Figure 3.23: Demodulator structure for the four sets of signals analyzed; a) first two sets, b) third set, c) fourth set.

3.4.2 The matched filter

Another possibility for the receiver design is the use of matched filters. In some cases, this option offers a more efficient alternative for the hardware implementation of the demodulator, depending on the shape of the basis functions that define the signal space of the system.

If the signal received at the input of the receiver, $r(t)$, is filtered with a certain filter, with impulse response $h_k(t)$, its output is

$$y_k(t) = r(t) * h_k(t) = \int_{-\infty}^{\infty} r(\tau)\, h_k(t-\tau)\, d\tau.$$

We are going to see that through a filtering process it is possible to implement the inner product operation that the receiver requires. Remember that we are looking for a structure to obtain

$$q_k[n] = \langle r(t), \phi_k(t-nT)\rangle = \int_{nT}^{(n+1)T} r(t)\phi_j^*(t-nT)\, dt. \tag{3.1}$$

If the two previous expressions are compared, for $y_k(t)$ and $q_k[n]$, respectively, it is observed that in the second one $\phi_k^*(t)$ appears instead of $h_k(t)$, a complex conjugate, that in the first one it integrates over τ while in the second one it integrates over t, and that the sign on the integration variable is negative in the response of the filter. If $h_k(t) = \phi^*(-t)$ is chosen, the expression for $y_k(t)$ becomes

$$y_k(t) = \int_{-\infty}^{\infty} r(\tau)\, \phi_k^*(-(t-\tau))\, d\tau.$$

If it is compared with the analytical expression for $q_k[n]$, Eq. (3.1), and taking into account that $-(t-\tau)$ can be written as $(\tau - t)$

$$y_k(t) = \int_{-\infty}^{\infty} r(\tau)\, \phi_k^*(\tau - t)\, d\tau,$$

we see that the two expressions are equal if the value $t = nT$ is taken. So if the signal at the ourput of the filter, $y_k(t)$, is sampled at $t = nT$

$$y_k(nT) = \int_{-\infty}^{\infty} r(\tau)\ \phi_k^*(\tau - nT)\ d\tau = \int_{nT}^{(n+1)T} r(\tau)\ \phi_k^*(\tau - nT)\ d\tau = q_k[n],$$

that is, the inner product of the received signal with $\phi_k(t)$ in the symbol interval of index n.

In general, a filter with impulse response $h(t) = x^*(-t)$ is said to be the *matched filter* to the signal $x(t)$. So the demodulator can be implemented using a bank of N matched filters, matched to the elements of the basis $\phi_k(t)$, with response $h_k(t) = \phi_k^*(-t)$, and sampling at $t = nT$ (at the beginning of the symbol interval), as shown in Figure 3.24.

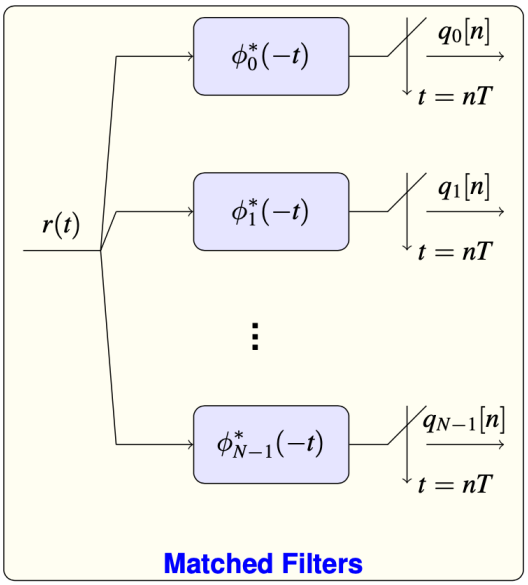


Figure 3.24: Demodulator structure based on matched filters.

For the practical implementation of this structure, it is necessary to note that the matched filters are anticausal: since $\phi_k(t)$ has by definition $0 \leq t < T$ as support, $\phi_i^*(-t)$ has as support $-T < t \geq 0$. In order to have an equivalent causal implementation, the response $\phi^*(-t)$ must be delayed T seconds, and the response is $\phi^*(T-t)$. In this case, the output of the filter is delayed by the same amount and it is necessary to sample at $t = (n+1)T$, at the end of the symbol interval, as shown in Figure 3.25. In this way, we have the same as with the anticausal filter sampling at $t = nT$, at the beginning of the symbol interval.

Analytically, shifting $\phi_k^*(-t)$ T seconds means having $\phi_k^*(-(t-T)) = \phi_k^*(T-t)$. The output of that shifted filter, denoted $y_k^T(t)$ is then

$$y_k^T(t) = r(t) * \phi_k^*(T-t) = \int_{-\infty}^{\infty} r(\tau)\ \phi_k^*(T-(t-\tau))\ d\tau = \int_{-\infty}^{\infty} r(\tau)\ \phi_k^*(T+\tau-t))\ d\tau.$$

Sampling at $t = (n+1)T$ we have

$$y_k^T((n+1)T) = \int_{-\infty}^{\infty} r(\tau)\ \phi_k^*(\tau - nT))\ d\tau = y_k(nT) = q_k[n].$$

This causal structure provides exactly the same result as the one that uses anticausal filters. For the implementation, obviously the structure with causal filters must be used, but since both are analytically equivalent, to make calculations the anticausal matched filters can be used, for ease of notation.

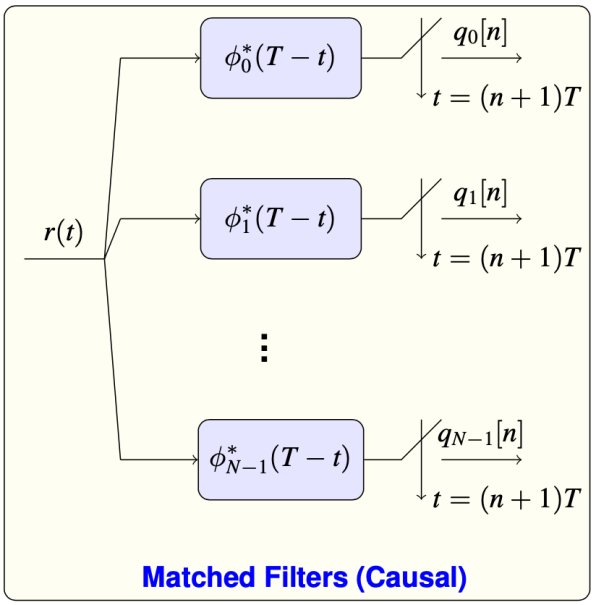


Figure 3.25: Demodulator structure based on causal matched filters.

Properties of the matched filter

The matched filter provides an alternative structure for the implementation of the inner product operations that the demodulator must carry out. Moreover, this structure allows to analyze in a simple way some of the properties of the demodulator.

Maximum signal to noise ratio This is surely the most important property. In order to demonstrate this property, the following general case is analyzed: a known signal $s(t)$, to which noise is added, is filtered, the filter has impulse response $h(t)$, and the output of this filter is evaluated at $t = 0$.. The model for the additive noise is the usual statistical model for thermal noise: stationary, ergodic, white, Gaussian random process, with zero mean and power spectral density, and autocorrelation function, respectively

$$S_n(j\omega) = \frac{N_0}{2}, \; R_n(\tau) = \frac{N_0}{2}\delta(\tau).$$

This general case is represented by the scheme that is shown in Fig. 3.26.

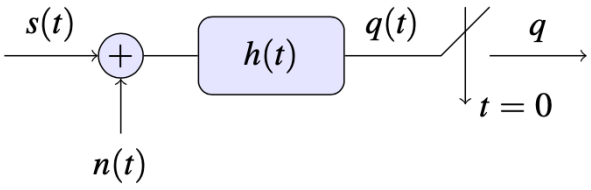


Figure 3.26: Model for the derivation of the maximum signal-to-noise ratio property of the matched filter.

We are going to look for the filter that provides the maximum signal-to-noise ratio at its sampled output, q, defined as

$$\left(\frac{S}{N}\right)_q \equiv \frac{\text{Energy in } q \text{ associated to } s(t)}{\text{Energy in } q \text{ associated to } n(t)}.$$

To simplify the notation, the case of real signals is considered, although the results are trivially extended to the case of complex signals. The output of the filter, $q(t)$, is defined as

$$\begin{aligned} q(t) =&(s(t)+n(t)) * h(t) = s(t) * h(t) + n(t) * h(t) \\ =& \int_{-\infty}^{\infty} s(\tau)\, h(t-\tau)\, d\tau + \int_{-\infty}^{\infty} n(\tau)\, h(t-\tau)\, d\tau. \end{aligned}$$

The value at $t = 0$ is

$$q = q(0) = \int_{-\infty}^{\infty} s(\tau)\, h(-\tau)\, d\tau + \int_{-\infty}^{\infty} n(\tau)\, h(-\tau)\, d\tau = s + n.$$

The signal-to-noise ratio of the output is defined as

$$\left(\frac{S}{N}\right)_q = \frac{E[|s|^2]}{E[|n|^2]} = \frac{|s|^2}{E[n^2]},$$

since the term s is deterministic if the signal $s(t)$ is known

$$E\left[|s|^2\right] = |s|^2 = \left| \int_{-\infty}^{\infty} s(\tau)\, h(-\tau)\, d\tau \right|^2 \text{ (deterministic value)}$$

The value for $E[|n|^2]$ is obtained as

$$\begin{aligned} E\left[|n|^2\right] =& E\left[\left(\int_{-\infty}^{+\infty} n(\tau)\, h(-\tau)\, d\tau\right)\left(\int_{-\infty}^{+\infty} n(\theta)\, h(-\theta)\, d\theta\right)\right] \\ =& \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} \underbrace{E[n(\tau)\, n(\theta)]}_{R_n(\tau-\theta)}\, h(-\tau)\, h(-\theta)\, d\tau\, d\theta \\ =& \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} \underbrace{\frac{N_0}{2}\delta(\tau-\theta)}_{R_n(\tau-\theta)}\, h(-\tau)\, h(-\theta)\, d\tau\, d\theta \\ =& \frac{N_0}{2}\int_{-\infty}^{+\infty} |h(-\tau)|^2\, d\tau = \frac{N_0}{2}\int_{-\infty}^{+\infty} |h(\tau)|^2\, d\tau = \frac{N_0}{2}\,\mathcal{E}\{h(t)\} \end{aligned}$$

In this expressions, the property of the integral of the product of a delta and a function has been applied.

$$\int_{-\infty}^{+\infty} f(x)\, \delta(x-x_0)\, dx = f(x_0).$$

It can be seen that the energy of the noise component only depends on the energy of the filter. Substituting this result in the expression for the signal-to-noise ratio, we have

$$\left(\frac{S}{N}\right)_q = \frac{|s|^2}{E[|n|^2]} = \frac{\left|\int_{-\infty}^{\infty} s(\tau)\, h(-\tau)\, d\tau\right|^2}{\frac{N_0}{2}\mathcal{E}\{h(t)\}}.$$

In order to find the maximum of this signal-to-noise ratio with respect to the impulse response of the filter, $h(t)$, the following reasoning is going to be done: it is assumed that the energy of $h(t)$ is constant, and the maximum of the relation $(S/N)_q$ with respect to $h(t)$ is sought; if that maximum does not depend on $\mathcal{E}\{h(t)\}$ we will have the maximum for any value of $\mathcal{E}\{h(t)\}$ and, therefore, the maximum that are we searching for.

If $\mathcal{E}\{h(t)\}$ is assumed constant, the maximum of the signal-to-noise ratio is limited to computing the maximum of the numerator. If we particularize the Cauchy-Schwarz inequality (see Section 3.2.2) for the signals $s(t)$ and $h(-t)$, and squaring each of the terms of the inequality

$$\left|\int_{-\infty}^{\infty} s(\tau)\, h(-\tau)\, d\tau\right|^2 \leq \left(\int_{-\infty}^{\infty} |s(\tau)|^2\, d\tau\right)\left(\int_{-\infty}^{\infty} |h(-\tau)|^2\, d\tau\right).$$

The equality, and thus the maximum, occurs when $h(-t) = \alpha \times s(t)$ for some value of the constant α. Introducing this result, we get

$$\begin{aligned}\max_{h(t)}\left(\frac{S}{N}\right)_q &= \left.\frac{\left(\int_{-\infty}^{\infty} s(\tau)h(-\tau)d\tau\right)^2}{\frac{N_0}{2}\mathcal{E}\{h(t)\}}\right|_{h(-t)=\alpha s(t)} \\ &= \frac{\left(\int_{-\infty}^{\infty} |s(\tau)|^2 d\tau\right)\left(\alpha^2 \int_{-\infty}^{\infty} |s(\tau)|^2 d\tau\right)}{\frac{N_0}{2}\alpha^2\mathcal{E}\{s(t)\}} \\ &= \frac{2}{N_0}\mathcal{E}\{s(t)\}.\end{aligned}$$

Two conclusions can be drawn from this result:

1. The signal-to-noise ratio becomes maximum when $h(t) = \alpha s(-t)$ for any value of α (except $\alpha = 0$) and, particularly, for the matched filter

   $$h(t) = s(-t).$$

   For complex signals, the same conclusion can be trivially reached for

   $$h(t) = s^*(-t).$$

2. The signal-to-noise ratio at the output of the matched filter does not depend on the specific shape of $s(t)$, but only on its energy and on the power spectral density of the additive noise term.

This proof for a generic signal can be applied to the set of M signals, $\{s_i(t) \mid i = 0, \cdots, M-1\}$, of a digital transmitter, and to the basis for these signals, $\{\phi_i(t) \mid i = 0, \cdots, M-1\}$, which shows that the matched filter-based demodulator structure (or its correlation equivalent) is the structure that allows to obtain the maximum signal-to-noise ratio in each component of the vector $\mathbf{q}[n]$.

3.4.3 Statistical characterization of the demodulator output in the case of transmission on a Gaussian channel

Once the structures that can be used for the implementation of the demodulator have been seen, then the statistical characterization of the sequence of observations $\mathbf{q}[n]$ will be carried out. In particular, the case of signal transmission over a Gaussian channel will be considered: The transmitted signal does not suffer any linear distortion (which implies assuming that the orthonormal basis is perfectly adapted to the channel response), and the only distortion is due to the additive noise. Therefore, the received signal is

$$r(t) = s(t) + n(t).$$

For ease of notation, we consider the first discrete instant, $n = 0$, and the dependency on the discrete time index n will be omitted. The objective is to obtain the statistical model of the observation $\mathbf{q}$ when the symbol that has been transmitted is known, which in general is assumed to be the i-th symbol, i.e. $\mathbf{A} \equiv \mathbf{A}[0] = \mathbf{a}_i$. This means that in the symbol interval of interest, the first one, $0 \leq t < T$, the transmitted signal is

$$s(t) = s_i(t) = \sum_{j=0}^{N-1} a_{i,j} \times \phi_j(t).$$

Now, the demodulator output under this situation is analyzed. Introducing the analytical expression of $r(t)$ to obtain the k-th coordinate, q_k, we have

$$\begin{aligned} q_k =& \langle r(t), \phi_k(t) \rangle = \int_0^T r(t)\ \phi_k^*(t)\ dt = \int_0^T (s(t) + n(t))\ \phi_k^*(t)\ dt \\ =& \int_0^T \left(\sum_{j=0}^{N-1} a_{i,j}\ \phi_j(t) \right) \phi_k^*(t)\ dt + \underbrace{\int_0^T n(t)\ \phi_k^*(t)\ dt}_{z_k} \\ =& \sum_{j=0}^{N-1} a_{i,j} \int_0^T \phi_j(t)\ \phi_k^*(t)\ dt + z_k = \sum_{j=0}^{N-1} a_{i,j}\ \delta[j-k] + z_k = a_{i,k} + z_k. \end{aligned}$$

It has been taken into account in the development that the elements of the basis are orthonormal, which analytically implies that

$$\langle \phi_j(t), \phi_k(t) \rangle = \int_0^T \phi_j(t)\ \phi_k^*(t)\ dt = \begin{cases} 0, & \text{if } k \neq j \\ 1, & \text{if } k = j \end{cases} \equiv \delta[j-k].$$

In this case $a_{i,k}$ is k-th coordinate of the symbol $\mathbf{a}_i$ and z_k is the contribution of the Gaussian noise to that coordinate. All the components (coordinates) of $\mathbf{q}$ can be expressed together using vector notation as

$$\mathbf{q} = \begin{bmatrix} a_{i,0} \\ a_{i,1} \\ \vdots \\ a_{i,N-1} \end{bmatrix} + \begin{bmatrix} z_0 \\ z_1 \\ \vdots \\ z_{N-1} \end{bmatrix} = \mathbf{a}_i + \mathbf{z}.$$

This expression can be generalized for any discrete time instant as

$$\mathbf{q}[n] = \mathbf{A}[n] + \mathbf{z}[n].$$

The vector component, $\mathbf{a}_i$, is deterministic (the assumption is that the transmitted symbol is known).The components of the noise vector are considered as a random variables. Since $n(t)$ is Gaussian, each component is Gaussian. The vector with the noise coordinates is therefore composed of N jointly Gaussian random variables. To characterize them, it is necessary to find their joint probability density function, which in the case of jointly Gaussian variables is defined by its vector of means and its covariance matrix. The mean of the k-th component is

$$m_{z_k} = E[z_k] = E\left[\int_0^T n(t)\ \phi_j^*(t)\ dt \right] = \int_0^T E[n(t)]\ \phi_j^*(t)\ dt = 0,$$

since $n(t)$ has zero mean. Next, the covariance matrix is calculated. Taking into account that all the components have zero mean, the covariance between two of the components is

$$\begin{aligned}\text{Cov}(z_j, z_k) =& E[z_j\ z_k^*] = E\left[\left(\int_0^T n(t)\ \phi_j^*(t)\ dt\right)\left(\int_0^T n^*(\tau)\ \phi_k(\tau)\ d\tau\right)\right] \\ =& \int_0^T \int_0^T \underbrace{E[n(t)\ n^*(\tau)]}_{R_n(t-\tau)=\frac{N_0}{2}\ \delta(t-\tau)}\ \phi_j^*(t)\ \phi_k(\tau)\ dt\ d\tau \\ =& \int_0^T \int_0^T \frac{N_0}{2}\ \delta(t-\tau)\ \phi_j^*(t)\ \phi_k(\tau)\ dt\ d\tau \\ =& \frac{N_0}{2}\int_0^T \phi_j^*(t)\ \phi_k(t)\ dt = \frac{N_0}{2}\ \delta[j-k].\end{aligned}$$

Therefore, the covariance between two different coordinates is zero, and the covariance of a coordinate with itself (variance) is

$$\sigma_{z_k}^2 = \frac{N_0}{2}.$$

This means that the random variables that model each of the elements of the noise vector are uncorrelated, which for Gaussian random variables is equivalent to saying that they are independent. Thus, the covariance matrix is a diagonal matrix with the variance of each component on the main diagonal. In light of these results, it can be concluded that the N noise components are uncorrelated (independent) Gaussian random variables with zero mean and variance $N_0/2$. Since a Gaussian random variable is uniquely determined from its mean and its variance, the probability density function of each component z_k is

$$f_{z_k}(z_k) = \mathcal{N}\left(0, \frac{N_0}{2}\right) = \frac{1}{\sqrt{\pi N_0}} e^{-\frac{z_k^2}{N_0}},$$

and as under Gaussian statistics uncorrelation implies independence, the joint probability density function of $\mathbf{z}$ is obtained by the product of the distributions of each of the components of the vector

$$f_{\mathbf{z}}(\mathbf{z}) = \prod_{k=0}^{N-1} f_{z_k}(z_k) = \frac{1}{(\pi N_0)^{N/2}} e^{-\sum_{k=0}^{N-1}\frac{z_k^2}{N_0}} = \frac{1}{(\pi N_0)^{N/2}} e^{-\frac{||\mathbf{z}||^2}{N_0}},$$

or equivalently

$$f_{\mathbf{z}}(\mathbf{z}) = \mathcal{N}^N\left(\mathbf{0}, \frac{N_0}{2}\right).$$

Once the PDF of the noise vector is obtained, it is easy to obtain the conditional distribution of the observation $\mathbf{q}$. The probability density function of $\mathbf{q}$ when the transmitted symbol is known, $\mathbf{A} = \mathbf{a}_i$, is a conditional distribution: distribution of q_k given that the transmitted symbol is $\mathbf{A} = \mathbf{a}_i$. For each of the components of the vector $\mathbf{q}$, q_k, there is also a conditional distribution: distribution of q_k given than the transmitted symbol is $\mathbf{A} = \mathbf{a}_i$. In this case, $q_k = a_{i,k} + z_k$ is a random variable formed by the sum of a deterministic constant $a_{i,k}$ and a random variable, z_k. Therefore, it has the same type of distribution as z_k but with the mean modified by adding the constant value, $a_{i,k}$, to the mean of z_k, which in this case is zero. Thus, the distribution is a Gaussian distribution, with mean equal to the k-th coordinate of the transmitted symbol, $a_{i,k}$, and variance $N_0/2$

$$f_{q_k|\mathbf{A}}(q_k|\mathbf{a}_i) = \mathcal{N}\left(a_{i,k}, \frac{N_0}{2}\right) = \frac{1}{\sqrt{\pi N_0}} e^{-\frac{(q_k - a_{i,k})^2}{N_0}}.$$

Because of the conditional independence, the joint conditional probability density function of the vector $\mathbf{q}$ given $\mathbf{A} = \mathbf{a}_i$ is obtained by the product of the conditional distributions of each of the vector components

$$f_{\mathbf{q}|\mathbf{A}}(\mathbf{q}|\mathbf{a}_i) = \prod_{k=0}^{N-1} f_{q_k|A_k}(q_k|a_{i,k}) = \frac{1}{(\pi N_0)^{N/2}} e^{-\sum_{k=0}^{N-1} \frac{(q_k - a_{i,k})^2}{N_0}}$$
$$= \frac{1}{(\pi N_0)^{N/2}} e^{-\frac{||\mathbf{q}-\mathbf{a}_i||^2}{N_0}}.$$

The conditional distribution of the observation vector, given a transmitted symbol, is an N dimensional Gaussian distribution, with independent components, with the mean equal to the transmitted symbol and variance $N_0/2$ in each of the N directions of the space

$$f_{\mathbf{q}|\mathbf{A}}(\mathbf{q}|\mathbf{a}_i) = \mathcal{N}^N\left(\mathbf{a}_i, \frac{N_0}{2}\right).$$

In the analysis of the demodulator we have started from the idea of recovering the coordinates of the received signal over the basis of the signal space of the transmitter. However, we have not analyzed if this observation contains all the relevant information to make the decision about the symbol that was transmitted. The answer to this question is that $\mathbf{q}$ contains all the relevant information to decide which symbol was transmitted. Strictly speaking, the vector $\mathbf{q}$ is said to be a *sufficient statistic for detection*. The demo can be found, for example at [Artés-Rodríguez et al., 2007].

3.4.4 Equivalent discrete channel

Some of the conclusions obtained from the analysis of the digital communication model are the following:

1. The reliability of a communication scheme, regarding the modulation process, is given by the characteristics of the signal constellation and not by the elements of the basis that is used in the modulation process.
2. The signal-to-noise ratio does not depend on the paritcular form of the elements of the bass, but on the energy of the signals and the power spectral density of the noise (and equivalently, on the noise power).
3. The output of the demodulator is a vector $\mathbf{q}$ that takes the form $\mathbf{q} = \mathbf{A} + \mathbf{z}$, where $\mathbf{A}$ is the transmitted symbol and $\mathbf{z}$ is a noise component with a jointly Gaussian probability density function.

These conclusions allow us to obtain a simplification of the general model of a communication system. The modulator, channel and demodulator can be grouped into a single element that is called the *equivalent discrete channel*. This model, which is represented in Fig. 3.27, is useful in analysis of the detecto, because it allows us to "hide" the analog nature of the channel and focus on the aspects that influence the reliability of the communication.

The equivalent discrete channel is in general vector, with the dimension given by the signal space of the system, N. The input-output relationship, for a Gaussian channel transmission model, is given by

$$\mathbf{q}[n] = \mathbf{A}[n] + \mathbf{z}[n],$$

as represented in Figure 3.28

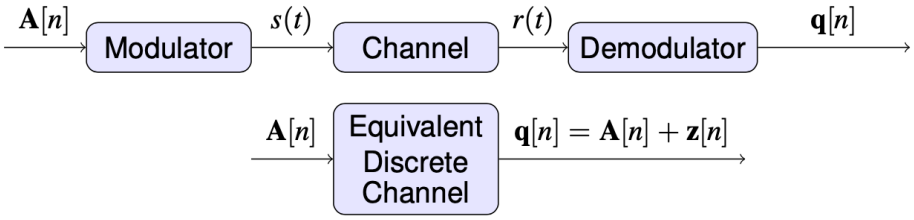


Figure 3.27: Definition of equivalent discrete channel.

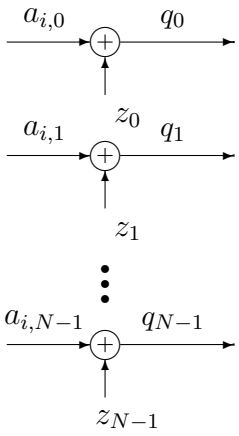


Figure 3.28: Relationships between transmitter and receiver vector representations given by the equivalent discrete channel, when the symbol $\mathbf{A}[n] = \mathbf{a}_i$ is transmitted over a Gaussian channel.

3.5 Detector

The detector is the final element of a digital demodulator, see Fig. 3.14, and its function is to provide the estimate of the symbol that has been transmitted at an instant n, $\hat{B}[n]$, based on the vector representation of the signal received in the symbol interval associated with this instant, $\mathbf{q}[n]$, as illustrated in Figure 3.29. With this estimate, there is an implicit estimate of the transmitted bits, $\hat{B}_b[\ell]$, due to the identification of each symbol, b_i, with an specific m-bit tuple.

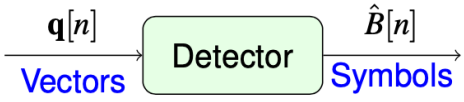


Figure 3.29: Detector in a digital communications system.

Given its function, it is a determining element for system performance, parameterized by the probability of error, defined at the symbol level or at the bit level. Therefore, the detector is designed according to the following criterion: to minimize the symbol error rate.

3.5.1 Detector Design - Decision Regions

Since $\mathbf{q}[n]$ is a sufficient statistic for the detection of $\mathbf{A}[n]$ (and thus $B[n]$), the decision will be made symbol-by-symbol, without memory. For each time instant n, $\hat{B}[n]$ is decided from $\mathbf{q}[n]$, always applying the same rule, since the statistics of $\mathbf{q}[n]$ are time invariant.

Given that the alphabet of symbols has M possible values

$$B[n] \in \{b_0, b_1, \cdots, b_{M-1}\},$$

the design of the detector consists in establishing for each possible value of $\mathbf{q}[n]$, which value of the alphabet, out of the possible M, is decided. In another way, establish what are the possible values of $\mathbf{q}[n]$ that will lead to the decision of each of the M possible values of $B[n]$. The way to set those values is by defining what are called *decision regions*. The domain of $\mathbf{q}[n]$ will be splitted into M disjoint regions that form a partition of the space of $\mathbf{q}[n]$

$$\{I_0, I_1, \cdots, I_{M-1}\}.$$

Each region is associated with one symbol, identifying the association through the subindex (I_k is associated with b_k). Once the regions are defined, in view of an observation $\mathbf{q}[n]$, the decision is $\hat{B}[n] = b_k$ when the observation is in the decision region of b_k, i.e., when $\mathbf{q}[n] \in I_k$

$$\text{Given } \mathbf{q}[n] = \mathbf{q}_0,\ \hat{B}[n] = b_k \text{ if } \mathbf{q}_0 \in I_k.$$

That is why they are called decision regions.

The design of the detector is stated as the problem of establishing the M decision regions that minimize the symbol error rate.

3.5.2 Obtaining the optimal detector

The rules or criteria to optimally establish the decision regions must minimize the symbol error rate. For ease of notation, the dependence on the time index is ignored

$$\hat{B} \equiv \hat{B}[n],\ \mathbf{q} \equiv \mathbf{q}[n].$$

When $\mathbf{q}$ takes a given value $\mathbf{q}_0$, the detector makes a decision about the transmitted symbol, for example $\hat{B} = b_i$. The error probability associated to that decision is denoted as $P_e(\mathbf{q} = \mathbf{q}_0 \to \hat{B} = b_i)$, and it is

$$P_e(\mathbf{q} = \mathbf{q}_0 \to \hat{B} = b_i) = P(B \neq b_i | \mathbf{q} = \mathbf{q}_0) = 1 - P(B = b_i | \mathbf{q} = \mathbf{q}_0) = 1 - p_{B|\mathbf{q}}(b_i|\mathbf{q}_0).$$

The error probability of such a decision, $\hat{B} = b_i$, from an observation, $\mathbf{q} = \mathbf{q}_0$, is equal to the probability of a different transmitted symbol, $B[n] \neq b_i$, given this observation. The conditional distribution $p_{B|\mathbf{q}}(b_i|\mathbf{q}_0)$ is called the *posterior probability* of the symbol b_i.

If the detector always made the same decision, $\hat{B} = b_i$, regardless of the observation value $\mathbf{q}$, the mean error probability, which will be denoted as $P_e(\hat{B} = b_i, \forall \mathbf{q})$, is obtained by calculating the average over the set of all possible values of the observation $\mathbf{q}$, which is statistically the mathematical expectation under the distribution of $\mathbf{q}$

$$\begin{aligned} P_e(\hat{B} = b_i, \forall \mathbf{q}) =& E_{f_{\mathbf{q}}(\mathbf{q}_0)}\left[P_e(\mathbf{q} = \mathbf{q}_0 \to \hat{B} = b_i)\right] = \int_{-\infty}^{\infty} \left[1 - p_{B|\mathbf{q}}(b_i|\mathbf{q}_0)\right]\ f_{\mathbf{q}}(\mathbf{q}_0)\ d\mathbf{q}_0 \\ =& \int_{-\infty}^{\infty} f_{\mathbf{q}}(\mathbf{q}_0)\ d\mathbf{q}_0 - \int_{-\infty}^{\infty} p_{B|\mathbf{q}}(b_i|\mathbf{q}_0)\ f_{\mathbf{q}}(\mathbf{q}_0)\ d\mathbf{q}_0 \\ =& 1 - \int_{-\infty}^{\infty} p_{B|\mathbf{q}}(b_i|\mathbf{q}_0)\ f_{\mathbf{q}}(\mathbf{q}_0)\ d\mathbf{q}_0. \end{aligned}$$

When the detector makes different decisions when the observation $\mathbf{q}$ takes values in each one of the M decision regions, the error probability of the system, taking into account that these are disjoint regions that form a partition of the domain of $\mathbf{q}$, is

$$P_e = 1 - \sum_{i=0}^{M-1} \int_{I_i} p_{B|\mathbf{q}}(b_i|\mathbf{q}_0)\ f_{\mathbf{q}}(\mathbf{q}_0)\ d\mathbf{q}_0.$$

To find the minimum of this expression, the following must be taken into account:

- The minimum is obtained when the second term is maximized.
- The function within the integrals of this term, $p_{B|\mathbf{q}}(b_i|\mathbf{q}_0)f_{\mathbf{q}}(\mathbf{q}_0)$, is always greater than or equal to zero because it is the product of non-negative functions ($0 \leq p_{B|\mathbf{q}}(b_i|\mathbf{q}_0) \leq 1$, $0 \leq f_{\mathbf{q}}(\mathbf{q}_0) \leq 1$). This implies that the error probability is minimized when this argument is maximized.
- $f_{\mathbf{q}}(\mathbf{q}_0)$ is independent of the decision. Therefore, the maximum of the summation is obtained when the value of $p_{B|\mathbf{q}}(b_i|\mathbf{q}_0)$ is maximized in each of the decision regions.
- Consequently, the decision region I_i, for which the decision is $\hat{B} = b_i$, is the one satisfying:

$$p_{B|\mathbf{q}}(b_i|\mathbf{q}_0) > p_{B|\mathbf{q}}(b_j|\mathbf{q}_0), \quad \forall\ j \neq i.$$

In other words, given an observation $\mathbf{q} = \mathbf{q}_0$, the detector must calculate the set of posterior probabilities $\{p_{B|\mathbf{q}}(b_j|\mathbf{q}_0) \mid j = 0, \cdots, M-1\}$ and the decision is the symbol b_i that maximizes the posterior probability. In the case in which two different symbols b_i, b_k, obtain the maximum value, that is to say

$$p_{B|\mathbf{q}}(b_i|\mathbf{q}_0) = p_{B|\mathbf{q}}(b_k|\mathbf{q}_0) > p_{B|\mathbf{q}}(b_j|\mathbf{q}_0), \quad \forall\ j \neq i, k,$$

the decision can be any of them, arbitrarily, and the choice does not affect the error probability.

This criterion is known as *maximum a posteriori* or MAP (from "*Maximum A Posteriori*"). Its name comes from the denomination of $p_{B|\mathbf{q}}(b_i|\mathbf{q}_0)$ as *posterior* probabilities of B given $\mathbf{q}$, since it represents the probability of the symbols once the transmission has been made, instead of the probabilities *a priori* $p_B(b_i)$. Figure 3.30 shows an example of how these posterior probabilities could be for a constellation of 4 symbols in a one-dimensional space. It can be verified that for each possible observation value q, the sum of the four posterior probabilities is one.

In order to obtain the decision regions, it is necessary to look for the symbol having the maximum poterior probability for each value of q. The maximum posterior probability value, for each possible observation value, is highlighted in black in Figure 3.31. The decision region of a symbol is formed by the set of observation values for which the posterior probability of that symbol is the largest one. The decision region of b_0 (or of $\mathbf{a}_0$), is the set of points for which $p_{B|\mathbf{q}}(b_0|q)$ is greater than the posterior probabilities of the other 3 symbols; the decision region of b_1 (or of $\mathbf{a}_1$), is the set of points for which $p_{B|\mathbf{q}}(b_1|q)$ is greater than the posterior probabilities of the other 3 symbols, and so on. The figure illustrates this way of obtaining the decision regions.

The analytical expression of the posterior probabilities can be obtained by means of Bayes' rule

$$p_{B|\mathbf{q}}(b_j|\mathbf{q}_0) = \frac{p_B(b_j)\ f_{\mathbf{q}|B}(\mathbf{q}_0|b_j)}{f_{\mathbf{q}}(\mathbf{q}_0)}.$$

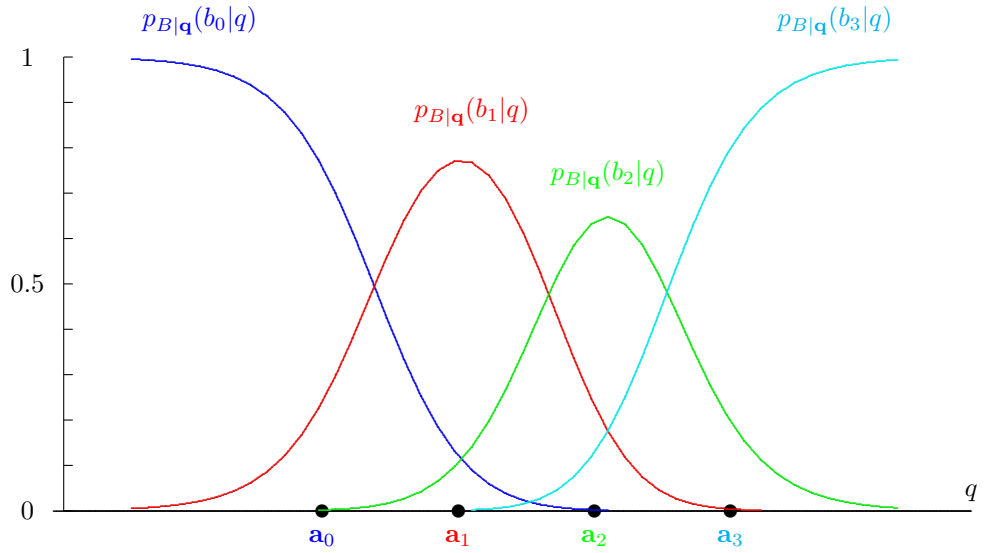


Figure 3.30: An example of the posterior probabilities for a 4-symbol constellation in one-dimensional space.

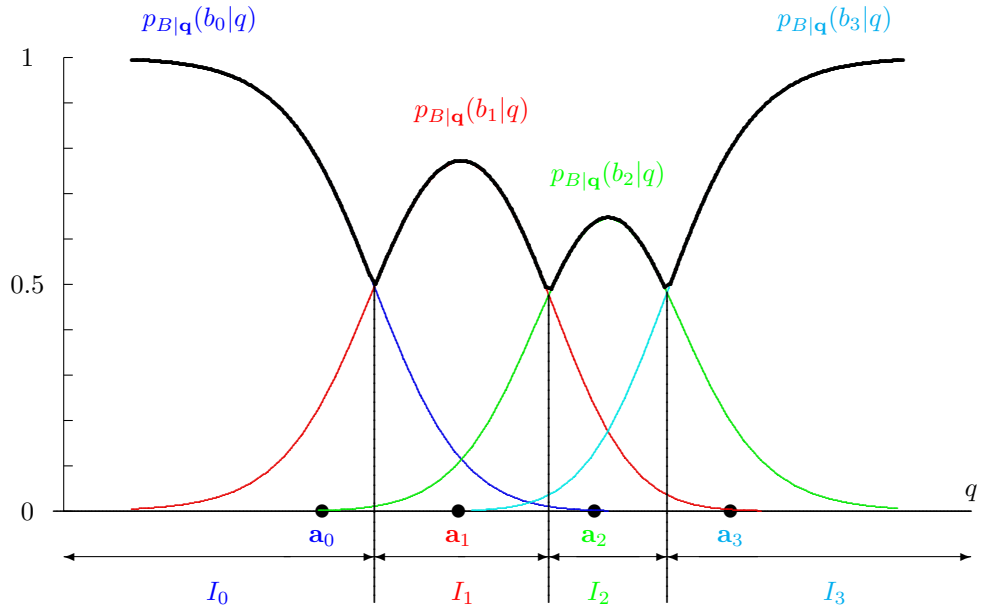


Figure 3.31: Obtaining the decision regions using the maximum a posteriori criterion for an example: a 4-ary constellation in a one-dimensional space.

Considering that $B = b_j$ implies that $\mathbf{A} = \mathbf{a}_j$ and vice versa,

$$f_{\mathbf{q}|B}(\mathbf{q}_0|b_j) = f_{\mathbf{q}|\mathbf{A}}(\mathbf{q}_0|\mathbf{a}_j).$$

Replacing this distribution into the expression for the posterior probabilities, the MAP criterion is reduced to finding the symbol b_i that satisfies

$$\frac{p_B(b_i)\ f_{\mathbf{q}|\mathbf{A}}(\mathbf{q}_0|\mathbf{a}_i)}{f_{\mathbf{q}}(\mathbf{q}_0)} > \frac{p_B(b_j)\ f_{\mathbf{q}|\mathbf{A}}(\mathbf{q}_0|\mathbf{a}_j)}{f_{\mathbf{q}}(\mathbf{q}_0)} \qquad j = 0, \cdots, M-1,\ j \neq i.$$

Since $f_{\mathbf{q}}(\mathbf{q}_0)$ is a non-negative quantity independent of the decision, this condition is equivalent to

$$p_B(b_i)\ f_{\mathbf{q}|\mathbf{A}}(\mathbf{q}_0|\mathbf{a}_i) > p_B(b_j)\ f_{\mathbf{q}|\mathbf{A}}(\mathbf{q}_0|\mathbf{a}_j) \qquad j = 0, \cdots, M-1,\ j \neq i,$$

or, equivalently

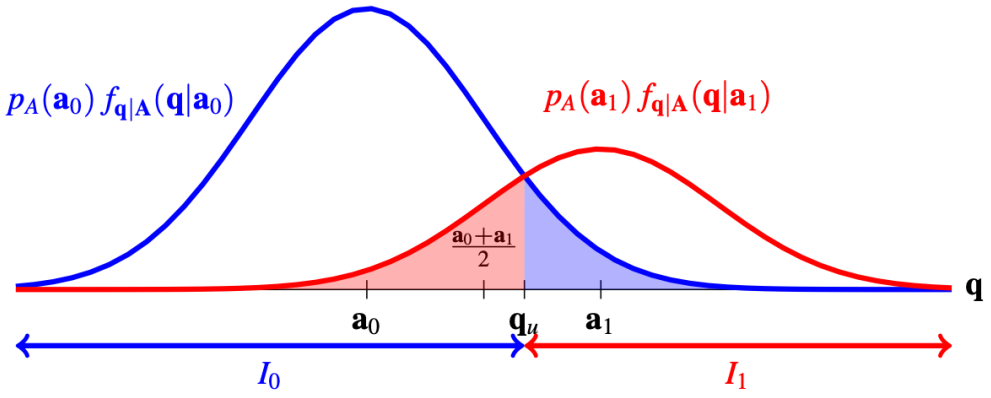
$$p_{\mathbf{A}}(\mathbf{a}_i)\ f_{\mathbf{q}|\mathbf{A}}(\mathbf{q}_0|\mathbf{a}_i) > p_{\mathbf{A}}(\mathbf{a}_j)\ f_{\mathbf{q}|\mathbf{A}}(\mathbf{q}_0|\mathbf{a}_j) \qquad j = 0, \cdots, M-1,\ j \neq i.$$

Below is an example that illustrates the application of this criterion under the conditions analyzed in Section 3.4.3, i.e. transmission over a Gaussian channel. Recall that in that case the conditional distributions of the observation were Gaussian, of the dimension of the signal space N, mean equal to the transmitted symbol and variance $N_0/2$ in all directions of the space

$$f_{\mathbf{q}|\mathbf{A}}(\mathbf{q}|\mathbf{a}_i) = \mathcal{N}^N\left(\mathbf{a}_i, \frac{N_0}{2}\right).$$

Example

A binary system ($M = 2$) employing a one-dimensional signal space ($N = 1$) is considered. In this case both $\mathbf{A}$ and $\mathbf{q}$ are scalars, and the symbol probabilities are $p_B(b_1) = \frac{2}{3}$ and $p_B(b_0) = \frac{1}{3}$. That is, the symbol b_1 is twice as likely. Figure 3.32 plots $p_B(b_1)\ f_{\mathbf{q}|\mathbf{A}}(\mathbf{q}|\mathbf{a}_1)$ y $p_B(b_0)\ f_{\mathbf{q}|\mathbf{A}}(\mathbf{q}|\mathbf{a}_0)$, when the distribution of $\mathbf{a}$ conditional on each symbol is Gaussian with the mean of the transmitted symbol and variance $N_0/2$.



$$\mathbf{q}_u = \frac{\mathbf{a}_0 + \mathbf{a}_1}{2} + \frac{N_0}{2}\frac{\ln\frac{p_A(\mathbf{a}_0)}{p_A(\mathbf{a}_1)}}{(\mathbf{a}_1 - \mathbf{a}_0)},\ I_0 = (-\infty, \mathbf{q}_u),\ I_1 = [\mathbf{q}_u, \infty)$$

Figure 3.32: Application of the MAP criterion with non-equiprobable symbols.

The decision regions are given by the threshold q_u, the point where the curves intersect

$$fp_A(\mathbf{a}_0) f_{\mathbf{q}|\mathbf{A}}(\mathbf{q}_u|\mathbf{a}_0) = fp_A(\mathbf{a}_1) f_{\mathbf{q}|\mathbf{A}}(\mathbf{q}_u|\mathbf{a}_1),$$

and therefore q_u is

$$\mathbf{q}_u = \frac{\mathbf{a}_0 + \mathbf{a}_1}{2} + \frac{N_0}{2} \frac{\ln \frac{p_A(\mathbf{a}_0)}{p_A(\mathbf{a}_1)}}{(\mathbf{a}_1 - \mathbf{a}_0)}.$$

If the value $\mathbf{q} = \mathbf{q}_0$ in the receiver is greater than $\mathbf{q}_u$, then

$$p_B(b_0)\ f_{\mathbf{q}|\mathbf{A}}(\mathbf{q}_0|\mathbf{a}_0) > p_B(b_1)\ f_{\mathbf{q}|\mathbf{A}}(\mathbf{q}_0|\mathbf{a}_1),$$

so such points will be part of the decision region I_0, while if it is less than $\mathbf{q}_u$

$$p_B(b_1)\ f_{\mathbf{q}|\mathbf{A}}(\mathbf{q}_0|\mathbf{a}_1) > p_B(b_0)\ f_{\mathbf{q}|\mathbf{A}}(\mathbf{q}_0|\mathbf{a}_0),$$

so such points will be part of the decision region I_1. Therefore, the decision regions are

$$I_0 = (\mathbf{q}_u, \infty)$$

$$I_1 = (-\infty, \mathbf{q}_u)$$

The areas that are highligted in the figure (in blue and red) define the symbol error rate

$$P_e = p_A(\mathbf{a}_0) \int_{I_1} f_{\mathbf{q}|\mathbf{A}}(\mathbf{q}|\mathbf{a}_0)\ dq + p_A(\mathbf{a}_1) \int_{I_0} f_{\mathbf{q}|\mathbf{A}}(\mathbf{q}|\mathbf{a}_1)\ dq,$$

as it will be shown later (see Section 3.5.3). Taking this into account, it is possible to see that these are the optimal decision regions, because if the decision threshold is modified, this error probability increases, as shown in Figure 3.33.

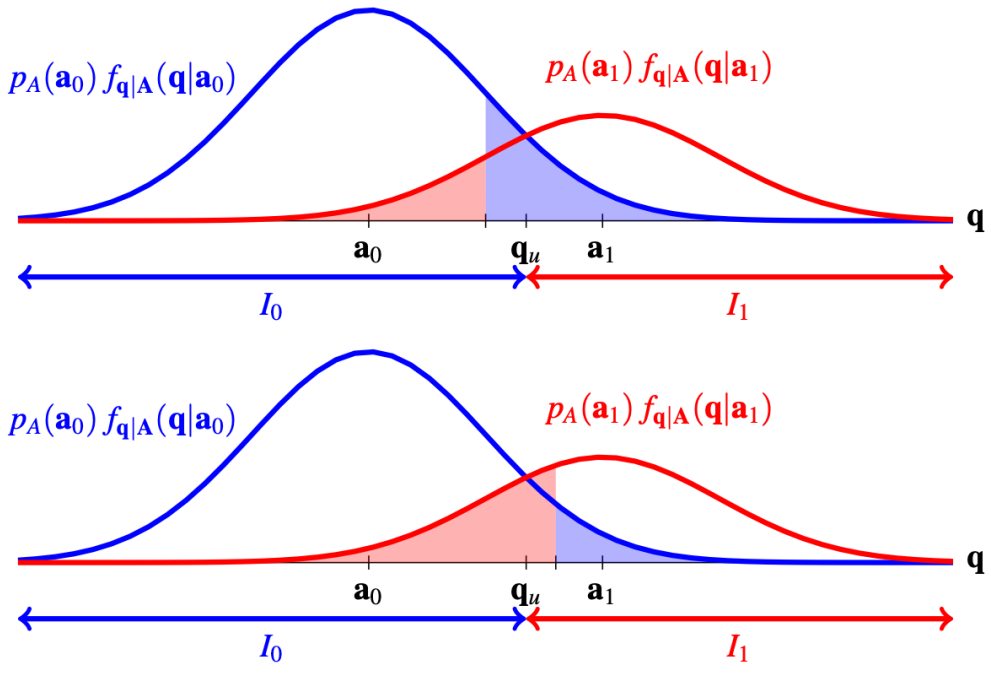


Figure 3.33: MAP criterion as the optimal decision with non-equiprobable symbols.

In many systems, all symbols are transmitted with the same probability. Under the equiprobable symbols hypothesis, $p_B(b_j) = 1/M$, the *a priori* probabilities of the symbols become irrelevant in the comparison, so the decision rule for an observation value $\mathbf{q} = \mathbf{q}_0$ reduces to finding the symbol b_i that satisfies

$$f_{\mathbf{q}|\mathbf{A}}(\mathbf{q}_0|\mathbf{a}_i) > f_{\mathbf{q}|\mathbf{A}}(\mathbf{q}_0|\mathbf{a}_j) \qquad j = 0, \cdots, M-1,\ j \neq i,$$

that is, the symbol that maximizes the conditional distribution of the observation, $f_{\mathbf{q}|\mathbf{A}}(\mathbf{q}_0|\mathbf{a}_i)$. This particularization of the MAP criterion for equiprobable symbols is called the *maximum likelihood criterion* or ML criterion (from "*Maximum Likelihood*"). The name comes from the function $f_{\mathbf{q}|\mathbf{A}}(\mathbf{q}_0|\mathbf{a}_i)$ itself, called the likelihood function since it represents a measure of certainty or likelihood that the true hypothesis is that the transmitted symbol was $\mathbf{a}_i$ given that $\mathbf{q} = \mathbf{q}_0$.

Example

The same binary system ($M = 2$) is considered in a one-dimensional signal space ($N = 1$) of the previous example, but in this case with equiprobable symbols. The figure 3.34 shows the curves of $p_B(b_1)\ f_{\mathbf{q}|\mathbf{A}}(\mathbf{q}|\mathbf{a}_1)$ and $p_B(b_0)\ f_{\mathbf{q}|\mathbf{A}}(\mathbf{q}|\mathbf{a}_0)$ respectively. Now, eliminating the a priori probabilities of each symbol in the comparison does not modify the representation by more than one scale factor, since the symbols are equiprobable $p_B(b_0) = p_B(b_1)$.

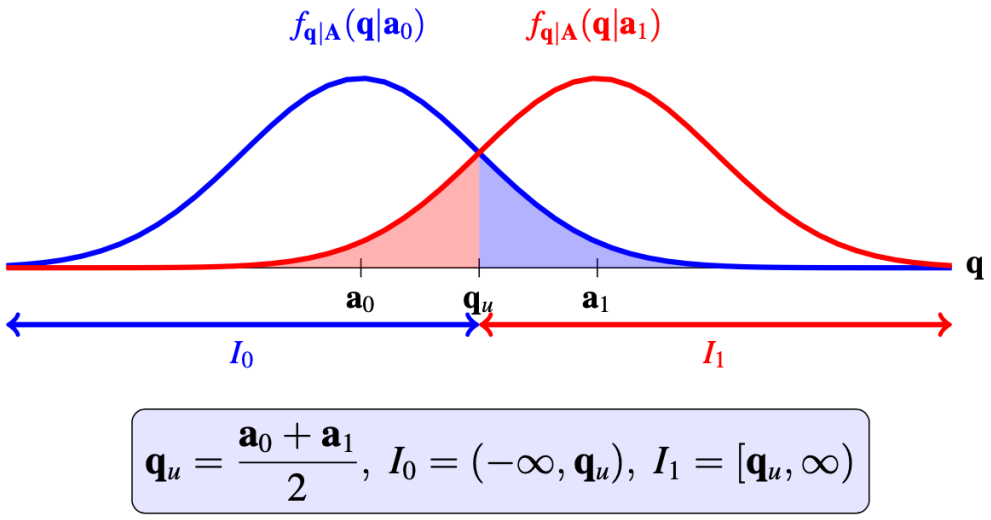


Figure 3.34: Application of the MAP criterion with equiprobable symbols, in which case it is specified in the maximum likelihood (ML) criterion.

If $\mathbf{q} = \mathbf{q}_0$ is lower than $\mathbf{q}_u$, then

$$p_B(b_0)\ f_{\mathbf{q}|\mathbf{A}}(\mathbf{q}_0|\mathbf{a}_0) > p_B(b_1)\ f_{\mathbf{q}|\mathbf{A}}(\mathbf{q}_0|\mathbf{a}_1),$$

or, equivalently

$$f_{\mathbf{q}|\mathbf{A}}(\mathbf{q}_0|\mathbf{a}_0) > f_{\mathbf{q}|\mathbf{A}}(\mathbf{q}_0|\mathbf{a}_1),$$

so such points will be part of the decision region I_0. If the observation is higher than $\mathbf{q}_u$

$$p_B(b_1)\ f_{\mathbf{q}|\mathbf{A}}(\mathbf{q}_0|\mathbf{a}_1) > p_B(b_0)\ f_{\mathbf{q}|\mathbf{A}}(\mathbf{q}_0|\mathbf{a}_0),$$

or, equivalently

$$f_{\mathbf{q}|\mathbf{A}}(\mathbf{q}_0|\mathbf{a}_1) > f_{\mathbf{q}|\mathbf{A}}(\mathbf{q}_0|\mathbf{a}_0),$$

so such points will be part of the decision region I_1. This leads us to see that the decision regions are

$$I_0 = (\mathbf{q}_1, \infty)$$

and

$$I_1 = (-\infty, \mathbf{q}_1).$$

Intuitively, it can be seen that in this case, since both Gaussian functions have the same variance and scale factor, the threshold is the midpoint between their means

$$\mathbf{q}_u = \frac{\mathbf{a}_0 + \mathbf{a}_1}{2}.$$

When the signal space is one-dimensional, as in the two examples seen above, the values that separate the decision regions are often called *decision thresholds*. If the signal space is multidimensional, the curves ($N = 2$) or surfaces that separate the decision regions are called *decision boundaries*. In the two previous examples the decision threshold is $\mathbf{q}_u$. It is the midpoint between the symbols $\mathbf{a}_0$ and $\mathbf{a}_1$, in the second example, when the symbols are equiprobable. It can be seen that when the symbols are not transmitted with the same probability, the optimal detector tends to increase the decision region of the symbols that are transmitted with the highest probability, which seems quite intuitively reasonable (as well as being supported by the relevant analytical developments). By modifying the *a priori* probabilities to make the symbol b_1 more likely, the threshold has been moved from $\frac{\mathbf{a}_0+\mathbf{a}_1}{2}$, increasing the decision region of the most likely symbol.

Finally, the decision rule for the ML criterion (or MAP with equiprobable symbols) will be specified for the case in which the conditional distributions of the observation are Gaussian, as in the case seen in Section 3.4.3. Introducing the probability density function at the input of the receiver in the formulation of the ML decision maker, it is obtained that the decision for an observation $\mathbf{q} = \mathbf{q}_0$ will be the symbol b_i (or equivalently $\mathbf{q}_0$ belongs to I_i) if it is fulfilled

$$\frac{1}{(\pi N_0)^{N/2}} e^{-\frac{||\mathbf{q}_0 - \mathbf{a}_i||^2}{N_0}} > \frac{1}{(\pi N_0)^{N/2}} e^{-\frac{||\mathbf{q}_0 - \mathbf{a}_j||^2}{N_0}} \qquad j = 0, \cdots, M-1,\ j \neq i.$$

Multiplying both terms of the inequality by $(\pi N_0)^{N/2}$ gives

$$e^{-\frac{||\mathbf{q}_0 - \mathbf{a}_i||^2}{N_0}} > e^{-\frac{||\mathbf{q}_0 - \mathbf{a}_j||^2}{N_0}} \qquad j = 0, \cdots, M-1,\ j \neq i.$$

Bearing in mind that the exponential function is a monotonic increasing function and, therefore, satisfies

$$e^a > e^b \Leftrightarrow a > b,$$

the above expression is equivalent to

$$-\frac{||\mathbf{q}_0 - \mathbf{a}_i||^2}{N_0} > -\frac{||\mathbf{q}_0 - \mathbf{a}_j||^2}{N_0} \qquad j = 0, \cdots, M-1,\ j \neq i,$$

and multiplying by N_0 and taking into account the negative sign, we arrive at the condition

$$||\mathbf{q}_0 - \mathbf{a}_i||^2 < ||\mathbf{q}_0 - \mathbf{a}_j||^2 \qquad j = 0, \cdots, M-1,\ j \neq i.$$

Applying the definition of the norm of a vector

$$||\mathbf{q}_0 - \mathbf{a}_i||^2 = \sum_{k=0}^{N-1} |q_{0,k} - a_{i,k}|^2 = |d(\mathbf{q}_0, \mathbf{a}_i)|^2 .$$

Therefore, it is finally established that the decision rule is limited to choosing the closest symbol to the observation vector $\mathbf{q} = \mathbf{q}_0$. Alternatively, it can be said that the decision region of a symbol, I_i, will be formed by all points in the space of $\mathbf{q}$ that are closer to the symbol, $\mathbf{a}_i$, than to any other symbol of the constellation. A scheme illustrating the decision maker resulting from this criterion, called *minimum Euclidean distance criterion*, is shown in Figure 3.35.

At this point it is necessary to make the following clarifications:

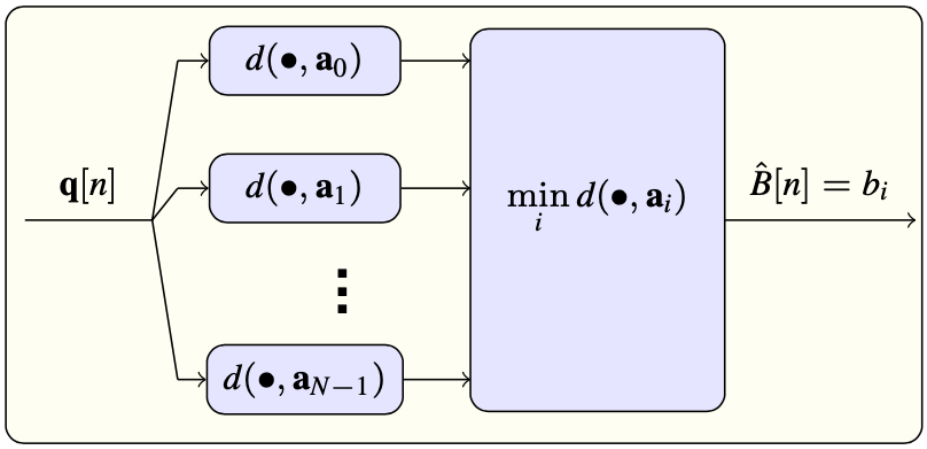


Figure 3.35: Structure of the minimum Euclidean distance detector.

- This last development that has led us to formulate the ML decider as a Euclidean minimum distance decider is based on a Gaussian pdf at the input of the decider, and this Gaussian pdf is given by the nature of the noise thermal usually present in communication channels. When the noise that appears in the channel does not have a Gaussian distribution, as it happens for example in some communication systems based on fiber optics, the development of the optimal decider will lead to different decision rules.

- The definition of dot product that was adopted in Section 3.2.2, is not the only possible one, and was adopted at the time without strict justification. This definition means that the distance measure on the resulting Hilbert space is the Euclidean distance. Now it is possible to justify that in the case of Gaussian statistics for noise, this definition is convenient since it allows us to considerably simplify the decision rule for the case of equiprobable symbols, which on the other hand is the most frequent case in digital communication systems.

3.5.3 Calculation of error probabilities

In the previous section, the design rules of a detector have been obtained. These rules allow obtaining the lowest error probability. In this section we will study how to evaluate the error probability in different types of systems. In a communication system the performance is determined by the transmitted constellation (together with the transmission probability for each symbol), and it is independent of the orthonormal basis that defines the signal space, as long as it is chosen appropriately taking into account the channel characteristics. For this reason, in this section, after stating the problem of evaluating the error probability of a system, the study will be particularized for different types of constellations.

Exact calculation of symbol error rate

The symbol error rate of a digital communications system is defined as the probability of deciding of a wrong symbol at an instant n

$$P_e = P(\hat{B}[n] \neq b_i | B[n] = b_i).$$

The calculation of this error probability is obtained by averaging the conditional error probabilities, i.e., the error probabilities given each of the M possible transmitted symbols. By notation, and taking into account the unique relationship between a symbol and the vector representation of its associated signal, these conditional probabilities will be represented as

$$P_{e|B[n]=b_i} = P_{e|\mathbf{A}[n]=\mathbf{a}_i} \equiv P_{e|\mathbf{a}_i}.$$

The symbol error probability, or symbol error rate, is obtained by averaging these conditional probabilities taking into account the probability with which each symbol is transmitted

$$P_e = \sum_{i=0}^{M-1} p_{\mathbf{A}}(a_i)\ P_{e|\mathbf{a}_i}.$$

Therefore, the basic problem is the calculation of the conditional error probabilities. To do this, it is only necessary to analyze under what circumstances an error occurs when a symbol has been transmitted, $\mathbf{A}[n] = \mathbf{a}_i$, and to evaluate the probability of occurrence of such circumstances. Once again, for ease of notation, the time index n will be omitted (which is possible due to the independence of the symbols and observations at different time instants). When the symbol $\mathbf{A} = \mathbf{a}_i$ has been transmitted, an erroneous decision occurs when $\hat{B} = b_j \neq b_i$ is decided, and this happens when the observation $\mathbf{q}$ is not in the decision region of the symbol, I_i. This means that the conditional error probability for $\mathbf{a}_i$ is the probability of the observation $\mathbf{q}$ being out the the decision region I_i, given that the transmitted symbol is $\mathbf{A} = \mathbf{a}_i$. Since the conditional distribution of the observation when $\mathbf{a}_i$ is transmitted is $f_{\mathbf{q}|\mathbf{A}}(\mathbf{q}|\mathbf{a}_i)$, the conditional error probability is

$$P_{e|\mathbf{a}_i} = \int_{\mathbf{q}\notin I_i} f_{\mathbf{q}|\mathbf{A}}(\mathbf{q}|\mathbf{a}_i)\ d\mathbf{q}.$$

Finally, it is interesting to remark that in order to obtain the symbol error rate of a system, as is clearly deduced from the previous expressions, the following parameters must be known:

- Prior probabilities of each symbol, $p_B(b_i) = p_{\mathbf{A}}(\mathbf{a}_i)$.
- Decision regions of each symbol, I_i.
- Conditional distributions of the observation for each symbol, $f_{\mathbf{q}|\mathbf{A}}(\mathbf{q}|\mathbf{a}_i)$.

In all cases, these parameters must be known for $i \in \{0, 1, \cdots, M-1\}$.

The calculation of the error probability for different types of constellations will be carried out. By default, if the contrary is not explicitly indicated, identical a priori probabilities (equiprobable symbols) will be assumed, as well as Gaussian conditional distributions for the observation, such as those obtained in the transmission over a Gaussian channel (see Section 3.4.3).

Binary constellation ($M = 2$) in one-dimensional space ($N = 1$)

As an initial example, a constellation of two symbols in a one-dimensional space is considered, the symbol $\mathbf{a}_0$ has the coordinate $-A$ and the symbol $\mathbf{a}_1$, $+A$. Under the assumption of equiprobable symbols and transmission on a Gaussian channel, the decision region I_1 will be formed by all

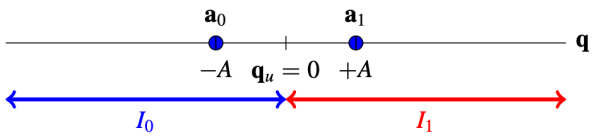


Figure 3.36: Constellation and decision regions for the binary constellation example in one-dimensional space.

the values of $\mathbf{q}$ closer to $\mathbf{a}_1$ than to $\mathbf{a}_0$ and vice versa for the region of decision I_0. The decision threshold is therefore zero, and the decision regions

$$\text{Threshold } q_u = 0 \rightarrow I_1 = [0, \infty),\ I_0 = (-\infty, 0),$$

as shown in Figure 3.36.

Let us now compute the conditional error probabilities, starting with the symbol $\mathbf{a}_0$. In this case, the conditional distribution of the observation for this symbol is the one shown in Figure 3.37.

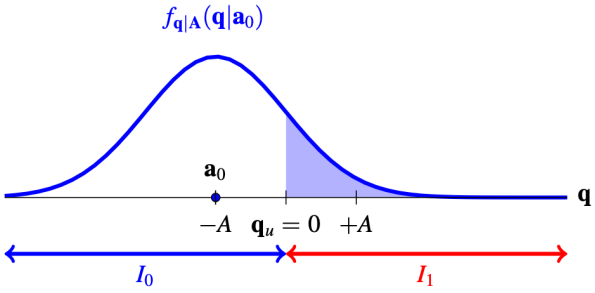


Figure 3.37: Conditional distribution of the observation for the symbol $\mathbf{a}_0$.

It is a Gaussian distribution $f_{\mathbf{q}|\mathbf{A}}(\mathbf{q}|\mathbf{a}_0)$ with mean $-A$, the vectorial representation of the transmitted symbol, and variance $N_0/2$. The conditional error probability for $\mathbf{a}_0$ is the integral of this conditional distribution outside the decision region of $\mathbf{a}_0$, which in this case is the highlighted area in the figure on the distribution

$$P_{e|\mathbf{a}_0} = \int_{\mathbf{q} \notin I_0} f_{\mathbf{q}|\mathbf{A}}(\mathbf{q}|\mathbf{a}_0)\, dq = Q\left(\frac{A}{\sqrt{N_0/2}}\right).$$

As when the symbol $\mathbf{A} = \mathbf{a}_0$ is transmitted the distribution of the observation is $f_{\mathbf{q}|\mathbf{A}}(\mathbf{q}|\mathbf{a}_0)$, and the decision error occurs when the observation $\mathbf{q}$ takes values that fall outside I_0 (in this binary case that means it falls in I_1), the conditional error probability is calculated by integrating the distribution of $\mathbf{q}$ outside I_0.

To compute the conditional error probability for the symbol $\mathbf{a}_1$, in this case the conditional distribution of the observation for this symbol is the one shown in Figure 3.38.

It is a Gaussian distribution $f_{\mathbf{q}|\mathbf{A}}(\mathbf{q}|\mathbf{a}_1)$ with mean $+A$ and variance $N_0/2$. The conditional error probability is the integral of this conditional distribution outside its decision region, which in this case is the highlighted area in the figure

$$P_{e|\mathbf{a}_1} = \int_{\mathbf{q} \notin I_1} f_{\mathbf{q}|\mathbf{A}}(\mathbf{q}|\mathbf{a}_1)\, dq = Q\left(\frac{A}{\sqrt{N_0/2}}\right).$$

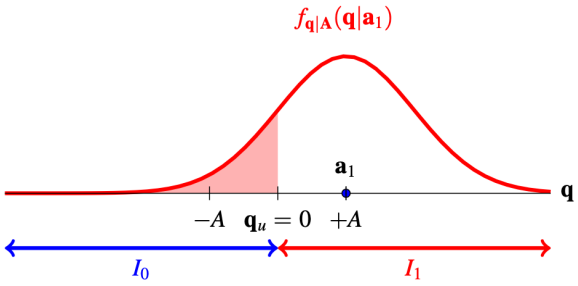


Figure 3.38: Conditional distribution of the observation for the symbol $\mathbf{a}_1$.

Now, as when the symbol $\mathbf{A} = \mathbf{a}_1$ is transmitted, the distribution of the observation is $f_{\mathbf{q}|\mathbf{A}}(\mathbf{q}|\mathbf{a}_1)$, and the decision error occurs when the observation $\mathbf{q}$ takes values that fall outside of I_1 (in this binary case that means that it falls in I_0), the probability of this occurring is calculated by integrating the distribution of $\mathbf{q}$ outside of I_1.

Once the conditional error probabilities are calculated, the symbol error probability is obtained by averaging them, which in this case means

$$P_e = \frac{1}{2}P_{e|a_0} + \frac{1}{2}P_{e|a_1} = \frac{1}{2}\int_{\mathbf{q}\notin I_0} f_{\mathbf{q}|\mathbf{A}}(\mathbf{q}|\mathbf{a}_0)\, dq + \frac{1}{2}\int_{\mathbf{q}\notin I_1} f_{\mathbf{q}|\mathbf{A}}(\mathbf{q}|\mathbf{a}_1)\, dq$$

Figure 3.39 shows the graphical interpretation of the meaning of this probability of error, and how the modification of the decision threshold would increase the error probability.

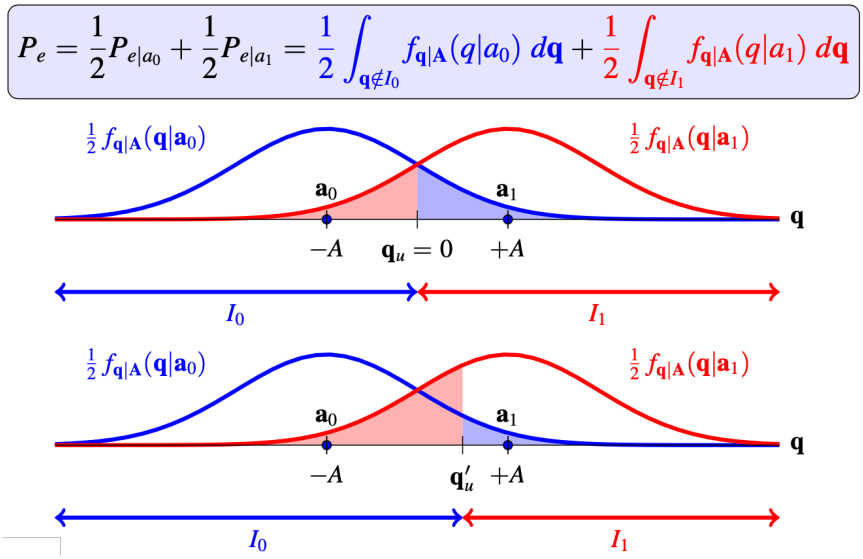


Figure 3.39: Graphical interpretation of the symbol error rate for the optimal detector compared with the symbol error rate for another detector.

From this example, the result can be immediately extrapolated to one-dimensional constellation if the two symbols have any arbitrary values, $\mathbf{a}_0$ and $\mathbf{a}_1$. Regardless of the values of the symbols, for equiprobable symbols and Gaussian conditional distributions the threshold is midway between them.

$$\mathbf{q}_u = \frac{\mathbf{a}_0 + \mathbf{a}_1}{2},$$

so that the distance from each symbol to the threshold, which defines the argument of the function $Q(x)$ to evaluate the integral of the Gaussian distribution, is half the distance between the two

symbols

$$d(\mathbf{a}_0, \mathbf{q}_u) = d(\mathbf{a}_1, \mathbf{q}_u) = \frac{d(\mathbf{a}_0, \mathbf{a}_1)}{2},$$

and, therefore, the conditional error probabilities are equal and the symbol error rate is

$$P_e = Q\left(\frac{d(\mathbf{a}_0, \mathbf{a}_1)}{2\sqrt{N_0/2}}\right).$$

Binary system ($M = 2$) in multidimensional space ($N > 1$)

To illustrate this case, consider the following constellation

$$\mathbf{a}_0 = \begin{bmatrix} A \\ 0 \end{bmatrix}, \ \mathbf{a}_1 = \begin{bmatrix} 0 \\ A \end{bmatrix}.$$

The symbols are 2-dimensional vectors. For the default case of equiprobable symbols and under transmission on a Gaussian channel, the minimum Euclidean distance criterion can be applied again: the decision region of each symbol is formed by the points in the space of $\mathbf{q}$ (in this case the two-dimensional plane formed by the two coordinates, q_0 and q_1) that are closer to that symbol than to the other. The decision frontier that separates the closest points of $\mathbf{a}_0$ from those of $\mathbf{a}_1$ is the line $q_0 = q_1$. Decision region I_0 is the right half plane, and decision region I_1 is the left half plane

$$I_0 = \left\{\mathbf{q} = \begin{bmatrix} q_0 \\ q_1 \end{bmatrix} \middle| q_0 \geq q_1\right\} \qquad I_1 = \left\{\mathbf{q} = \begin{bmatrix} q_0 \\ q_1 \end{bmatrix} \middle| q_0 < q_1\right\},$$

as represented in Figure 3.40.

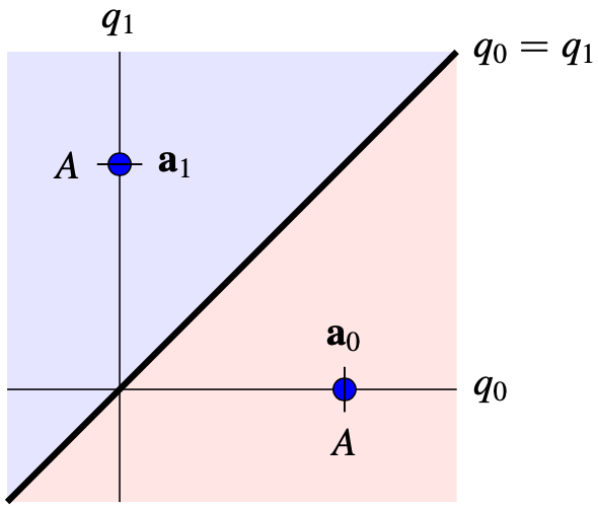


Figure 3.40: Decision boundary and decision regions for the example of a binary system in two-dimensional space.

If the transmitted symbol is $\mathbf{a}_0$ the conditional PDF of the observation, $f_{\mathbf{q}|\mathbf{A}}(\mathbf{q}|\mathbf{a}_0)$, is a Gaussian of mean $\mathbf{a}_0 = \begin{bmatrix} A \\ 0 \end{bmatrix}$, with independent components and variance of each of those components equal to $N_0/2$, and for the symbol $\mathbf{a}_1$ $f_{\mathbf{q}|\mathbf{A}}(\mathbf{q}|\mathbf{a}_0)$ is a Gaussian of mean $\mathbf{a}_1 = \begin{bmatrix} 0 \\ A \end{bmatrix}$, again with

independent components and variance $N_0/2$

$$f_{\mathbf{q}|\mathbf{A}}(\mathbf{q}|\mathbf{a}_i) = \frac{1}{\pi N_0} e^{-\frac{||\mathbf{q}-\mathbf{a}_i||^2}{N_0}} = \frac{1}{\pi N_0} e^{-\frac{(q_0-\sqrt{2T})^2+q_1^2}{N_0}},$$

as shown in Figure 3.41 (in this case they are represented for $A = 1$, above separated for each symbol, and below both together, highligting in colors the support of the decision regions).

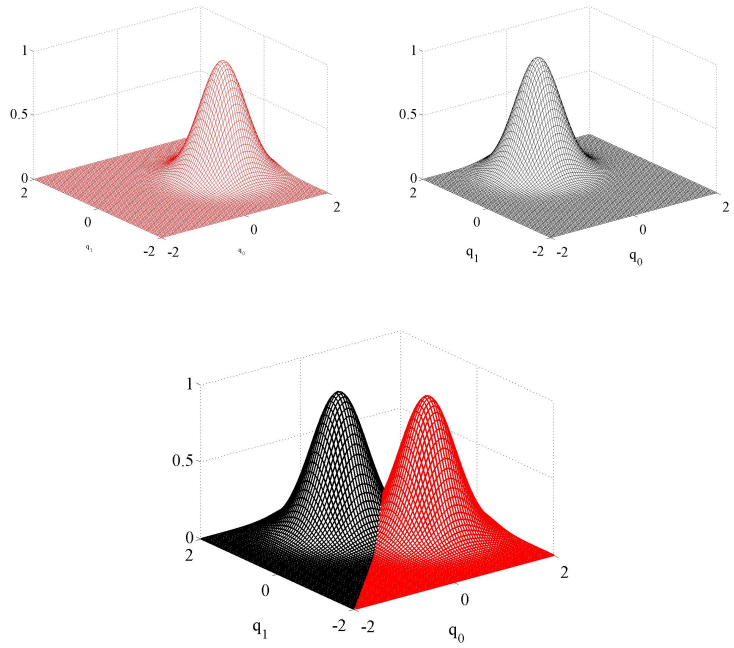


Figure 3.41: Conditional probability density functions for the example of a binary system in two-dimensional space. Each function separately, and both represented within their decision regions.

To calculate the conditional error probabilities, $P_{e|\mathbf{a}_i}$, the conditional distribution for each symbol $\mathbf{a}_i$ must be integrated outside of its decision region, I_i, in this case the defined half-plane for $q_0 > q_1$ or $q_0 < q_1$, respectively for $\mathbf{a}_0$ and $\mathbf{a}_1$, as shown in Figure 3.42.

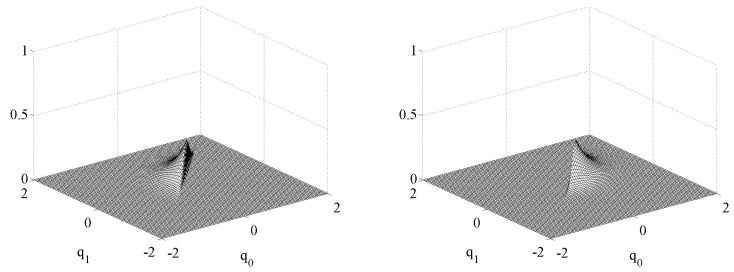


Figure 3.42: Graphical interpretation of the conditional error probabilities for the example of a binary system in two-dimensional space.

As is well known, the integral of a Gaussian function does not have a definite analytic expression. In the one-dimensional case, the tabulated function $Q(x)$, which can be calculated numerically, is used to obtain the value of the integral of a one-dimensional Gaussian distribution. But now, for two-dimensional spaces, there is no tabulated function that computes integrals of Gaussian distributions over half-planes. However, taking advantage of the fact that with two symbols it

is always possible to find a line going through those points, a transformation of the coordinate system can convert the 2-D problem into another equivalent 1-D problem. To do this, the following change of variables is made:

$$q_0' = \frac{1}{\sqrt{2}}(q_0 - q_1),$$
$$q_1' = \frac{1}{\sqrt{2}}(q_0 + q_1 - A).$$

This transformation rotates the constellation 45º and shifts the resulting system $\frac{A}{\sqrt{2}}$ downwards, as illustrated in Figure 3.43, so that the points of the constellation now happen to be on a one-dimensional space, since its coordinate on the second axis, q_1', is null for both symbols,

$$\mathbf{a}_0' = \begin{bmatrix} +\frac{A}{\sqrt{2}} \\ 0 \end{bmatrix}, \ \mathbf{a}_1' = \begin{bmatrix} -\frac{A}{\sqrt{2}} \\ 0 \end{bmatrix}.$$

Therefore, we could consider eliminating the second coordinate and solving the problem for the new resulting one-dimensional constellation, as shown in Figure 3.44. In order to do this, the noise components on the new axes need to be independent, and the noise distribution at coordinate q_0' needs to be checked when performing the transformation, and see if it is still a mean-zero Gaussian distribution. and variance $N_0/2$. Let's see if these conditions are met.

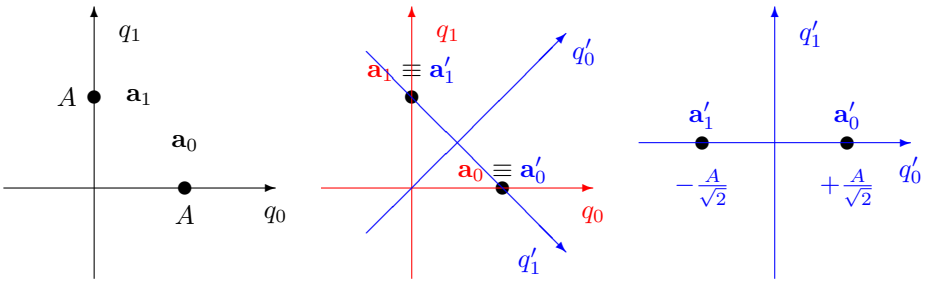


Figure 3.43: Illustration of the geometric transformation implicit in the proposed change of coordinates.

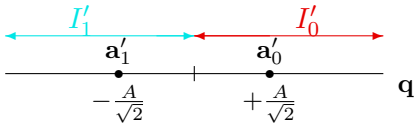


Figure 3.44: Decision regions on the new reference system after the change of coordinates.

Let's start by remembering that when the symbol $\mathbf{A} = \mathbf{a}_i$ is transmitted

$$q_0 = a_{i,0} + z_0, \ q_1 = a_{i,1} + z_1.$$

In the new system, the received observation is

$$q_0' = a_{i,0}' + z_0', \ q_1' = a_{i,1}' + z_1'.$$

If the transmitted symbol is $\mathbf{a}_0$, q_0' takes the form

$$\begin{aligned} q_0'|_{\mathbf{A}=\mathbf{a}_0} &= \frac{1}{\sqrt{2}}\left((a_{0,0}+z_0)-(a_{0,1}+z_1)\right) \\ &= \frac{1}{\sqrt{2}}\left((A+z_0)-(0+z_1)\right) \\ &= \underbrace{\frac{A}{\sqrt{2}}}_{a_{0,0}'} + \underbrace{\frac{1}{\sqrt{2}}(z_0-z_1)}_{z_0'}, \end{aligned}$$

and q_1'

$$\begin{aligned} q_1'|_{\mathbf{A}=\mathbf{a}_0} &= \frac{1}{\sqrt{2}}\left((a_{0,0}+z_0)+(a_{0,1}+z_1)-A\right) \\ &= \frac{1}{\sqrt{2}}\left((A+z_0)+(0+z_1)-A\right) \\ &= \underbrace{\frac{1}{\sqrt{2}}(z_0+z_1)}_{z_1'}, \end{aligned}$$

Similarly, if the transmitted symbol is $\mathbf{a}_1$, q_0' and q_1' take the form

$$\begin{aligned} q_0'|_{\mathbf{A}=\mathbf{a}_1} &= \frac{1}{\sqrt{2}}\left((a_{1,0}+z_0)-(a_{1,1}+z_1)\right) \\ &= \frac{1}{\sqrt{2}}\left((0+z_0)-(A+z_1)\right) \\ &= \underbrace{-\frac{A}{\sqrt{2}}}_{a_{1,0}'} + \underbrace{\frac{1}{\sqrt{2}}(z_0-z_1)}_{z_0'}, \end{aligned}$$

and

$$\begin{aligned} q_1'|_{\mathbf{A}=\mathbf{a}_1} &= \frac{1}{\sqrt{2}}\left((a_{1,0}+z_0)+(a_{1,1}+z_1)-A\right) \\ &= \frac{1}{\sqrt{2}}\left((0+z_0)+(A+z_1)-A\right) \\ &= \underbrace{\frac{1}{\sqrt{2}}(z_0+z_1)}_{z_1'} \end{aligned}$$

Of the two components, only q_0' contains information about the transmitted symbol; q_1' contains only noise. This implies that the dimension of the signal space has been reduced from 2 to 1. The coordinates of the new elements of the constellation with respect to q_0', which we will denote a_0' and a_1', are

$$a_0' = +\frac{A}{\sqrt{2}},$$

and

$$a_1' = -\frac{A}{\sqrt{2}}.$$

Regarding noise, it is easy to show that the terms z_0' and z_1' are independent. The first is proportional to $z_0 - z_1$ and the second to $z_0 + z_1$, since the sum (or subtraction) of two random variables with Gaussian probability density function is another Gaussian random variable and

$$E[(z_0 - z_1)(z_0 + z_1)] = E[z_0^2 - z_1^2] = \frac{N_0}{2} - \frac{N_0}{2} = 0.$$

This implies that the value of q_1' is irrelevant to the decision and that q_0' is a sufficient statistic for detection.

To determine the probability of error we must know the statistics of the noise component, $\frac{1}{\sqrt{2}}(z_0 - z_1)$, since for now we only know that it is Gaussian. The mean value is

$$E[z_0'] = E\left[\frac{1}{\sqrt{2}}(z_0 - z_1)\right] = \frac{1}{\sqrt{2}}E[z_0] - \frac{1}{\sqrt{2}}E[z_1] = 0.$$

And the variance is

$$\begin{aligned}\mathrm{Var}(z_0') = E\left[\left(\frac{1}{\sqrt{2}}(z_0 - z_1)\right)^2\right] =& \frac{1}{2}E[z_0^2] + \frac{1}{2}E[z_1^2] - E[z_0 z_1] \\ =& \frac{1}{2}\frac{N_0}{2} + \frac{1}{2}\frac{N_0}{2} - 0 \\ =& \frac{N_0}{2}.\end{aligned}$$

Finally, if the new constellation is the same as in the one-dimensional case and the noise statistics are the same (zero mean and variance $N_0/2$), so will the error probability,

$$P_e = Q\left(\frac{d(\mathbf{a}_0', \mathbf{a}_1')}{2\sqrt{N_0/2}}\right).$$

The distance between the symbols is the same as in the original constellation, $d(\mathbf{a}_0', \mathbf{a}_1') = d(\mathbf{a}_0, \mathbf{a}_1)$, so it makes no difference to measure in the original space than in the transformed one. In fact, the transformation does not even have to be computed to evaluate the mean probability of error in this case.

This result can be extended to any binary constellation in a space of arbitrary dimension, since it is always possible to define a direction of space through the line that passes through the two points. Thus, in general, the probability of error for binary systems with equiprobable symbols and with transmission on a Gaussian channel is

$$P_e = Q\left(\frac{d(\mathbf{a}_0, \mathbf{a}_1)}{2\sqrt{N_0/2}}\right),$$

regardless of the dimension of the signal space.

M-ary detector in one-dimensional space

To illustrate this case, consider an example of a system with a constellation of $M = 4$ symbols that are transmitted with the same probability over a Gaussian channel. The coordinates of the vector representation of the symbols are

$$\mathbf{a}_0 = -3,\ \mathbf{a}_1 = -1,\ \mathbf{a}_2 = +1,\ \mathbf{a}_3 = +3$$

In this case, since the criterion of minimum Euclidean distance can be applied for the design of the decider, the thresholds of the decider appear at the midpoints between each two symbols of the constellation

$$q_{u1} = -2,\ q_{u2} = 0,\ q_{u3} = +2,$$

and the decision regions are

$$I_0 = (-\infty, -2],\ I_1 = (-2, 0],\ I_2 = (0, +2],\ I_3 = (+2, +\infty),$$

as illustrated in Figure 3.45.

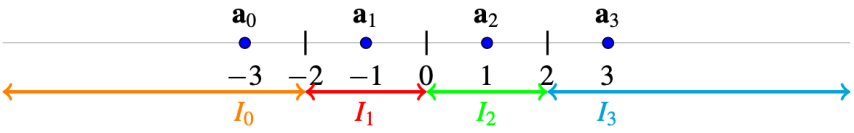


Figure 3.45: Constellation and decision regions for the example of a 4-symbol constellation in one-dimensional space.

The conditional error probability for the symbol $\mathbf{a}_0$ requires to know the conditional distribution of the observation for this symbol, $f_{\mathbf{q}|\mathbf{A}}(\mathbf{q}|\mathbf{a}_0)$, which is the one shown in Figure 3.46. It is a Gaussian distribution with mean -3, the vector representation of the transmitted symbol, and variance $N_0/2$.

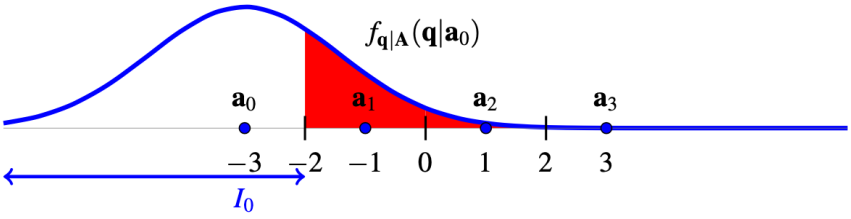


Figure 3.46: Conditional distribution of the observation for the symbol $\mathbf{a}_0$.

The conditional error probability is the integral of this conditional distribution outside of its decision region, which in this case is the highlighted area in the figure

$$P_{e|\mathbf{a}_0} = \int_{q \notin I_0} f_{\mathbf{q}|\mathbf{A}}(q|a_0)\ dq = Q\left(\frac{1}{\sqrt{N_0/2}}\right).$$

The conditional error probability for the symbol $\mathbf{a}_1$ is based on the conditional distribution of the observation for this symbol, $f_{\mathbf{q}|\mathbf{A}}(\mathbf{q}|\mathbf{a}_1)$, which is the one shown in Figure 3.47. It is a Gaussian distribution with mean -1, the vector representation of the transmitted symbol, and variance $N_0/2$.

The conditional error probability is the integral of this conditional distribution outside of its decision region, which corresponds with the highlighted area in the figure

$$P_{e|\mathbf{a}_1} = \int_{q \notin I_1} f_{\mathbf{q}|\mathbf{A}}(q|a_1)\ dq = 2Q\left(\frac{1}{\sqrt{N_0/2}}\right).$$

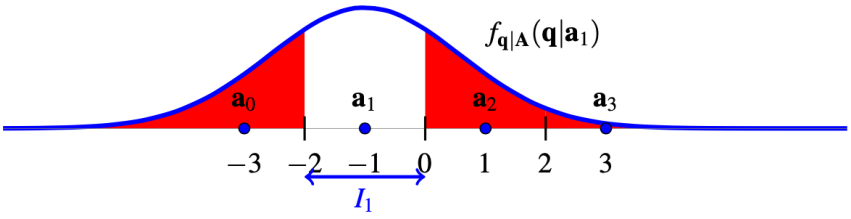


Figure 3.47: Conditional distribution of the observation for the symbol $\mathbf{a}_1$.

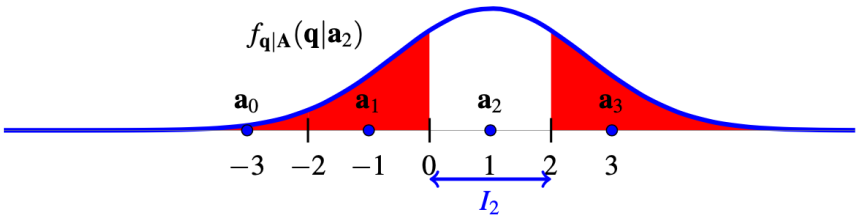


Figure 3.48: Conditional distribution of the observation for the symbol $\mathbf{a}_2$.

The conditional distribution of the observation for the symbol $\mathbf{a}_2$, $f_{\mathbf{q}|\mathbf{A}}(\mathbf{q}|\mathbf{a}_2)$, is a Gaussian distribution with mean +1, the vector representation of the transmitted symbol, and variance $N_0/2$, as shown in Figure 3.48.

The conditional error probability is the integral of this conditional distribution outside its decision region (the highlighted area in the figure)

$$P_{e|\mathbf{a}_2} = \int_{q \notin I_2} f_{\mathbf{q}|\mathbf{A}}(q|a_2)\, dq = 2Q\left(\frac{1}{\sqrt{N_0/2}}\right).$$

Finally, for the symbol $\mathbf{a}_3$, the conditional distribution of the observation for this symbol, $f_{\mathbf{q}|\mathbf{A}}(\mathbf{q}|\mathbf{a}_3)$, is a Gaussian distribution ith mean +3 and variance $N_0/2$, as the one shown in Figure 3.49.

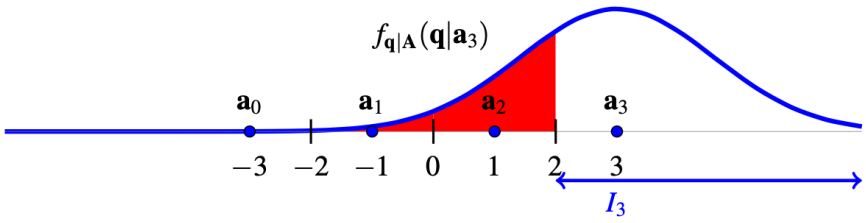


Figure 3.49: Conditional distribution of the observation for the symbol $\mathbf{a}_3$.

In this case, the conditional error probability is

$$P_{e|\mathbf{a}_3} = \int_{q \notin I_3} f_{\mathbf{q}|\mathbf{A}}(q|a_3)\, dq = Q\left(\frac{1}{\sqrt{N_0/2}}\right).$$

In this example, given the symmetry of the decision regions, it can clearly be seen that

$$P_{e|\mathbf{a}_0} = P_{e|\mathbf{a}_3} \text{ and } P_{e|\mathbf{a}_1} = P_{e|\mathbf{a}_2},$$

which could have been used to simplify the calculation.

Once the conditional error probabilities are calculated, the symbol error probability is obtained by averaging them

$$P_e = \sum_{i=0}^{M-1} p_A(\mathbf{a}_i) P_{e|\mathbf{a}_i} = \frac{1}{4} \sum_{i=0}^{M-1} P_{e|\mathbf{a}_i} = \frac{3}{2} Q\left(\frac{1}{\sqrt{N_0/2}}\right).$$

M-ary detector in a multidimensional space

The general case of calculating the symbol error rate in multidimensional constellations is a complex problem due to the shapes that the decision regions can take. For binary constellations it is always possible to transform the problem into a one-dimensional one, thus simplifying the resolution of the integrals necessary to calculate the error probability of each symbol. However, for constellations of more than two symbols this is not always possible. As an example, it is enough to consider the constellations in Figure 3.50, which presents two examples of two-dimensional constellations commonly used in communication systems and their associated decision regions. In some cases it will be possible to calculate the exact probability of error, and in other cases this will not be possible analytically, and it will be necessary to turn to numerical calculations, approximations or bounds of the error probability.

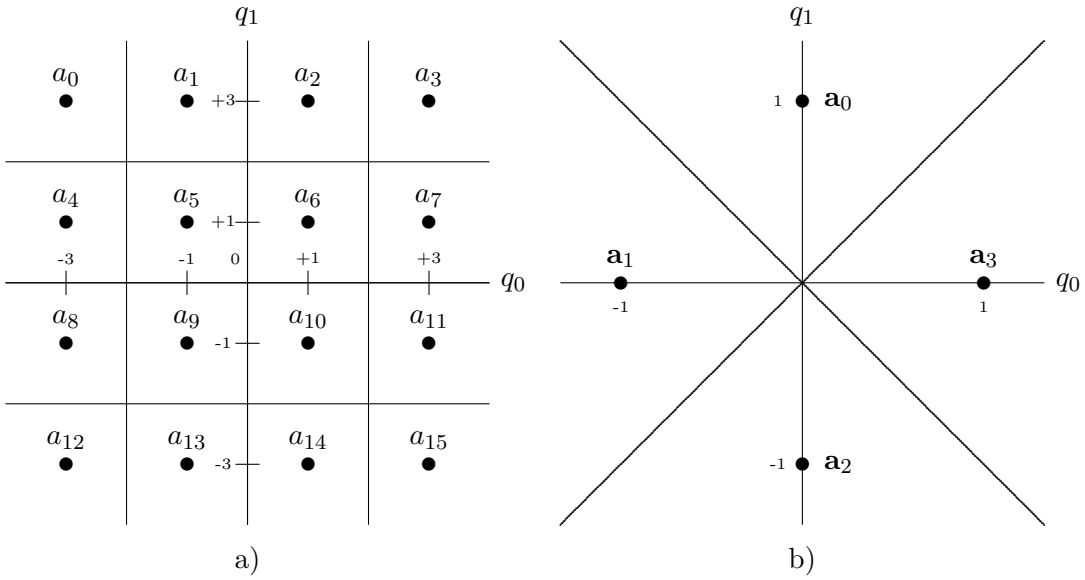


Figure 3.50: Examples of two-dimensional constellations and their decision boundaries.

For systems with constellations like the ones in the figure, it is still possible to easily calculate the symbol error rate. This is possible when the boundaries of the decision regions form a ractangular grid or lattice over space, as occurs with the two constellations in the figure.

For example, first consider the constellation of Figure 3.50 a). Let's see how the conditional error probability would be calculated for one of the symbols, for example $\mathbf{a}_6$. This symbol has coordinates

$$\mathbf{a}_6 = \begin{bmatrix} +1 \\ +1 \end{bmatrix},$$

so the conditional distribution of the observation, $f_{\mathbf{q}|\mathbf{A}}(\mathbf{q}|\mathbf{a}_6)$, is a two-dimensional Gaussian distribution with mean $\mathbf{a}_6$ and variances $\sigma^2 = N_0/2$, as shown in Figure 3.51.

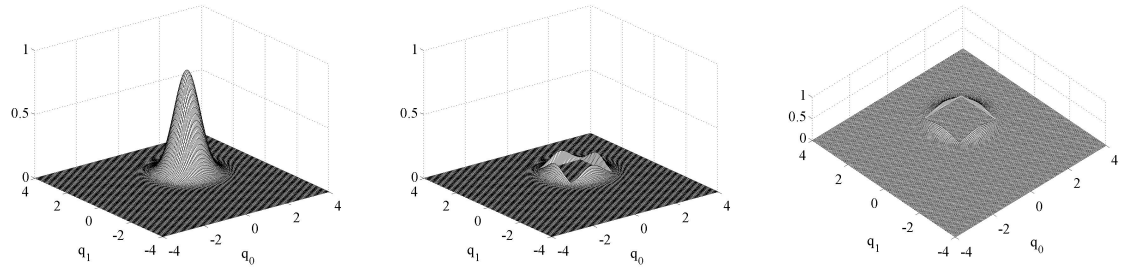


Figure 3.51: Conditional distribution of the observation for the symbol $\mathbf{a}_6$, $f_{\mathbf{q}|\mathbf{A}}(\mathbf{q}|\mathbf{a}_6)$: complete distribution, and distribution outside its decision region (two views from different perspectives).

The decision region of this symbol is a grid, which in this case is aligned with the two axes of the representation.Because of this, it can be written by two independent conditions (one for each axis) that must occur simultaneously. In this case

$$0 \leq q_0 < 2 \text{ and } 0 \leq q_1 < 2.$$

There are no analytic expressions to directly compute the integral of a Gaussian outside of a square or rectangle, but it is possible to transform this 2D problem into two coupled one-dimensional problems. When the decision region can be parameterized into two independent conditions, one for each dimension of space, that must hold simultaneously, the conditional error probability of a symbol can be written as

$$P_{e|\mathbf{a}_i} = 1 - P_{a|\mathbf{a}_i} = 1 - P_{a|a_{i,0}} \times P_{a|a_{i,1}} = 1 - \left[\left(1 - P_{e|a_{i,0}}\right) \times \left(1 - P_{e|a_{i,1}}\right)\right],$$

where the following steps have been followed:

- The conditional probability of error can be written as 1 minus the conditional accuracy (probability of a correct decision), $P_{a|\mathbf{a}_i}$.
- The conditional accuracy $P_{a|\mathbf{a}_i}$ can be written as the product of the accuracies for the two space directions, $P_{a|a_{i,0}} \times P_{a|a_{i,1}}$, since the decision region is established with two independent conditions, one on each direction of space.
- The accuracy in one of the directions of space ($P_{a|a_{i,0}}$ or $P_{a|a_{i,1}}$) can be written as 1 minus the error probability in that direction ($P_{e|a_{i,0}}$ or $P_{e|a_{i,1}}$). And these probabilities are equal to the error probabilities in one-dimensional spaces that have been analyzed previously.

This allows to calculate the conditional error probabilities if the decision regions form a grid or a lattice aligned with the axes of the observation space $\mathbf{q}$.

Let us now see with numerical examples how the total error probability would be calculated for the previous constellation. Decision regions can be grouped into three types, depending on the number of boundaries defining the region.

- Type 1: $\{I_0, I_3, I_{12}, I_{15}\}$
  - A single decision boundary in each direction of space.
- Type 2: $\{I_5, I_6, I_9, I_{10}\}$

  - Two decision boundaries in each direction of space.

- Type 3: $\{I_1, I_2, I_4, I_7, I_8, I_{11}, I_{13}, I_{14}\}$
  - A boundary in one of the directions of space.
  - Two borders in the other direction.

All symbols of the same type have the same conditional probability of error, since the Gaussian distribution centered on the symbol must be integrated outside a region of the same dimensions. Therefore, it is possible to take an example of each type, and extrapolate the results. The chosen examples can be, for example: $\mathbf{a}_0$ (Type 1), $\mathbf{a}_5$ (Type 2), $\mathbf{a}_7$ (Type 3). Thus, the probability of total error will be

$$\begin{aligned} P_e &= \sum_{i=0}^{M-1} p_{\mathbf{A}}(\mathbf{a}_i)\ P_{e|\mathbf{a}_i} \\ &= 4 \times \frac{1}{16} P_{e|\mathbf{a}_0} + 4 \times \frac{1}{16} P_{e|\mathbf{a}_5} + 8 \times \frac{1}{16} P_{e|\mathbf{a}_7}, \end{aligned}$$

which will lead, as we will see below, to

$$P_e = 3Q\left(\frac{1}{\sqrt{N_0/2}}\right) - \frac{9}{4}Q^2\left(\frac{1}{\sqrt{N_0/2}}\right). \tag{3.2}$$

We will start the calculation with the symbol that exemplifies the Type 1 regions, which is $\mathbf{a}_0$. The calculation procedure for this case, which is illustrated in Figure 3.52, is as follows:

- Axis q_0
  - Mean of the 1-D Gaussian distribution: $a_{0,0} = -3$
  - Decision region : $-\infty < q_0 < -2$

$$P_{a|a_{0,0}} = 1 - P_{e|a_{0,0}} = 1 - Q\left(\frac{1}{\sqrt{N_0/2}}\right)$$

- Axis q_1
  - Mean of the 1-D Gaussian distribution: $a_{0,1} = +3$
  - Decision region : $+2 \leq q_1 < +\infty$

$$P_{a|a_{0,1}} = 1 - P_{e|a_{0,1}} = 1 - Q\left(\frac{1}{\sqrt{N_0/2}}\right)$$

- Conditional error probability

$$P_{e|\mathbf{a}_0} = 1 - \left[1 - Q\left(\frac{1}{\sqrt{N_0/2}}\right)\right]^2 = 2Q\left(\frac{1}{\sqrt{N_0/2}}\right) - Q^2\left(\frac{1}{\sqrt{N_0/2}}\right)$$

Next we will perform the calculation for the symbol that exemplifies the Type 2 regions, which is $\mathbf{a}_5$. The calculation procedure for this case, which is illustrated in Figure 3.53, is:

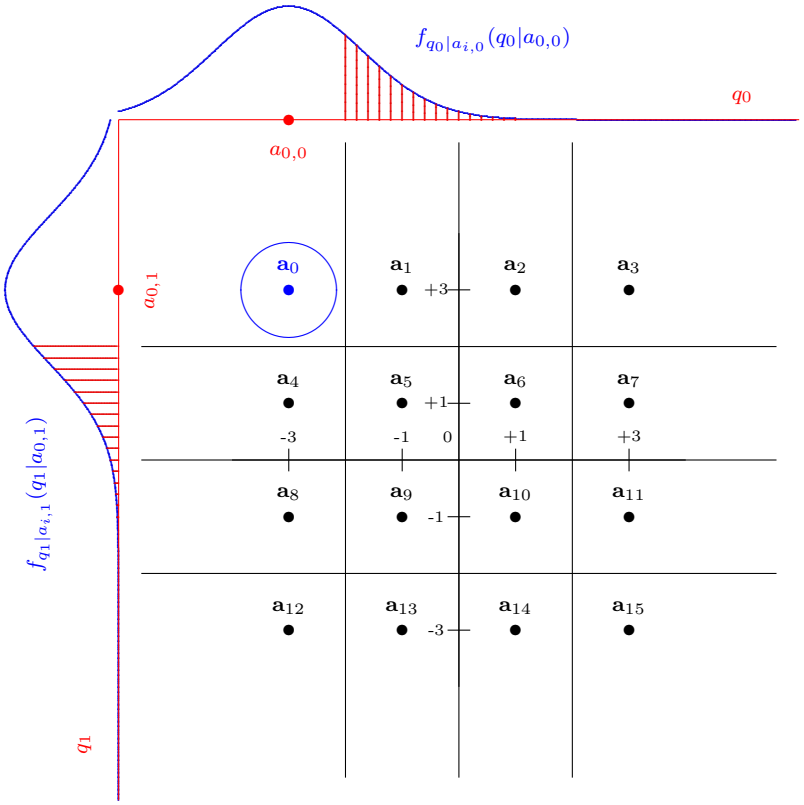


Figure 3.52: Distribuciones condicionales marginales para el símbolo $\mathbf{a}_0$.

- Axis q_0
  - Mean of the 1-D Gaussian distribution: $a_{5,0} = -1$
  - Decision region : $-2 \leq q_0 < 0$

$$P_{a|a_{5,0}} = 1 - P_{e|a_{5,0}} = 1 - 2Q\left(\frac{1}{\sqrt{N_0/2}}\right)$$

- Axis q_1
  - Mean of the 1-D Gaussian distribution: $a_{5,1} = +1$
  - Decision region : $0 \leq q_1 < +2$

$$P_{a|a_{5,1}} = 1 - P_{e|a_{5,1}} = 1 - 2Q\left(\frac{1}{\sqrt{N_0/2}}\right)$$

- Conditional error probability

$$P_{e|\mathbf{a}_5} = 1 - \left[1 - 2Q\left(\frac{1}{\sqrt{N_0/2}}\right)\right]^2 = 4Q\left(\frac{1}{\sqrt{N_0/2}}\right) - 4Q^2\left(\frac{1}{\sqrt{N_0/2}}\right)$$

We will finish with the calculation for the symbol that exemplifies the Type 3 regions, $\mathbf{a}_7$. The calculation procedure for this case, illustrated in Figure 3.52, is now:

- Axis q_0

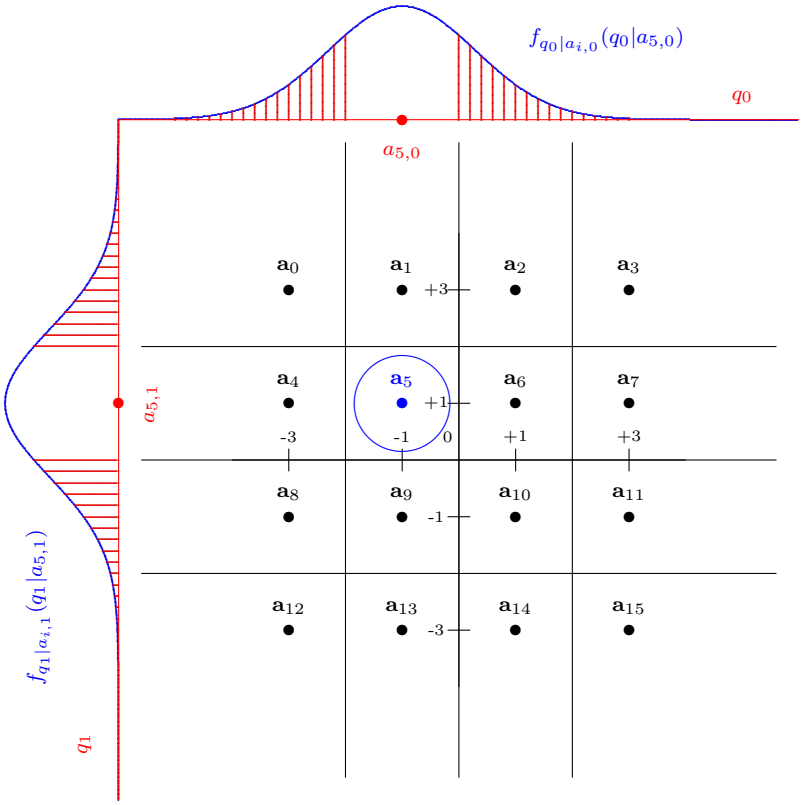


Figure 3.53: Marginal conditional distributions for the symbol $\mathbf{a}_5$.

- Mean of the 1-D Gaussian distribution: $a_{7,0} = +3$
- Decision region : $+2 \leq q_0 < +\infty$

$$P_{a|a_{7,0}} = 1 - P_{e|a_{7,0}} = 1 - Q\left(\frac{1}{\sqrt{N_0/2}}\right)$$

- Axis q_1
  - Mean of the 1-D Gaussian distribution: $a_{7,1} = +1$
  - Decision region : $0 \leq q_1 < +2$

$$P_{a|a_{7,1}} = 1 - P_{e|a_{7,1}} = 1 - 2Q\left(\frac{1}{\sqrt{N_0/2}}\right)$$

- Conditional error probability

$$P_{e|\mathbf{a}_7} = 1 - \left[1 - Q\left(\frac{1}{\sqrt{N_0/2}}\right)\right]\left[1 - 2Q\left(\frac{1}{\sqrt{N_0/2}}\right)\right] = 3Q\left(\frac{1}{\sqrt{N_0/2}}\right) - 2Q^2\left(\frac{1}{\sqrt{N_0/2}}\right)$$

Finally, averaging the conditional error probabilities for the 16 symbols gives the final result shown above, in Eq. (3.2).

We have seen how the probability of error can be calculated for decision regions that form a lattice aligned with the axes of the observation space (given by vector $\mathbf{q}$). The 2D problem can

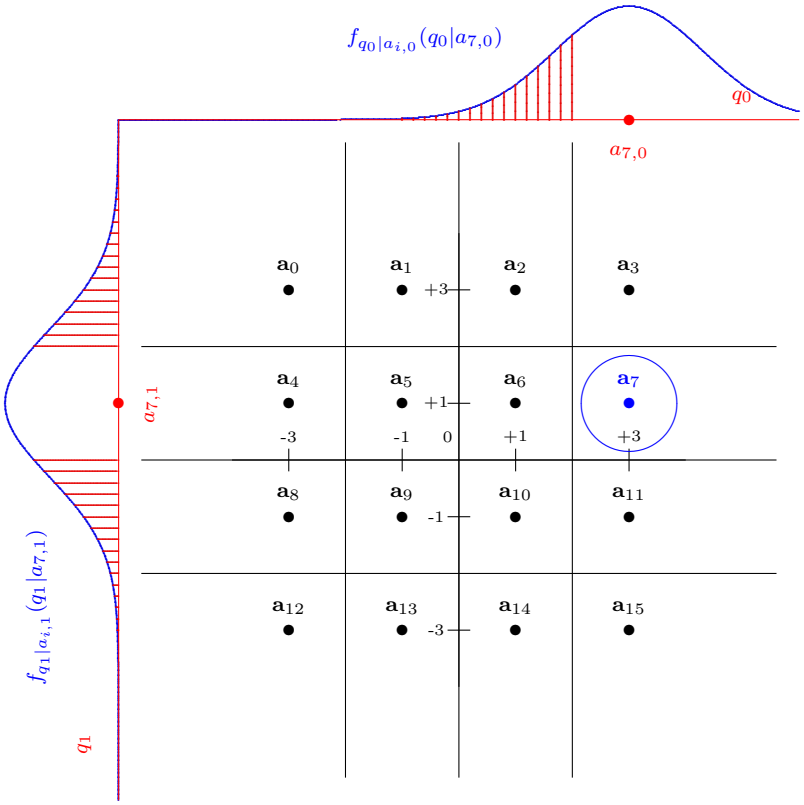


Figure 3.54: Marginal conditional distributions for the symbol $\mathbf{a}_7$.

be turned into two coupled 1D problems that can be solved independently. For the constellation in Figure 3.50 b), the decision regions form a lattice, but this is not aligned with the space axes of $\mathbf{q}$, q_0 and q_1. However, it is easy to see that if a change of variables is applied that produces a 45 degrees rotation, as was done in the example of a binary constellation in two-dimensional space, the decision regions become aligned with the new axes and the symbol error rate can be obtained. In this case, all regions are of Type 1, and since the distance between symbols in this case is $\sqrt{2}$, it is easy to check that the conditional error probability is equal for of all symbols, so its value coincides with the symbol error rate

$$P_e = 1 - \left(1 - Q\left(\frac{1}{\sqrt{N_0}}\right)\right)^2 = 2Q\left(\frac{1}{\sqrt{N_0}}\right) - Q^2\left(\frac{1}{\sqrt{N_0}}\right). \tag{3.3}$$

In the constellations represented in Figure 3.50, error probabilities can be easily calculated because either there are decision regions that form a grid aligned with the axes of space, or there is a simple transformation that converts them into those types of regions. But in other cases this may not be possible, as in the case of a system with the constellation of Figure 3.55, where the eight resulting decision regions are shown. In this case, it is not possible to analytically evaluate the integral of a Gaussian distribution outside of such decision regions. To calculate the symbol error rate, it is necessary to solve the integral of the conditional probability density function of the observation outside the decision region using other procedures. In cases like this, to calculate the probability of error exactly it is necessary to resort to numerical calculations. If analytical expressions are required, approximations or bounds of the error probability can be useful. These analytical tools will be discussed below.

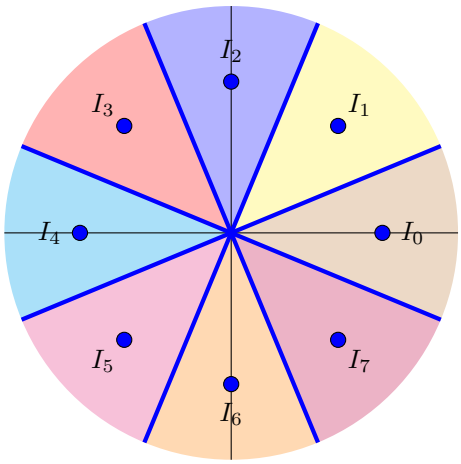


Figure 3.55: Example of a constellation in a two-dimensional signal space in which it is not possible to calculate the probability of error analytically.

3.5.4 Approximation and bounds for the probability of error

Approximations and bounds are useful when it is not possible to accurately assess the probability of error analytically, or when an idea of the magnitude of the probability of error is required, without needing to know the exact value.

Approximation of the probability of error

The symbol error rate depends on the distance between symbols. In the event of an error, it is most likely to erroneously decide on a symbol that is at a minimum distance from the transmitted symbol, while the probability of an error occurring with symbols that are farther apart is considerably lower. The most common approximation assumes that errors will only be made with symbols that are at minimum distance, and that all symbols have the same number of symbols at minimum distance, this number being the largest possible for the constellation. These considerations lead to the approximation

$$P_e \approx k\, Q\left(\frac{d_{min}}{2\sqrt{N_0/2}}\right),$$

where the two parameters appearing in the expression are:

- d_{min}: minimum distance between two symbols in the constellation symbols;
- k: maximum number of symbols at minimum distance of a symbol in the constellation.

These two parameters are very easy to calculate for any constellation. For example, for the constellation of 16 symbols in Figure 3.50 (a), these parameters would be

$$d_{min} = 2,\ k = 4.$$

For the parameter k, in the constellation there are symbols that have 2 symbols at minimum distance (those at the corners, Type 1), symbols that have 4 symbols at minimum distance (the 4 in the center, Type 2) and others that have 3 symbols at minimum distance (symbols of Type

3). The value to be included in the parameter is the maximum number of symbols at minimum distance from a certain symbol, in this case 4. For the constellation in figure (b), the parameters would be

$$d_{min} = \sqrt{2},\ k = 2.$$

Union bound

The approximation of the probability of error is useful when you want to have an idea of the magnitude of the probability of error, and it is not relevant whether the total probability of error is greater or less than the approximate value. However, on certain occasions it is necessary to have an approximate idea of the probability of error, but with the certainty that the probability of error is below the specified (bounding) value. In this case we resort to error probability bounds, values that satisfy that

$$P_e \leq \text{Bound}.$$

Here we will see two bounds: the *union bound*, which delimits the probability of error in a relatively tight way, that is, that the probability of error is not far from the value of the bound, especially for high signal-to-noise ratios. Its calculation becomes involved for constellations with many symbols, and in this case the *loose bound*, a bound with a much simpler analytical expression, can be useful. It is computed easily, but provides values that are somewhat further from the exact error probability than with the union bound.

We will start with the union bound. It is a bound that is expressed as the sum of error probabilities of binary systems. The idea is to limit the conditional error probability of a symbol by the sum of the error probabilities of the $M-1$ binary systems resulting from always using that symbol and each of the remaining $M-1$ symbols of the constellation, that is

$$P_{e|\mathbf{a}_i} \leq \sum_{\substack{j=0 \\ j\neq i}}^{M-1} Q\left(\frac{d(\mathbf{a}_i, \mathbf{a}_j)}{2\sqrt{N_0/2}}\right),$$

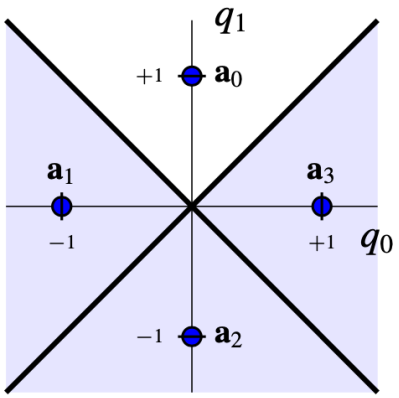
which leads to delimit symbol error rate as

$$P_e \leq \sum_{i=0}^{M-1} p_{\mathbf{A}}(\mathbf{a}_i) \sum_{\substack{j=0 \\ j\neq i}}^{M-1} Q\left(\frac{d(\mathbf{a}_i, \mathbf{a}_j)}{2\sqrt{N_0/2}}\right).$$

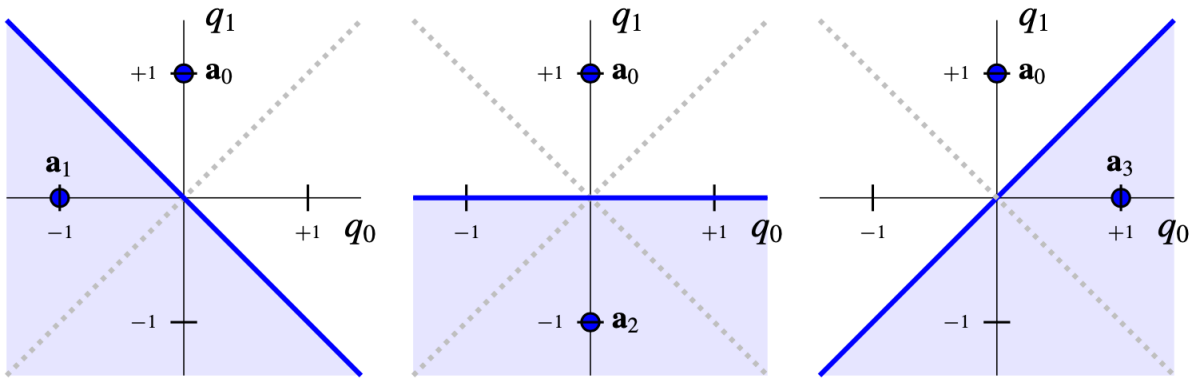
To illustrate the procedure, the constellation in Figure 3.50 b) will be used as an example. In this constellation, the error probability when the transmitted symbol is $\mathbf{a}_0$, $P_{e|\mathbf{a}_0}$, is obtained by integrating $f_{\mathbf{q}|\mathbf{A}}(\mathbf{q}|\mathbf{a}_0)$ out of the decision region I_0, which corresponds to the highlighted area in Figure 3.56 (a).

Instead of calculating this integral, we calculate the error probability that would be obtained using three binary detectors to decide between the symbol $\mathbf{a}_0$ and each of the other three symbols, that is, considering the following binary cases

- $\mathbf{a}_0$ and $\mathbf{a}_1$;
- $\mathbf{a}_0$ and $\mathbf{a}_2$;
- $\mathbf{a}_0$ and $\mathbf{a}_3$;



(a)



(b)

Figure 3.56: The union bound. Above (a), area where $f_{\mathbf{q}|\mathbf{A}}(\mathbf{q}|\mathbf{a}_0)$ must be integrated. Below (b), areas that are added in the union bound.

This involves integrating the probability density function in each of the three regions shown in Figure 3.56 (b). Starting with the symbol $\mathbf{a}_1$, we denote the error probability of the binary decider between $\mathbf{a}_0$ and $\mathbf{a}_1$ as $P_e(\mathbf{a}_0, \mathbf{a}_1)$, and is obtained by integrating the highligted area in the left figure of Fig. 3.56 (b). For a binary detector

$$P_e(\mathbf{a}_0, \mathbf{a}_1) = Q\left(\frac{d(\mathbf{a}_0, \mathbf{a}_1)}{2\sqrt{N_0/2}}\right) = Q\left(\frac{\sqrt{2}}{2\sqrt{N_0/2}}\right) = Q\left(\frac{1}{\sqrt{N_0}}\right).$$

Proceeding in the same way with the other two symbols

$$P_e(\mathbf{a}_0, \mathbf{a}_2) = Q\left(\frac{d(\mathbf{a}_0, \mathbf{a}_2)}{2\sqrt{N_0/2}}\right) = Q\left(\frac{2}{2\sqrt{N_0/2}}\right) = Q\left(\frac{\sqrt{2}}{\sqrt{N_0}}\right),$$

corresponds to the integral of $f_{\mathbf{q}|\mathbf{A}}(\mathbf{q}|\mathbf{a}_0)$ in the area that is highlighted in the central figure of Fig. 3.56 (b), and

$$P_e(\mathbf{a}_0, \mathbf{a}_3) = Q\left(\frac{d(\mathbf{a}_0, \mathbf{a}_3)}{2\sqrt{N_0/2}}\right) = Q\left(\frac{\sqrt{2}}{2\sqrt{N_0/2}}\right) = Q\left(\frac{1}{\sqrt{N_0}}\right),$$

corresponds to the integral of $f_{\mathbf{q}|\mathbf{A}}(\mathbf{q}|\mathbf{a}_0)$ in the area that is highlighted in the right figure of Fig. 3.56 (b).

From these error probabilities with binary deciders, the bound of the union for $P_{e|\mathbf{a}_0}$ is defined as

$$\begin{aligned} P_{e|\mathbf{a}_0} \leq & \sum_{j=1}^{M-1} P_e(\mathbf{a}_0, \mathbf{a}_j) \\ = & P_e(\mathbf{a}_0, \mathbf{a}_1) + P_e(\mathbf{a}_0, \mathbf{a}_2) + P_e(\mathbf{a}_0, \mathbf{a}_3) \\ = & 2Q\left(\frac{1}{\sqrt{N_0}}\right) + Q\left(\frac{2}{2\sqrt{N_0/2}}\right). \end{aligned}$$

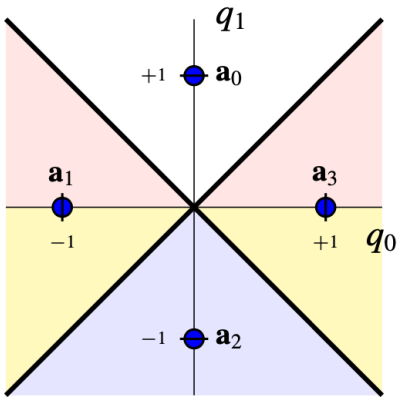


Figure 3.57: The union bound is an upper bound: $f_{\mathbf{q}|\mathbf{A}}(\mathbf{q}|\mathbf{a}_0)$ is integrated once in the red areas, twice in the yellow areas and three times in the blue area.

Figure 3.57 illustrates why the union is an upper bound. It must be taken into account that to obtain $P_{e|\mathbf{a}_0}$, the conditional distribution $f_{\mathbf{q}|\mathbf{A}}(\mathbf{q}|\mathbf{a}_0)$ must be integrated in the blue region of Figure 3.56 (a), i.e., outside of I_0. The three integrals of the same function in the regions highlighted in the three figures of Fig. 3.56 (b) cover all the space outside of I_0. However, some regions are

covered multiple times. This aspect is shown in Figure 3.57. In the red areas, $f_{\mathbf{q}|\mathbf{A}}(\mathbf{q}|\mathbf{a}_0)$ has only been integrated once, in the yellow areas it has been integrated twice and in the blue area it has been integrated three times.

Proceeding in the same way with the rest of the symbols of the constellation, we obtain the so-called *the union for the symbol error rate*, which for a generic constellation of M symbols is written as

$$P_e \leq \sum_{i=0}^{M-1} p_{\mathbf{A}}(\mathbf{a}_i) \sum_{\substack{j=0 \\ j\neq i}}^{M-1} P_e(\mathbf{a}_i, \mathbf{a}_j) = \sum_{i=0}^{M-1} p_{\mathbf{A}}(\mathbf{a}_i) \sum_{\substack{j=0 \\ j\neq i}}^{M-1} Q\left(\frac{d(\mathbf{a}_i, \mathbf{a}_j)}{2\sqrt{N_0/2}}\right)$$

For constellations with equiprobable symbols

$$P_e \leq \frac{1}{M} \sum_{i=0}^{M-1} \sum_{\substack{j=0 \\ j\neq i}}^{M-1} Q\left(\frac{d(\mathbf{a}_i, \mathbf{a}_j)}{2\sqrt{N_0/2}}\right)$$

Particularizing for the example at hand, given the symmetry of the constellation, the union bound results in

$$P_e \leq \frac{1}{4} 4 \left(2Q\left(\frac{1}{\sqrt{N_0}}\right) + Q\left(\frac{2}{2\sqrt{N_0/2}}\right)\right) \tag{3.4}$$

It has been proven that the union bound is an upper bound of the symbol error rate, but we do not know if the bound is close or far from the true value of the real rate. For the previous constellation, both the exact probability and the union bound have been calculated, in Eq. (3.3) and Eq. (3.4), respectively.

Figure 3.58 compares both values as a function of the E_s/N_0 ratio. It can be seen that the union bound provides a value that is higher than the exact error probability and that is also quite tight to the exact value, especially for high values of signal-to-noise ratio (parameterized by E_s/N_0).

The loose bound

In constellations with a large number of symbols, the evaluation of the union bound can be cumbersome due to the number of terms to be evaluated, which grows approximately with the square of the number of symbols in the constellation. For a constellation of 4 symbols, 6 distances must be evaluated (taking advantage of the fact that $d(\mathbf{a}_i, \mathbf{a}_j) = d(\mathbf{a}_j, \mathbf{a}_i)$ and the union bound requires to evaluate the $Q(x)$ funtion for 6 distances. In general, the minimum number of distances to consider in the expression, taking into account the symmetry $d(\mathbf{a}_i, \mathbf{a}_j) = d(\mathbf{a}_j, \mathbf{a}_i)$, is

$$N_{distances} = \sum_{k=1}^{M-1} k$$

This, for some numerical examples supposes

- $M = 4$ symbols lead to $N_{distances} = 6$.
- $M = 8$ symbols lead to $N_{distances} = 28$.
- $M = 16$ symbols lead to $N_{distances} = 120$.

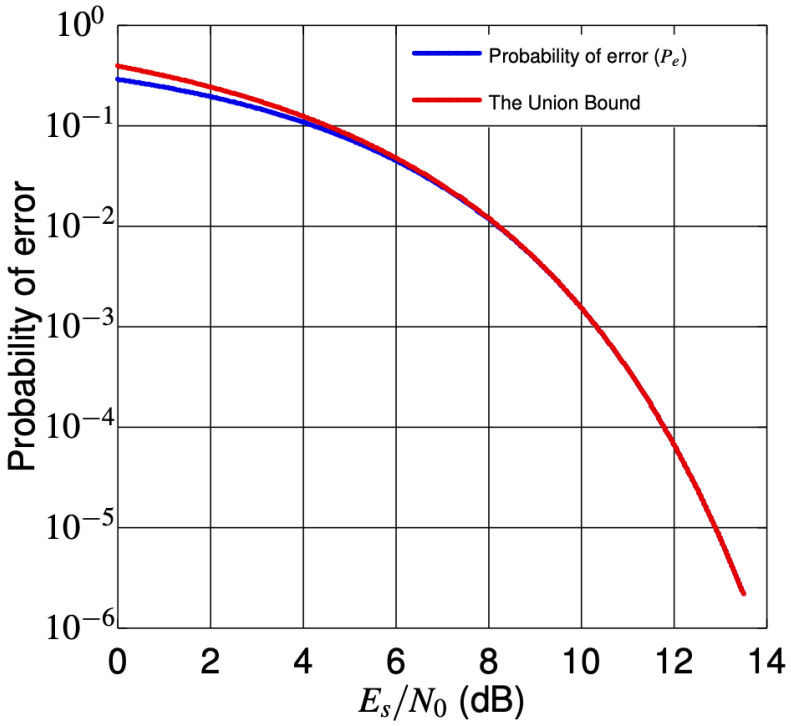


Figure 3.58: Exact error probability and the union bound for the constellation of Figure 3.50 (b), as a function of the E_s/N_0 ratio.

- $M = 64$ symbols lead to $N_{distances} = 2016$.

For constellations with a large number of symbols, the analytical computation without using a computer is relatively expensive, so a more analytically compact bound should be sought. A bound with a compact analytic expression that is commonly used is called the *loose bound*. This bound assumes that all the symbols are at a distance d_{min} from the rest, which is a pessimistic approximation for constellations with a large number of symbols: given a symbol, only a reduced number of the remaining symbols are at minimum distance, and the other ones are at higher distances. Therefore, error probability is bounded by the bound obtained by computing the error probability under this assumption, i.e.

$$P_e \leq (M-1)Q\left(\frac{d_{min}}{2\sqrt{N_0/2}}\right).$$

Figure 3.59 compares the exact error probability with the loose bound for the four-symbol constellation of Figure 3.50 (b), as a function of the relation E_s/N_0. The loose bound provides a value greater than the exact error probability which is now less tight to the exact value than the union bound.

3.5.5 Expressions of the probability of error as a function of E_s/N_0

On many occasions, as in Figures 3.58 or 3.59, it is interesting to express the probability of error (either the exact value, an approximation or a bound), as a function of the signal-to-noise ratio of the system, parameterized by the ratio between the average energy per symbol of the transmitted signal, and the power spectral density of the noise, i.e. E_s/N_0. Obtaining expressions of the

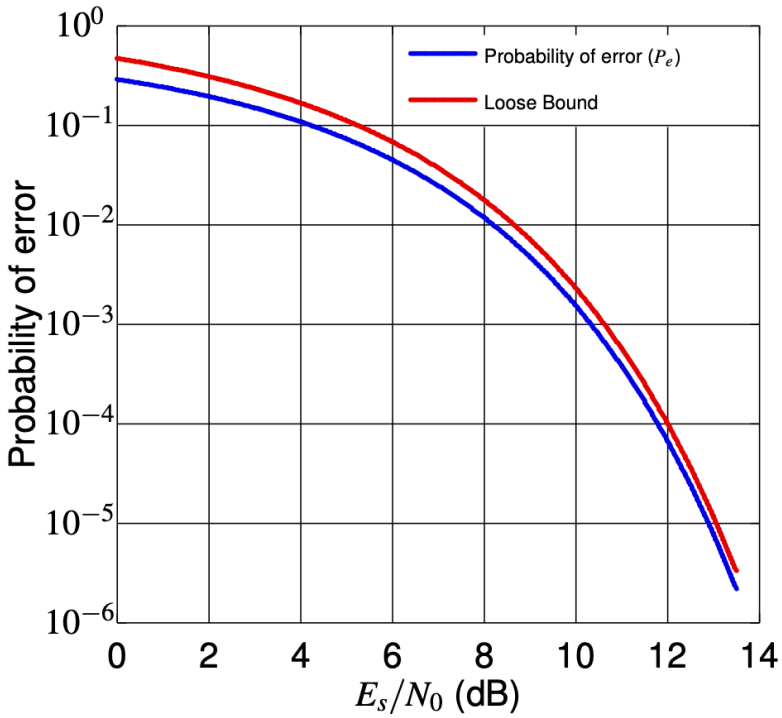


Figure 3.59: Exact error probability and loose bound for the constellation of Figure 3.50 (b), as a function of the E_s/N_0 ratio.

probability of error as a function of this ratio is very simple in most cases. Terms that commonly appear in error probabilities are often written as

$$Q\left(\frac{d_{min}}{2\sqrt{N_0/2}}\right) \text{ or } Q\left(\frac{A}{\sqrt{N_0/2}}\right).$$

In either case, by multiplying and dividing by $\sqrt{E_s}$, they can be rewritten as

$$Q\left(v\sqrt{\frac{E_s}{N_0}}\right), \text{ donde } Q\left(\underbrace{\frac{d_{min}}{\sqrt{2}\sqrt{E_s}}}_{v}\sqrt{\frac{E_s}{N_0}}\right) \text{ ó } Q\left(\underbrace{\frac{A\sqrt{2}}{\sqrt{E_s}}}_{v}\sqrt{\frac{E_s}{N_0}}\right).$$

The factor v is a constant value, which depends on the constellation and can be evaluated in a simple way: it is enough to know the numerical value of E_s and to include it in the previous expression. This factor can be seen as a measure of the constellation efficiency, since the higher v is, the smaller the value of the $Q(x)$ function is for that argument, and the more efficient the constellation is. Next, an example is introduced in which the probability of error is calculated as a function of E_s/N_0 for two different binary constellations.

Example

It has been seen that for any dimension N, any binary system with equiprobable symbols that transmits over a Gaussian channel has a probability of error

$$P_e = Q\left(\frac{d(\mathbf{a}_0, \mathbf{a}_1)}{2\sqrt{N_0/2}}\right)$$

Two cases will be compared:

- Case (a): Symmetric binary constellation ($N = 1$)

$$\mathbf{a}_0 = -A,\ \mathbf{a}_1 = +A$$

- Case (b): Orthogonal constellation ($N = 2$)

$$\mathbf{a}_0 = \begin{bmatrix} A \\ 0 \end{bmatrix},\ \mathbf{a}_1 = \begin{bmatrix} 0 \\ A \end{bmatrix}$$

Distances and average energies per symbol are obtained for each constellation:

- Case (a): $E_s = A^2$, $d(\mathbf{a}_0, \mathbf{a}_1) = 2\ A$
- Case (b): $E_s = A^2$, $d(\mathbf{a}_0, \mathbf{a}_1) = \sqrt{2}\ A$

From these expressions, the efficiency factor v is calculated as

- Case (a): Symmetric binary constellation

$$v = \frac{d(\mathbf{a}_0, \mathbf{a}_1)}{\sqrt{2}\sqrt{E_s}} = \sqrt{2} \rightarrow P_e = Q\left(\sqrt{2\frac{E_s}{N_0}}\right)$$

- Case (b): Orthogonal constellation (extends for $N > 2$)

$$v = \frac{d(\mathbf{a}_0, \mathbf{a}_1)}{\sqrt{2}\sqrt{E_s}} = 1 \rightarrow P_e = Q\left(\sqrt{\frac{E_s}{N_0}}\right)$$

It can be seen that the symmetric binary constellation is more efficient. For a given value of E_s, it performs better than the orthogonal constellation. Figure 3.60 plots the error probability as a function of the E_s/N_0 ratio.

3.6 Encoder

So far, the description of the encoder has been reduced to saying that it performs a transformation of a sequence of symbols, $B[n]$, to a vector representation of the signals that will be associated with each symbol in the sequence, $\mathbf{A}[n]$. This function is illustrated in Figure 3.61. Each of the symbols will be made up of a block of m bits.

Once the receiver has been analyzed, and in view of the expressions obtained in the calculation of the symbol error rate, the design of the encoder to have an efficient communication is now possible.

3.6.1 Encoder design

The encoder design consists of two parts:

1. Design or choice of the constellation to transmit.
2. Binary assignment for each of the symbols.

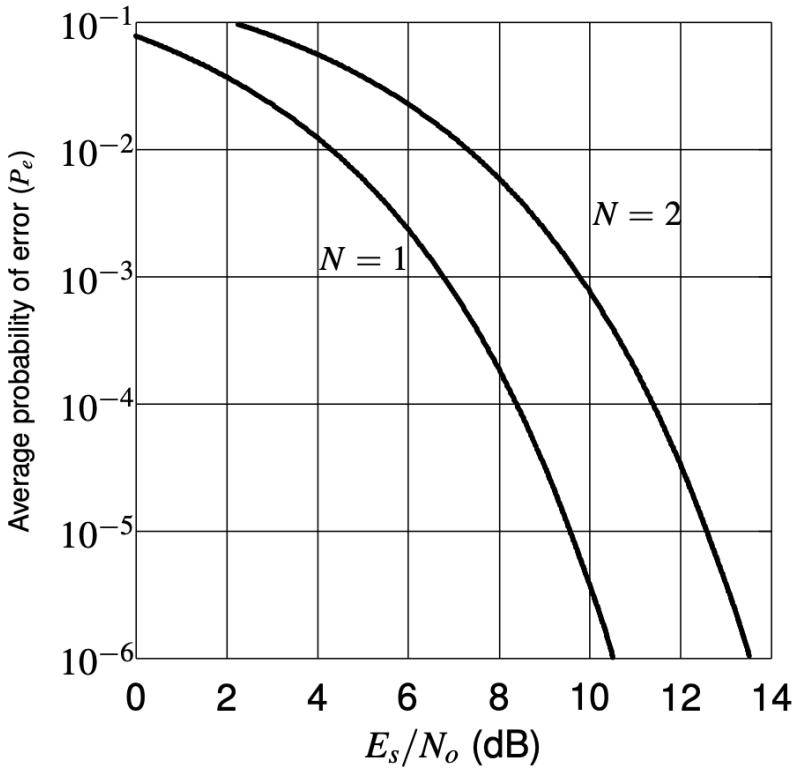


Figure 3.60: Error probability of 1D ($N = 1$) and 2D ($N = 2$) binary constellations as a function of the E_s/N_0 ratio.

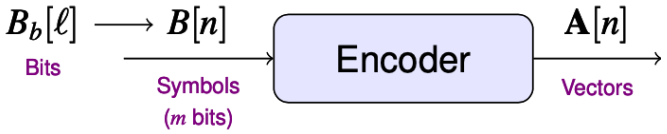


Figure 3.61: Encoder in the transmitter of a digital communications system.

The first part consists of choosing the constellation of M points in a N-dimensional space. This constellation defines the vector representation of the M signals that are used in the transmitter to transport each one of the $M = 2^m$ values of the alphabet of symbols (blocks of m bits). Thus, each possible combination of m bits, symbol b_i, will have an associated signal $s_i(t)$, and what the encoder does is to define the vector representation of the signal, $\mathbf{a}_i$. The modulator will then convert this vector representation into a continuous-time signal $s_i(t)$.

It is important to remember that the vector representation of a set of M signals determines two important factors:

1. The energy of each signal, and therefore, the average energy of the M signals.
2. The distance between each pair of signals, which is related to the energy of the difference signal, and which determines the performance of the system, as seen in previous sections.

Therefore, the choice of the constellation determines these two factors: energy and performance. And in fact, the design of the constellation will be based on these two factors: the best tradeoff between energy and performance will be sought.

The second part of the encoder design is the binary assignment. This assignment consists in assigning to each of the possible values of the alphabet of $\mathbf{A}$ (or equivalently of $B[n]$, since there is a one-to-one assignment $b_i \leftarrow \mathbf{a}_i$) one of the M possible combinations of m bits. As we will see later, the assignment determines the error probability at the bit level, that is, the performance of the system. Therefore, the criteria that should guide the design of the encoder are:

1. Performance (symbol and bit error probabilities).
2. Energy.

Next, we will analyze how the constellation design is carried out for the case of equiprobable symbol transmission when the noise is Gaussian, which is the most frequent case in communication systems.

3.6.2 Design of the constellation

Constellation design depends on two factors: performance and energy. Performaces is related to the distance between symbols, depending fundamentally on the minimum distance between two symbols of the constellation. Remember the approximation of the probability of error

$$P_e \approx k\, Q\left(\frac{d_{min}}{2\sqrt{N_0/2}}\right).$$

So, depending on the type of a priori limitations of the system, the design problem can be posed in two ways:

- If there is an energy limitation, to achieve the lowest possible error probability. The encoder must generate a constellation that has the greatest distance between symbols, but with the limitation that it sets the maximum value of symbol energy or average energy per symbol, E_s that is admissible taking into account the energy limitations of the system.

- If what is limited is the maximum admissible probability of error, this in fact is limiting the mininmum distance between symbol. sit will be necessary to find a constellation with this minimum distance by using the least amount of energy, so as to minimize the cost of the transmitter system.

In any case, the two factors that determine the choice of constellation work in opposite directions: if the distance between symbols in a constellation is increased to decrease the error probability, the energy that is required to transmit each symbol will generally increase. It is therefore necessary to find a compromise between both requirements. It should also be remembered that the energy of a signal, in terms of its vector representation, is related to the distance from the origin (its norm), since

$$\mathcal{E}\{\mathbf{a}_i\} = ||\mathbf{a}_i||^2 = \sum_{k=0}^{N-1} |a_{i,k}|^2 ,$$

that is, that the energy of a signal is the squared norm of its vector representation (the squared distance from the origin of coordinates). In this way, the problem of the optimal design of an encoder can be stated as the problem of placing M points in a space of dimension N so that they have a minimum separation between them, while at the same time are all as close as possible to the origin of coordinates, looking for a compromise between the two design factors.

Related to the idea of placing the points as close as possible to the origin is the following property, which determines one of the conditions that an optimal encoder must meet from the performance-vs-energy compromise point of view: For given intersymbol distances, the mean energy per symbol is minimized when the constellation mean is zero

$$E\left[\mathbf{a}_i\right] = \begin{bmatrix} E[a_{i,0}] \\ E[a_{i,1}] \\ \vdots \\ E[a_{i,N-1}] \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} = \mathbf{0}$$

This property will be illustrated with a simple example of a binary system in a one-dimensional space, which is easily extensible to any other situation. Two symbols are considered in a 1D space, specifically the symbols

$$\mathbf{a}_0 = B - A,\ \mathbf{a}_1 = B + A$$

which are represented in Figure 3.62

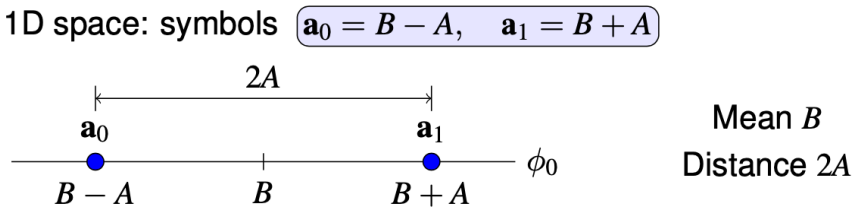


Figure 3.62: Example for a binary and one-dimensional case of the relation of the energy of a constellation with its mean and with the distance between symbols of the constellation.

The two symbol parameters, A and B, separately parameterize the distance between symbols and their mean:

- Mean: B

- Distance between symbols: $2A$

If the average energy per symbol is calculated (assuming equiprobable symbols) for this constellation, we have

$$\begin{aligned} E_s =& \frac{1}{2}\,\mathcal{E}\{\mathbf{a}_0\} + \frac{1}{2}\,\mathcal{E}\{\mathbf{a}_1\} = \frac{1}{2}\,(B-A)^2 + \frac{1}{2}\,(B+A)^2 \\ =& \frac{1}{2}\,\left(B^2 + A^2 - 2AB\right) + \frac{1}{2}\,\left(B^2 + A^2 + 2AB\right) = B^2 + A^2. \end{aligned}$$

It can be seen that two independent terms appear, one related to the mean of the constellation, and the other to the distance between symbols.

- Contribution of the mean: B^2
- Contribution of the distance between symbols: A^2

Therefore, for any distance between symbols, the minimum energy per symbol is obtained when the mean is zero.

$$\textbf{Zero mean}\ (B=0) \rightarrow E_s = A^2.$$

It is trivial to extend this development for higher-dimensional spaces or for a larger number of symbols. So an optimal design will always have this feature: the mean of the constellation symbols will be null (being the mean of dimension N).

Sphere Packing Technique

Taking all these considerations into account, the problem of designing an optimal constellation can be stated. Here, by optimal we mean the one with the best possible compromise between performance and energy consumption. In this case, the design can be posed as the search for constellations that, for a given minimum distance between symbols, require the minimum energy, which implies that the symbols are as close as possible to the origin (having a zero mean). Based on this description, the constellation design problem can be solved by the so-called *sphere packing technique*.

In the sphere packing technique, a symbol is modeled as a sphere of diameter d_{min}, so that two spheres that are in contact are at distance d_{min}, as shown in Figure 3.63. With this model, the design problem can be stated as follows: how to pack M such spheres in an N-dimensional space occupying the smallest possible volume.

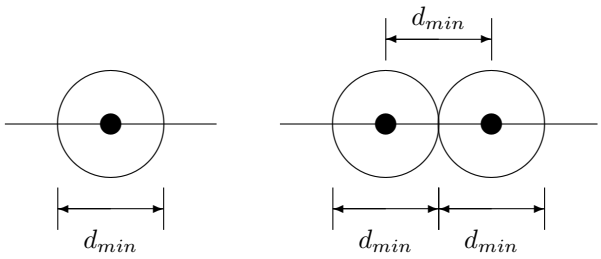


Figure 3.63: Model of a symbol in the sphere packing technique.

Sphere Packing Technique - 1D Space

In a one-dimensional space, the application of this methodology is very simple. The M symbols must be on a line, separated by at least d_{min}, and as close as possible to the origin. This is achieved by placing equispaced symbols, centered on the origin, with half on each side, so that the average is null, with which the coordinates of the M symbols are

$$\mathbf{a}_i \in \left\{ \pm\frac{d_{min}}{2}, \pm 3\,\frac{d_{min}}{2}, \cdots, \pm(M-1)\,\frac{d_{min}}{2} \right\},$$

that is, symbols with coordinates $\pm$ odd numbers times half the minimum distance between symbols. Figure 3.64 shows an example for $M = 4$.

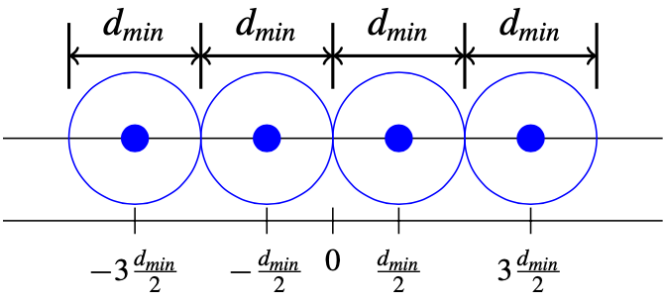


Figure 3.64: Example of application of the sphere packing technique for a 1D space.

Sphere Packing Technique - 2D Space

In 2D space, the best way to pack spheres, as illustrated in Figure 3.65, is in a hexagonal configuration, which is more efficient than a lattice configuration.

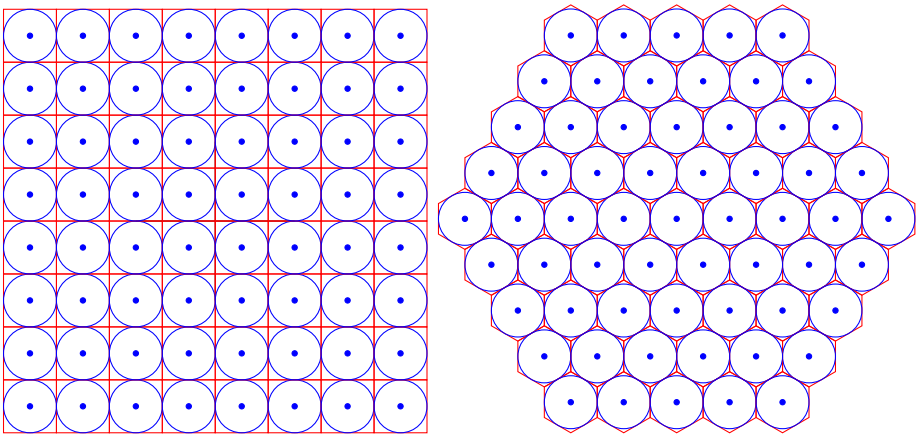
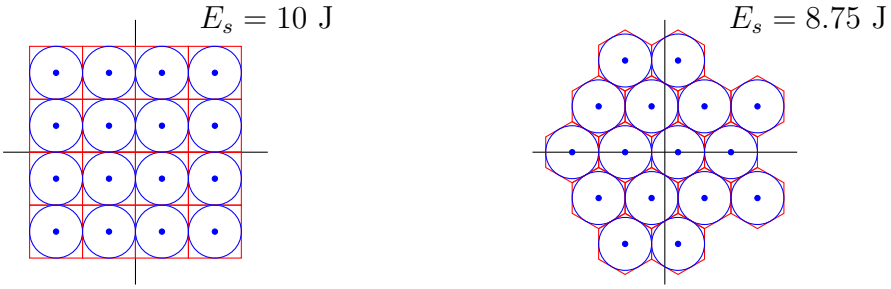


Figure 3.65: Example of rectangular vs hexagonal package.

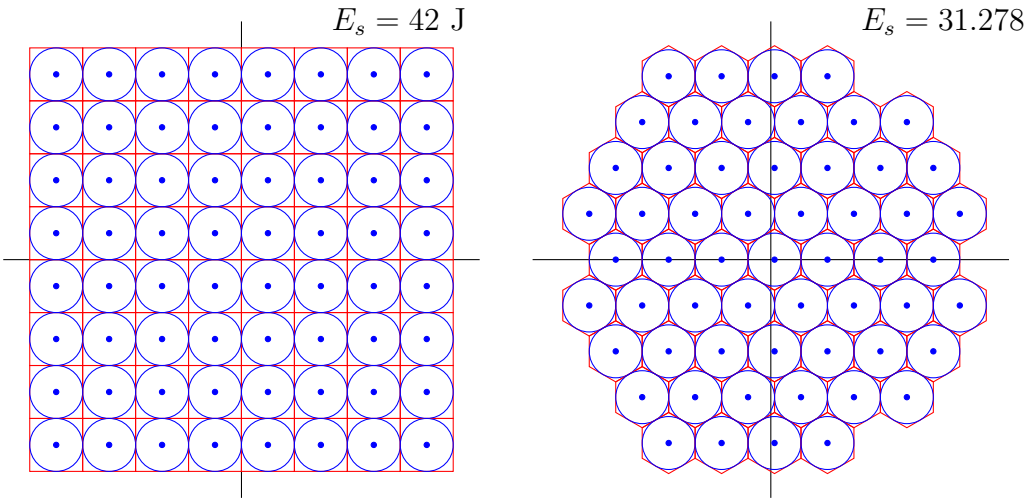
The option of placing a sphere in the center, and placing a circle of spheres touching it around it, results in a much more efficient hexagonal configuration than a rectangular configuration with the symbols arraged in a grid. If there are more symbols, a second ring will be placed on top of the first, also having a hexagonal configuration, and this will continue until M spheres are completed.

Since M is a power of 2, the last ring may not be complete, as in the examples shown in Figure 3.66. In this case, if the initial sphere remains at the origin, the mean of the constellation will not be zero, so it will be necessary to *shift* the constellation so that it has zero mean, as has already been done for the constellation in the figure. It can be seen how the value of the mean energy per symbol, using normalized levels ($d_{min} = 2$) is becomen progresively much lower (proportionally) using the hexagonal package.

QAM and Hexagonal constellations for $M = 4$ (unit ratio in E_s)



QAM and Hexagonal constellations for $M = 16$ (0.875 ratio in E_s)



QAM and Hexagonal constellations for $M = 64$ (0.7447 ratio in E_s)

Figure 3.66: Example of application of the sphere packing technique for a 2D space, for constellations of 4, 16 and 64 points (and comparison with a QAM packing). The mean energy per symbol for each constellation is shown, as well as the ratio among the value of E_s for the hexagonal and QAM package, respectively.

The way in which a constellation with a non-zero mean must be modified so that maintaining the relationship of distances between symbols (which is geometrically equivalent to a shift) it has a zero mean is by subtracting from each symbol the constellation mean. The mean of the

constellation is

$$E[\mathbf{a}_i] = \sum_{i=0}^{M-1} p_A(\mathbf{a}_i)\ \mathbf{a}_i.$$

and the new constellation modified to have zero mean will have symbols with coordinates

$$\mathbf{a}'_i = \mathbf{a}_i - E[\mathbf{a}_i]$$

so that the mean of the new constellation is

$$E[\mathbf{a}'_i] = \sum_{i=0}^{M-1} \underbrace{p_{A'}(\mathbf{a}'_i)}_{p_A(\mathbf{a}_i)}\ \mathbf{a}'_i = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} = \mathbf{0}$$

Below is an example of how to modify a constellation so that, maintaining the relation of distances between symbols, the mean of the new constellation is zero, and therefore, the mean energy per symbol is minimal for these relative distances.

Example

Specifically, we consider a constellation with $M = 4$ symbols, $p_A(\mathbf{a}_i) = \frac{1}{4}$, $\forall i$, with coordinates

$$\mathbf{a}_0 = \begin{bmatrix} -\frac{1}{2} \\ 0 \end{bmatrix},\ \mathbf{a}_1 = \begin{bmatrix} +\frac{3}{2} \\ 0 \end{bmatrix},\ \mathbf{a}_2 = \begin{bmatrix} -\frac{1}{2} \\ +2 \end{bmatrix},\ \mathbf{a}_3 = \begin{bmatrix} +\frac{3}{2} \\ +2 \end{bmatrix}$$

The constellation is represented in Fig. 3.67.

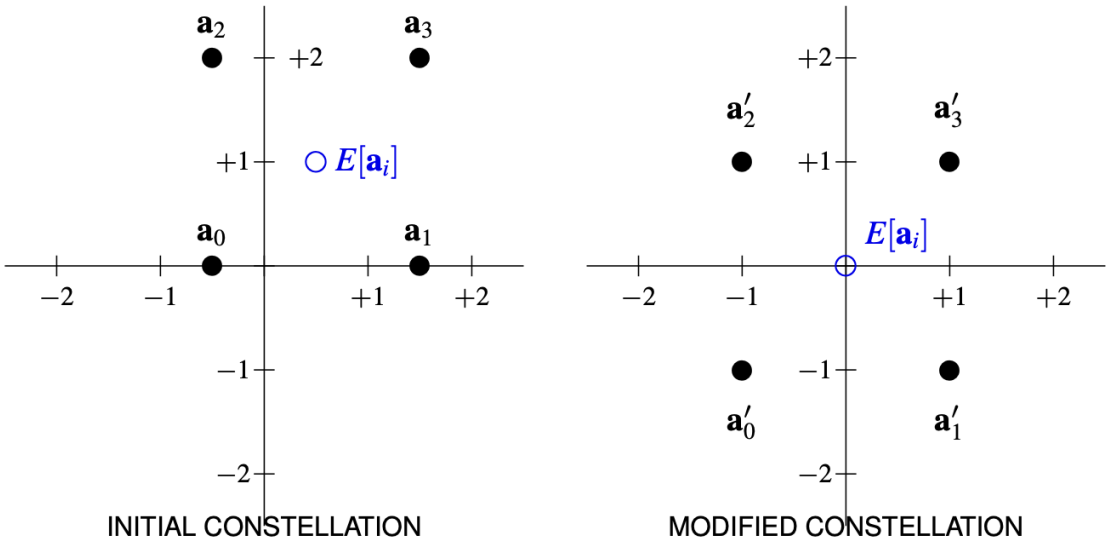


Figure 3.67: Example of modification of a constellation to minimize the average energy per symbol while maintaining the performance (relative distances between symbols).

The average energy per symbol of the constellation in this case is

$$E_s = \frac{1}{4}\left[\left(-\frac{1}{2}\right)^2 + 0^2\right] + \frac{1}{4}\left[\left(\frac{3}{2}\right)^2 + 0^2\right] + \frac{1}{4}\left[\left(-\frac{1}{2}\right)^2 + 2^2\right] + \frac{1}{4}\left[\left(\frac{3}{2}\right)^2 + 2^2\right] = \frac{13}{4} = 3.25$$

In order to minimize the average energy per symbol, the average of the constellation is first calculated, which in this case is equal to

$$E[\mathbf{a}_i] = \frac{1}{4}\begin{bmatrix} -\frac{1}{2} \\ 0 \end{bmatrix} + \frac{1}{4}\begin{bmatrix} +\frac{3}{2} \\ 0 \end{bmatrix} + \frac{1}{4}\begin{bmatrix} -\frac{1}{2} \\ +2 \end{bmatrix} + \frac{1}{4}\begin{bmatrix} +\frac{3}{2} \\ +2 \end{bmatrix} = \begin{bmatrix} +\frac{1}{2} \\ +1 \end{bmatrix}.$$

This average is illustrated with a circle in the figure. Once the average of the constellation is calculated, this average is subtracted from each symbol, so that the following modified constellation is obtained

$$\mathbf{a}'_0 = \mathbf{a}_0 - E[\mathbf{a}_i] = \begin{bmatrix} -1 \\ -1 \end{bmatrix},\ \mathbf{a}'_1 = \mathbf{a}_1 - E[\mathbf{a}_i] = \begin{bmatrix} +1 \\ -1 \end{bmatrix},\ \mathbf{a}'_2 = \begin{bmatrix} -1 \\ +1 \end{bmatrix},\ \mathbf{a}'_3 = \begin{bmatrix} +1 \\ +1 \end{bmatrix},$$

which is also represented in Figure 3.67, and which has a null mean. If we now calculate the average energy per symbol of the modified constellation, we have

$$E'_s = \frac{1}{4}\left[(-1)^2 + (-1)^2\right] + \frac{1}{4}\left[(+1)^2 + (-1)^2\right] + \frac{1}{4}\left[(-1)^2 + (+1)^2\right] + \frac{1}{4}\left[(+1)^2 + (+1)^2\right] = 2.$$

It can be seen that it is smaller than in the original constellation, as expected.

3.6.3 Constellations used in communication systems

Although this is the optimal form of constellations from the point of view of the best compromise between performance and energy consumption, there are other factors to take into account when choosing a constellation. In practices, this means that this strategy is not always chosen. In fact, hexagonal constellations are not the most frequently used constellations in practical communications systems. Among the factors that lead to choosing other types of constellations, the following stand out:

Simplicity of the transmitter The coordinates of the different symbols must be able to be expressed with finite precision numbers (not irrational) since it is necessary to store them in the hardware or software of the transmitter. Coordinates of symbols of the irrational type appear in a hexagonal packing. If the value of the coordinates is truncated, the position of the symbols is modified, having a change in performance.

Simplicity of the receiver There are applications in which the cost of the receiver is decisive, to such an extent that it is preferable to sacrifice a lower probability of error in reception to achieve a simplification of the necessary circuitry in the receiver. Depending on the technology used, this can be a determining factor for using constellations that, although they do not have the best performance/energy compromise, allow cheaper implementations for the receiver. An example of a problem that can appear in some technologies is the recovery of the sign of the coordinates of the received symbol (implementing the sign function in an asic is relatively complicated) and this forces the use of constellations where the coordinates are not negative, commonly called *unipolar constellations*. Another example is the implementation of the decider. In hexagonal constellations, the decision regions are hexagonal, and determining if a point is within a given decision region involves comparing that point with 6 lines and checking that it is simultaneously on the correct side of all six lines. Constellations with simpler decision regions may be more convenient in some cases.

Constant energy per symbol In some cases, it can be convenient to transmit a constant energy in all symbol intervals. In this case, all the symbols must have the same norm.

Peak energy/average energy ratio There are amplification technologies that work well when the energy of the signals does not vary too much between symbol intervals, but that generate serious distortions when there are abrupt changes in energy from one interval to another. Constellations like the one in Figure 3.66 have symbols with relatively different energies (very low for symbols close to the origin, and several orders of magnitude higher for symbols further away). Constellations in which the ratio between the peak energy (maximum energy of a symbol), and the average energy per symbol is high, are not suitable for use with this type of amplifiers.

For practical reasons, especially those mentioned above, the most frequently used constellations are those listed below.

QAM Constellations

These constellations, whose name comes from the English acronym for "*Quadrature Amplitude Modulation*", have as their main characteristic that the points of the constellation form a grid on a 2D space, just like the examples shown in the Figure 3.68. To do this, they have the particularity that the number of bits per symbol m is even.

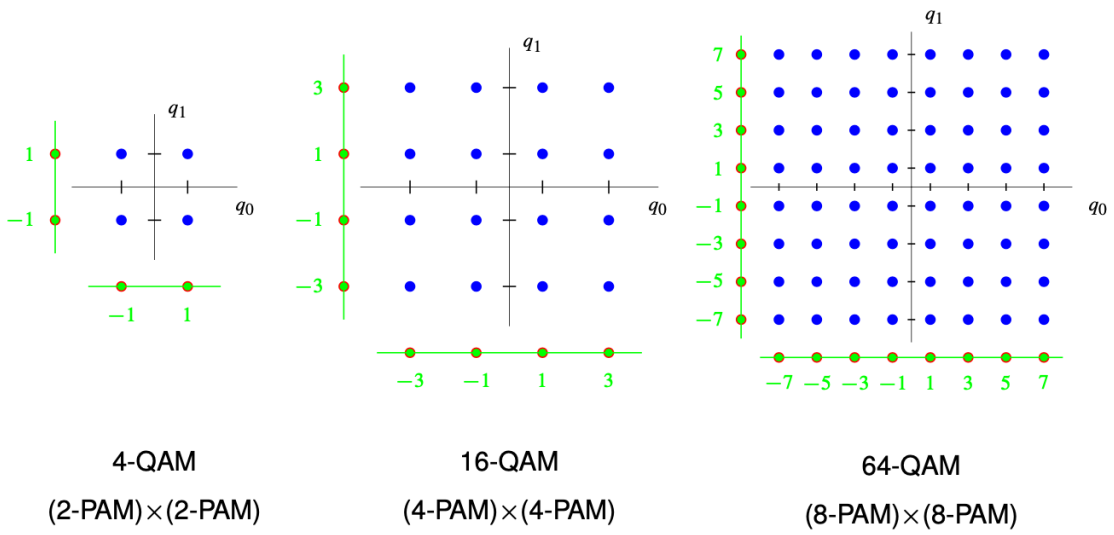


Figure 3.68: QAM constellations (with 4, 16 and 64 symbols).

This type of constellation has the advantage that the transmitter and receiver implementation is relatively simple. For example, the coordinate values take the same values in each direction of space, and the decision regions of each symbol can be expressed as independent conditions for each of the directions of space, so that the processing can be done in the same way. independent for each direction of space in both the transmitter and the receiver.

PSK constellations

These constellations, whose name comes from the English acronym for "*Phase Shift Keying*", have as their main characteristic that the points of the constellation form a circle on a 2D space, just like the examples shown in Figure 3.69.

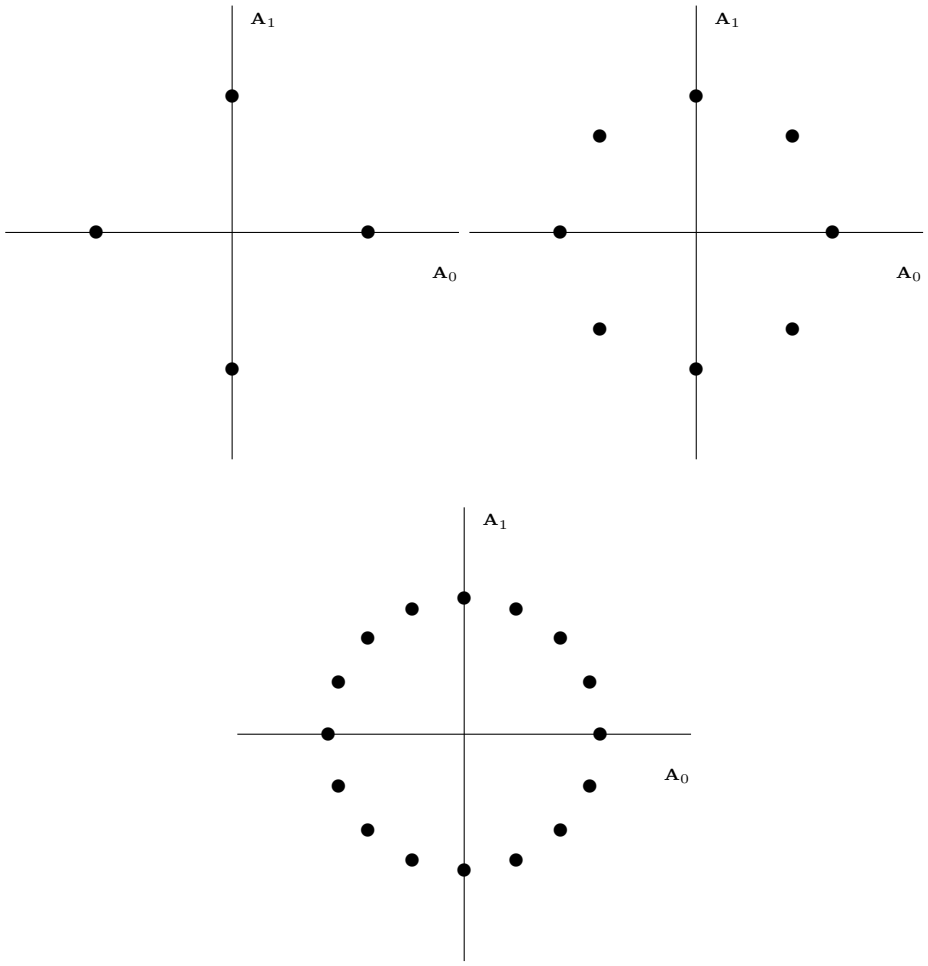


Figure 3.69: Examples of PSK constellations (with 4, 16 and 64 symbols).

Its main advantage is that since the symbols all have the same norm (distance from the origin), they all have the same energy, so the energy remains constant at all symbol intervals during transmission. The transmitter and receiver implementation is also not complex, since the symbols can be encoded as

$$\mathbf{A}[n] = \sqrt{E_s} \times e^{j\phi[n]},$$

that is, that the information of each symbol is in its phase. In the detector, the decision regions are defined from thresholds on the phases, so the only thing to estimate is the phase of the received observation and compare it with the values that define the decision regions.

Unipolar orthogonal constellations

These constellations are constellations in which the dimension of the signal space coincides with the number of signals, such that M orthogonal signals are transmitted whose vectorial representation has a single identical non-zero component, that is

$$\mathbf{a}_0 = \begin{bmatrix} \sqrt{E_s} \\ 0 \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \ \mathbf{a}_1 = \begin{bmatrix} 0 \\ \sqrt{E_s} \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \ \mathbf{a}_2 = \begin{bmatrix} 0 \\ 0 \\ \sqrt{E_s} \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \ \cdots, \ \mathbf{a}_{M-1} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ \vdots \\ \sqrt{E_s} \end{bmatrix}.$$

In this case, all the coordinates are positive, the energy of all the symbols is the same, and the receiver is very simple, since it is enough to find the dimension with the largest component.

3.6.4 Binary Assignment - Gray Coding and BER Calculation

Once a constellation has been selected, it is necessary to carry out the binary assignment: assign to each of the M symbols of the constellation, one of the M possible combinations of m bits. The objective when making the assignment will be to minimize the probability of error at bit level, which we will denote as BER (from "*Bit Error Rate*"), and which is defined as

$$BER = P(\hat{B}_b[\ell] \neq B_b[\ell]).$$

In order to obtain a rule that allows minimizing this error probability, the first thing is to know what it depends on, and how having a certain binary assignment in the encoder affects it. The calculation of this probability of error is done in a similar way to that of the symbol error rate. The conditional bit error probabilities, $BER_{\mathbf{a}_i}$, must be calculated and averaged taking into account the probability with which each symbol in the constellation is transmitted

$$BER = \sum_{i=0}^{M-1} p_A(a_i)\ BER_{a_i}.$$

Regarding the conditional probabilities, we must average the probability of a given wrong symbol decision, taking into account the number of erroneous bits associated to this error. Mathematically, it is

$$BER_{a_i} = \sum_{\substack{j=0 \\ j\neq i}}^{M-1} P_{e|a_i \to a_j} \times \frac{m_{e|a_i \to a_j}}{m},$$

where the parameters involved are defined as follows:

- $P_{e|a_i \to a_j}$: probability of deciding $\hat{\mathbf{A}} = \mathbf{a}_j$ if $\mathbf{A} = \mathbf{a}_i$ is transmitted

$$P_{e|a_i \to a_j} = \int_{\mathbf{q}_0 \in I_j} f_{\mathbf{q}|\mathbf{A}}(\mathbf{q}_0|\mathbf{a}_i)\, d\mathbf{q}_0$$

- $m_{e|a_i \to a_j}$: number of erroneous bits associated to this decision
- m: number of bits per symbol in the constellation

An example of BER calculation for a one-dimensional system

The BER will be calculated for a one-dimensional constellation of 4 equiprobable symbols with coordinates

$$\mathbf{a}_0 = -3,\ \mathbf{a}_1 = -1,\ \mathbf{a}_2 = +1,\ \mathbf{a}_3 = +3.$$

The decision regions are defined by the thresholds $q_{u1} = -2,\ q_{u2} = 0,\ q_{u3} = +2$

$$I_0 = (-\infty, -2],\ I_1 = (-2, 0],\ I_2 = (0, +2],\ I_3 = (+2, +\infty),$$

and the chosen binary assignment (at the moment arbitrarily is)

$$\mathbf{a}_0 \equiv 01,\ \mathbf{a}_1 \equiv 00,\ \mathbf{a}_2 \equiv 10,\ \mathbf{a}_3 \equiv 11.$$

Figure 3.70 illustrates these characteristics of the system under evaluation.

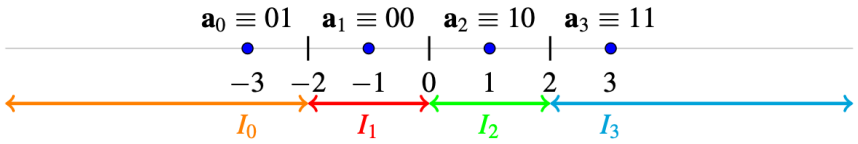


Figure 3.70: Constellation, decision regions, and binary assignment for BER calculation.

First, the conditional BERs for each of the 4 symbols of the constellation will be evaluated. We start with the first symbol, $\mathbf{a}_0$. For this symbol, the conditional distribution $f_{\mathbf{q}|\mathbf{A}}(\mathbf{q}|\mathbf{a}_0)$ is Gaussian with mean $\mathbf{a}_0$ and variance $N_0/2$, so the conditional BER is given by

$$BER_{\mathbf{a}_0} = \underbrace{\left[Q\left(\frac{1}{\sqrt{N_0/2}}\right) - Q\left(\frac{3}{\sqrt{N_0/2}}\right)\right]}_{P_{e|\mathbf{a}_0 \to \mathbf{a}_1}} \times \underbrace{\frac{1}{2}}_{\frac{m_{e|\mathbf{a}_0 \to \mathbf{a}_1}}{m}}$$
$$+ \underbrace{\left[Q\left(\frac{3}{\sqrt{N_0/2}}\right) - Q\left(\frac{5}{\sqrt{N_0/2}}\right)\right]}_{P_{e|\mathbf{a}_0 \to \mathbf{a}_2}} \times \underbrace{\frac{2}{2}}_{\frac{m_{e|\mathbf{a}_0 \to \mathbf{a}_2}}{m}} + \underbrace{\left[Q\left(\frac{5}{\sqrt{N_0/2}}\right)\right]}_{P_{e|\mathbf{a}_0 \to \mathbf{a}_3}} \times \underbrace{\frac{1}{2}}_{\frac{m_{e|\mathbf{a}_0 \to \mathbf{a}_3}}{m}}$$

Figure 3.71 graphically represents the meaning of the probabilities $P_{e|\mathbf{a}_0 \to \mathbf{a}_j}$ in different colors.

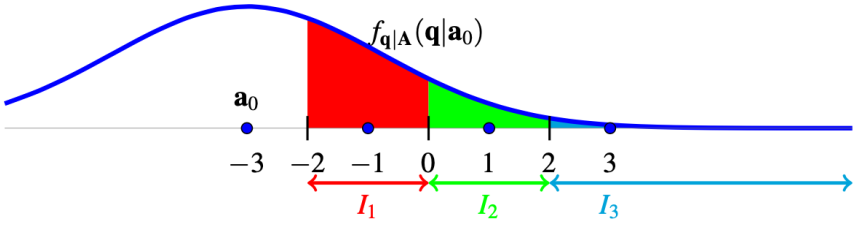


Figure 3.71: Graphical illustration of the probabilities of error between symbols, $P_{e|\mathbf{a}_0\to\mathbf{a}_j}$, para $\mathbf{A}=\mathbf{a}_0$.

For the second symbol, $\mathbf{a}_1$, the conditional distribution $f_{\mathbf{q}|\mathbf{A}}(\mathbf{q}|\mathbf{a}_1)$ is Gaussian with mean $\mathbf{a}_1$ and variance $N_0/2$, so the probability of conditional bit error is

$$BER_{\mathbf{a}_1}=\underbrace{\left[Q\left(\frac{1}{\sqrt{N_0/2}}\right)\right]}_{P_{e|\mathbf{a}_1\to\mathbf{a}_0}}\times\underbrace{\frac{1}{2}}_{\frac{m_{e|\mathbf{a}_1\to\mathbf{a}_0}}{m}}+\underbrace{\left[Q\left(\frac{1}{\sqrt{N_0/2}}\right)-Q\left(\frac{3}{\sqrt{N_0/2}}\right)\right]}_{P_{e|\mathbf{a}_1\to\mathbf{a}_2}}\times\underbrace{\frac{1}{2}}_{\frac{m_{e|\mathbf{a}_1\to\mathbf{a}_2}}{m}}$$
$$+\underbrace{\left[Q\left(\frac{3}{\sqrt{N_0/2}}\right)\right]}_{P_{e|\mathbf{a}_1\to\mathbf{a}_3}}\times\underbrace{\frac{2}{2}}_{\frac{m_{e|\mathbf{a}_1\to\mathbf{a}_3}}{m}}$$

.

Figure 3.72 graphically represents the meaning of the probabilities $P_{e|\mathbf{a}_1\to\mathbf{a}_j}$ in different colors.

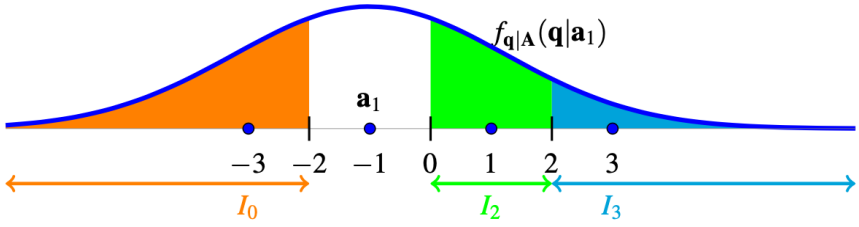


Figure 3.72: Graphical illustration of the probabilities of error between symbols, $P_{e|\mathbf{a}_1\to\mathbf{a}_j}$, para $\mathbf{A}=\mathbf{a}_1$.

For the third symbol, the conditional distribution $f_{\mathbf{q}|\mathbf{A}}(\mathbf{q}|\mathbf{a}_2)$ is Gaussian with mean $\mathbf{a}_2$ and variance $N_0/2$, so that

$$BER_{\mathbf{a}_2}=\underbrace{\left[Q\left(\frac{3}{\sqrt{N_0/2}}\right)\right]}_{P_{e|\mathbf{a}_2\to\mathbf{a}_0}}\times\underbrace{\frac{2}{2}}_{\frac{m_{e|\mathbf{a}_2\to\mathbf{a}_0}}{m}}+\underbrace{\left[Q\left(\frac{1}{\sqrt{N_0/2}}\right)-Q\left(\frac{3}{\sqrt{N_0/2}}\right)\right]}_{P_{e|\mathbf{a}_2\to\mathbf{a}_1}}\times\underbrace{\frac{1}{2}}_{\frac{m_{e|\mathbf{a}_2\to\mathbf{a}_1}}{m}}$$
$$+\underbrace{\left[Q\left(\frac{1}{\sqrt{N_0/2}}\right)\right]}_{P_{e|\mathbf{a}_2\to\mathbf{a}_3}}\times\underbrace{\frac{1}{2}}_{\frac{m_{e|\mathbf{a}_2\to\mathbf{a}_3}}{m}}$$

.

Figure 3.73 graphically represents the meaning of the probabilities $P_{e|\mathbf{a}_2\to\mathbf{a}_j}$ in different colors.

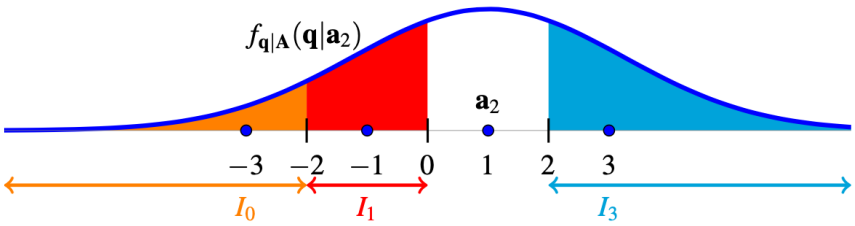


Figure 3.73: Graphical illustration of the probabilities of error between symbols, $P_{e|\mathbf{a}_2\to\mathbf{a}_j}$, para $\mathbf{A}=\mathbf{a}_2$.

Finally, for the last symbol of the constellation, $\mathbf{a}_3$, the conditional distribution of the observation, $f_{\mathbf{q}|\mathbf{A}}(\mathbf{q}|\mathbf{a}_3)$, is Gaussian with mean $\mathbf{a}_3$ and variance $N_0/2$, which means that

$$BER_{\mathbf{a}_3}=\underbrace{\left[Q\left(\frac{5}{\sqrt{N_0/2}}\right)\right]}_{P_{e|\mathbf{a}_3\to\mathbf{a}_0}}\times\underbrace{\frac{1}{2}}_{\frac{m_{e|\mathbf{a}_3\to\mathbf{a}_0}}{m}}+\underbrace{\left[Q\left(\frac{3}{\sqrt{N_0/2}}\right)-Q\left(\frac{5}{\sqrt{N_0/2}}\right)\right]}_{P_{e|\mathbf{a}_3\to\mathbf{a}_1}}\times\underbrace{\frac{2}{2}}_{\frac{m_{e|\mathbf{a}_3\to\mathbf{a}_1}}{m}}$$
$$+\underbrace{\left[Q\left(\frac{1}{\sqrt{N_0/2}}\right)-Q\left(\frac{3}{\sqrt{N_0/2}}\right)\right]}_{P_{e|\mathbf{a}_3\to\mathbf{a}_2}}\times\underbrace{\frac{1}{2}}_{\frac{m_{e|\mathbf{a}_3\to\mathbf{a}_2}}{m}}$$

.

Figure 3.74 graphically represents the meaning of the probabilities $P_{e|\mathbf{a}_3\to\mathbf{a}_j}$ in different colors.

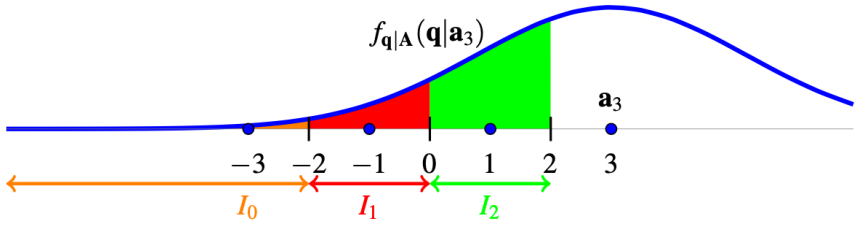


Figure 3.74: Graphical illustration of the probabilities of error between symbols, $P_{e|\mathbf{a}_3\to\mathbf{a}_j}$, para $\mathbf{A}=\mathbf{a}_3$.

Once the conditional BERs are calculated, the total BER is obtained by averaging them, which for this example leads to

$$\begin{aligned}BER=&\frac{1}{4}\times BER_{a_0}+\frac{1}{4}\times BER_{a_1}+\frac{1}{4}\times BER_{a_2}+\frac{1}{4}\times BER_{a_3}\\=&\frac{3}{4}Q\left(\frac{1}{\sqrt{N_0/2}}\right)+\frac{1}{2}Q\left(\frac{3}{\sqrt{N_0/2}}\right)-\frac{1}{4}Q\left(\frac{5}{\sqrt{N_0/2}}\right).\end{aligned}$$

Let's see what happens if the binary assignment is modified. If we were to use now, for example, the binary assignment

$$\mathbf{a}_0\equiv 11,\ \mathbf{a}_1\equiv 00,\ \mathbf{a}_2\equiv 10,\ \mathbf{a}_3\equiv 01.$$

If the previous process were repeated to calculate the BER, a different result would be reached, specifically

$$BER=\frac{5}{4}Q\left(\frac{1}{\sqrt{N_0/2}}\right)-\frac{1}{4}Q\left(\frac{3}{\sqrt{N_0/2}}\right).$$

It can be seen that this value is larger (the dominant term is the one with the smallest argument of the function $Q(x)$, in this case $Q\left(\frac{1}{\sqrt{N_0/2}}\right)$, which in the first case is multiplied by $\frac{3}{4}$, and in the second by $\frac{5}{4}$. What makes an assignment better or worse with respect to the BER it produces?

If we observe the calculation procedure, we will see that modifying the binary assignment does not modify the error probability terms between one symbol and another, that is

$$P_{e|a_i \to a_j} : \text{do not change}$$

Sin embargo, si que cambia la probabilidad de error de bit asignada a cada error de símbolo

$$m_{e|a_i \to a_j} : \text{depend on the assignment}$$

This means, that depending on the chosen binary assignment, the terms

$$\frac{m_{e|\mathbf{a}_i \to \mathbf{a}_j}}{m}$$

They can take one of two values: $\frac{1}{2}$ or $\frac{2}{2} = 1$. Changing the binary assignment implies changing the terms that are associated with one or another value. What is interesting here is to associate these values in such a way that the bit error probability is minimized. This implies trying to assign the lowest value, 1/2, to those terms with higher values for $P_{e|a_i \to a_j}$. And the terms that have a higher symbol error probability value $P_{e|a_i \to a_j}$ are those associated with symbols that are at the minimum distance in the constellation. From here arises the rule that optimizes the binary assignment, and which is called *Gray coding*.

Gray coding The binary assignment of symbols that are at minimum distance must only differ by 1 bit, so that the terms that weigh the most in the BER (for higher values of $P_{e|a_i \to a_j}$) are multiplied by the value $\frac{1}{m}$.

In other words. When a symbol error occurs, it is most likely an error with a symbol that is at minimum distance from the transmitted one. For this reason, it is best to make the most frequent type of error only produce a single bit error over the m coded bits. In this way, using a Gray coding, the bit error probability can be approximated for, reasonably high signal-to-noise ratios, as

$$BER \approx \frac{1}{m} P_e,$$

where $m = \log_2(M)$ is the number of bits per symbol in the constellation.

Any encoder must use a Gray encoding, or if this is not possible, a pseudo-Gray encoding, where the rule is enforced for as many symbols at the minimum distance as possible. For the most common constellations, QAM and PSK, it is always possible to find a Gray coding. In the case of QAM this is easy because the two-dimensional encoding can be done by establishing a one-dimensional Gray encoding with half the bits assigned independently in each direction of space, as shown in Figure 3.75.

For PSK constellations, since the points are on a circle, it is also relatively easy to find a Gray encoding. It is enough to consider a one-dimensional Gray encoding in which the extreme symbols differ by a single bit, and convert it into a circular assignment in 2D space, as in the example of Figure 3.76.

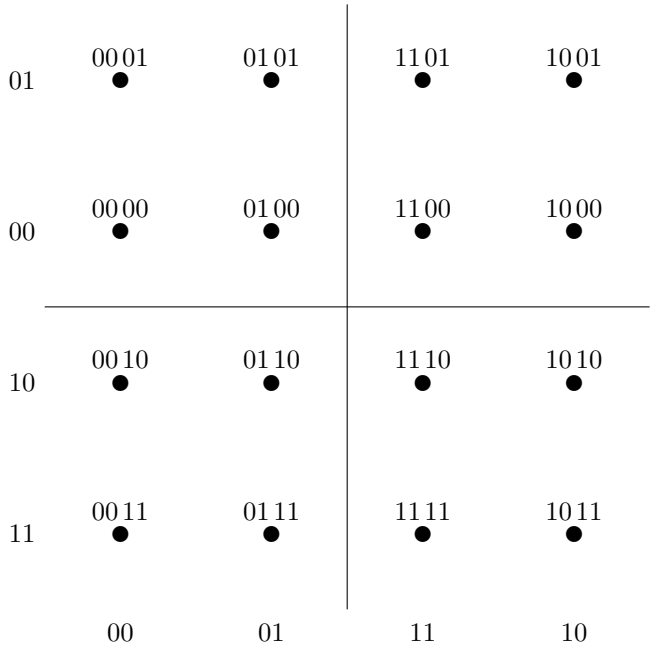


Figure 3.75: Gray coding example for a 16-QAM constellation.

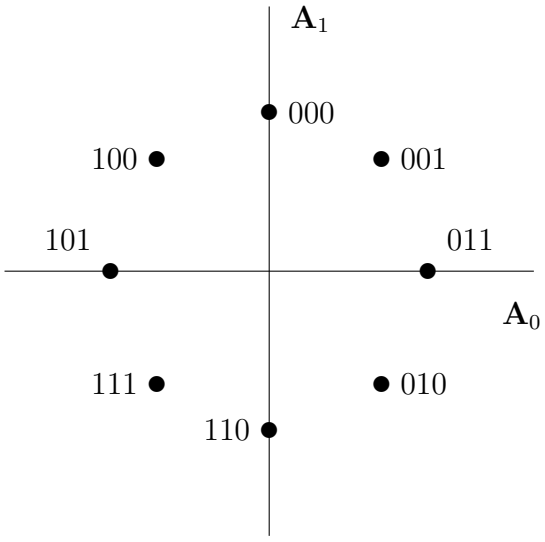


Figure 3.76: Example of Gray encoding for an 8-PSK constellation.

M (symbols)	m (bits/symbol)	E_s with normalized levels ($d_{min} = 2$)	d_{min} with $E_s = 2$
4	2	2	2
16	4	10	0.8944
64	8	42	0.4364
256	16	170	0.2169

Table 3.1: Transmission with M-QAM constellations.

3.6.5 Relationship between bit rate and symbol rate

To finish with the encoder, we just remember the relationship between symbol rate and bit rate. Since each symbol carries m bits, the relationship between bit and symbol rates is obvious.

$$R_b = m \times R_s \text{ bits/s},$$

or

$$R_s = \frac{R_b}{m} \text{ bauds (symbols/s)}.$$

This means that a system with a greater number of bits per symbol (or what is the same, with a greater number of symbols, $M = 2^m$), will have a greater transmission capacity for the same symbol rate. This fact could invite to consider the idea of transmitting very dense constellations (with many symbols). However, another factor must be taken into account when choosing a certain constellation density. If the system is energy limited, denser constellations will imply smaller distances between symbols, which in turn will lead to poorer performance. To illustrate it with an example, let's analyze a system with M-QAM constellations. The Table 3.1 shows for different sizes of the constellation, how the necessary energy increases if the minimum distance is maintained (normalized levels) or how the minimum distance is equivalently reduced if the energy is kept constant average energy per constellation symbol. In Figure 3.77, the constellations are shown to scale for that average energy level per constant symbol, $E_s = 2$ in this case. It can be seen that when choosing denser constellations, although the number of bits per symbol increases, the ratio between performance and energy decreases, so in practice, once again, a compromise must be sought when establishing the optimal density for a constellation.

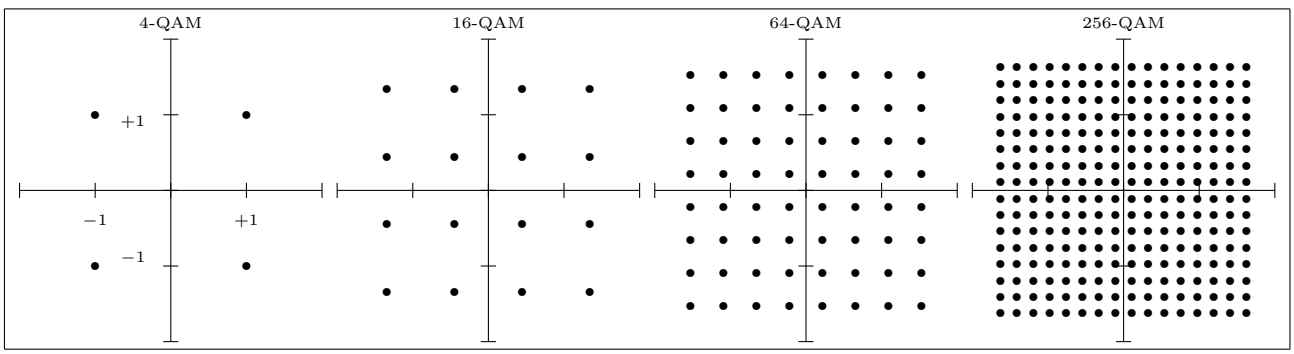


Figure 3.77: Different M-QAM constellations with the same mean energy per symbol ($E_s = 2$).

3.7 Modulator

The modulator is the second functional element of the transmitter, after the encoder. Its function is to transform the sequence of vector representations of the signals to be transmitted, $\mathbf{A}[n]$, into the modulated signal $s(t)$ that contains the information to be transmitted. This function is illustrated in Figure 3.78. The way to do it will be transforming each vector of the sequence $\mathbf{A}[n]$ into a signal of duration T that will define the shape of the signal in the corresponding symbol interval, $nT \leq t < (n+1)T$, in such a way that if $\mathbf{A}[n] = \mathbf{a}_k$ then $s(t) = s_k(t - nT)$ in $nT \leq (n+1)T$. This means that the modulated signal is generated piecewise, by symbol intervals.

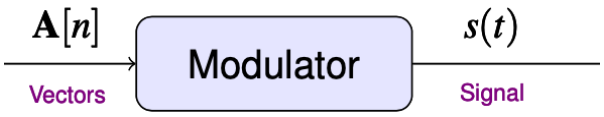


Figure 3.78: Modulator in the transmitter of a digital communications system.

3.7.1 Design of the modulator

The basic function of the modulator, as has been said, is to generate a signal from its discrete representation. Since the relationship between signal and discrete representation is given by the expression

$$s_i(t) = \sum_{j=0}^{N-1} a_{i,j} \times \phi_j(t),$$

what is needed to perform this conversion is to define an orthonormal basis of N elements (dimension of the signal space). And that is what the design of the modulator consists of, in the choice of that orthonormal basis of dimension N

$$\text{Modulator design: } \{\phi_0(t), \phi_1(t), \cdots, \phi_{N-1}(t)\}.$$

If $\mathbf{A}[n] = \mathbf{a}_k$ then $s(t) = s_k(t - nT)$ in $nT \leq (n+1)T$, so that the analytical expression of the complete signal is

$$s(t) \sum_n \sum_{j=0}^{N-1} A_j[n] \times \phi_j(t - nT),$$

where $A_j[n]$ denotes the j-th coordinate of symbol $\mathbf{A}[n]$, i.e.

$$\mathbf{A}[n] = \begin{bmatrix} A_0[n] \\ A_1[n] \\ \vdots \\ A_{N-1}[n] \end{bmatrix}.$$

Just as the design of the modulator was made taking into account two factors, performance and energy, the design of the modulator is made based on a single factor: the characteristics of the channel. An orthonormal basis must be sought whose elements minimize the linear distortion suffered by the signal during its transmission through the channel. If the channel response is $h(t)$ in the time domain, or equivalently its Fourier transform $H(j\omega)$ in the frequency domain, ideally

the basis must be such that there is no linear distortion. This means that the following condition must be satisfied, expressed in the time domain and in the frequency domain

$$\phi_i(t) * h(t) = \phi_i(t) \Leftrightarrow \Phi_i(j\omega) \times H(j\omega) = \Phi_i(j\omega).$$

Although in many cases in this subject it will be considered that a perfect adaptation to the channel conditions can be achieved, in practice this will not be possible, so we will have to settle for selecting signals whose frequency response is in the passband of the transmission channel, which gives rise to a distinction between baseband channels and bandpass channels.

3.7.2 Some examples of modulators and modulated signals

This section shows some examples of modulators, and the type of modulated signals that are generated with them when it is desired to transmit a certain sequence of information bits.

The first example is a system whose constellation (ENCODER) and orthonormal basis (MODULATOR) are those shown in Figure 3.79

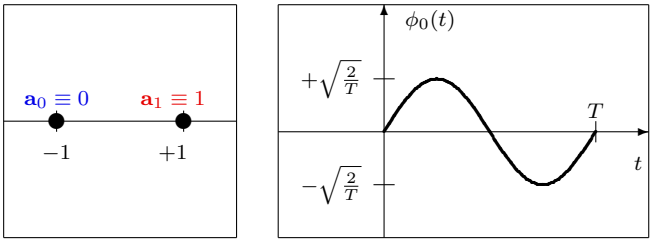


Figure 3.79: Constellation (ENCODER) and orthonormal basis (MODULATOR) of a communications system. Example A.

Taking into account that it is a one-dimensional system, the signals associated with each of the symbols are

$$s_0(t) = -1 \times \phi_0(t),\ s_1(t) = +1 \times \phi_0(t).$$

Being a space of dimension $N = 1$, the two signals are scaled replicas, with different scale factors, of the same signal, the orthonormal basis that defines the modulator, $\phi_0(t)$. The two signals are shown in Figure 3.80.

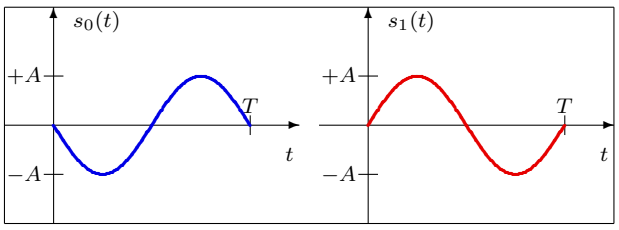


Figure 3.80: Signals associated to each symbol in Example A.

For the following binary sequence of information

$$B_b[\ell] = 0\ 0\ 1\ 0\ 1\ 1\ 0\ 1\ 1\ 0\ 0\ \cdots$$

the corresponding modulated signal is the one shown in Figure 3.81.

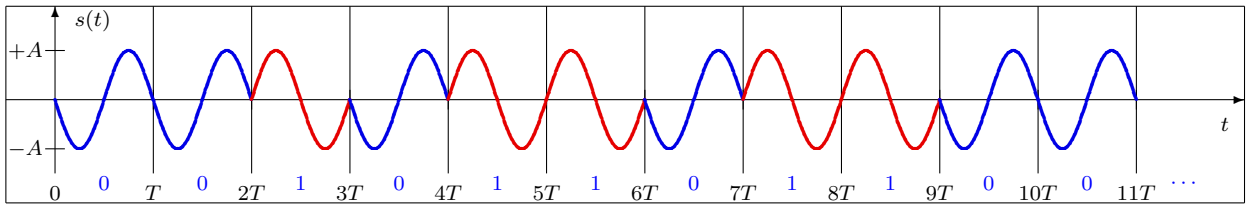


Figure 3.81: Modulated signal for the information sequence transmitted in Example A.

In each symbol interval, one bit is transmitted, which is carried in one of two waveforms, $s_0(t)$ (in blue) carries the bit 0, and $s_1(t)$ (in red) carries the bit 1.

In the second example, a binary system is considered but now in a two-dimensional space. The constellation (ENCODER) and orthonormal basis (MODULATOR) used in this case are those shown in Figure 3.82

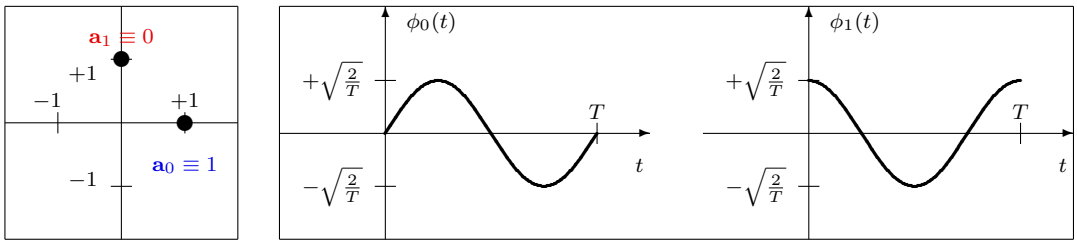


Figure 3.82: Constellation (ENCODER) and orthonormal basis (MODULATOR) of a communications system. Example B.

Taking into account that it is a two-dimensional system, the signals associated with each of the symbols are

$$s_0(t) = +1 \times \phi_0(t) + 0 \times \phi_1(t),\ s_1(t) = 0 \times \phi_0(t) + 1 \times \phi_1(t),$$

which, unlike the previous case, is now formed from the linear combination of two signals, the two elements of the orthonormal basis that defines the modulator. The two signals are shown in Fig. 3.83

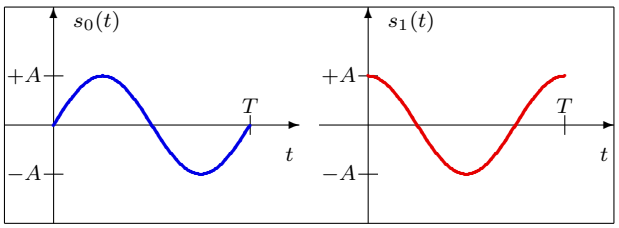


Figure 3.83: Signals associated with each symbol in Example B.

For the same binary sequence of information as in the previous case

$$B_b[\ell] = 0\ 0\ 1\ 0\ 1\ 1\ 0\ 1\ 1\ 0\ 0\ \cdots,$$

the corresponding modulated signal is the one shown in Figure 3.84

As in the previous example, in each symbol interval one bit is transmitted, which is carried in one of two waveforms, $s_0(t)$ (in blue) carries the bit 0, and $s_1(t)$ (in red) carries the 1 bit. What changes from the previous example is the shape of the signals used to transmit each symbol now.

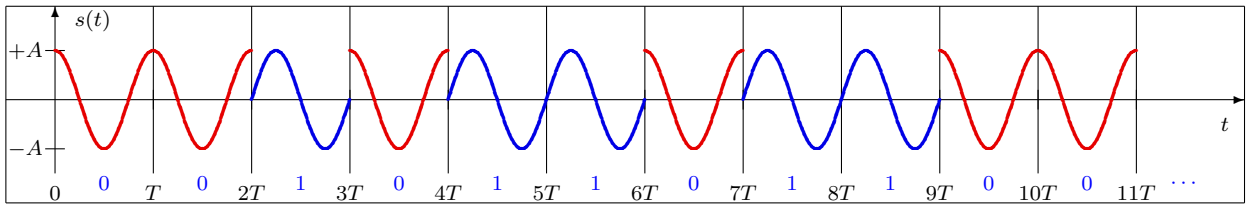


Figure 3.84: Modulated signal for the information sequence transmitted in Example B.

In the third example, using the same modulator as in the previous case, a constellation of $M = 4$ symbols will now be used. The constellation (ENCODER) and orthonormal basis (MODULATOR) used in this case are those shown in Figure 3.85

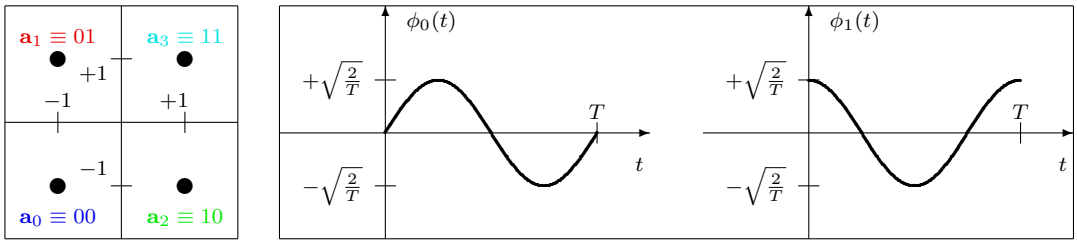


Figure 3.85: Constellation (ENCODER) and orthonormal basis (MODULATOR) of a communications system. Example C.

Once again, it is a two-dimensional system, where the signals associated with each of the symbols are

$$s_0(t) = -1 \times \phi_0(t) - 1 \times \phi_1(t),$$
$$s_1(t) = -1 \times \phi_0(t) + 1 \times \phi_1(t),$$
$$s_2(t) = +1 \times \phi_0(t) - 1 \times \phi_1(t),$$

y

$$s_3(t) = +1 \times \phi_0(t) + 1 \times \phi_1(t),$$

which again are formed from the linear combination of two signals, the two elements of the orthonormal basis that defines the modulator. The four signals are shown in Figure 3.86

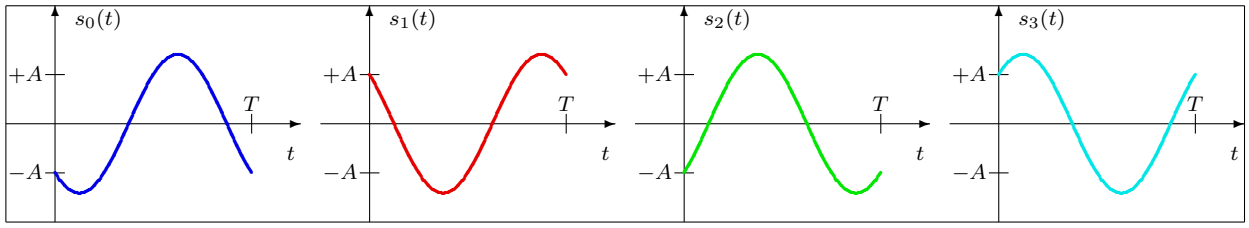


Figure 3.86: Signals associated with each symbol in Example C.

If the binary sequence of information to be transmitted is now

$$B_b[\ell] = 00\ 10\ 11\ 01\ 10\ 00\ 11\ 00\ 01\ 00\ 10\ \cdots,$$

where for convenience the bits have been separated into blocks of size $m = 2$ bits, the corresponding modulated signal is the one shown in Figure 3.87

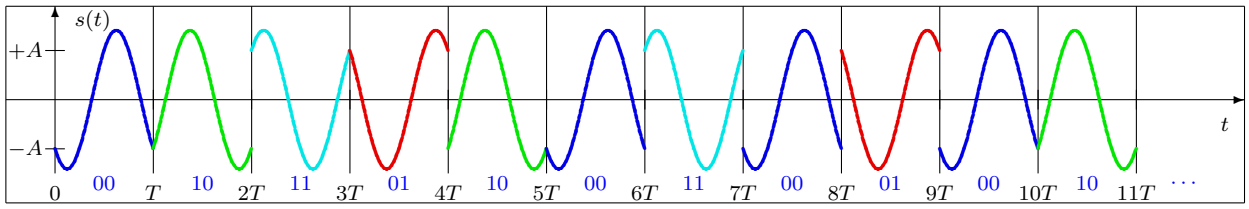


Figure 3.87: Modulated signal for the information sequence transmitted in Example C.

Unlike the previous two examples, which were binary ($M = 2$), now in each symbol interval not one but two bits are transmitted, which are carried in one of the four waveforms, $s_0(t)$ (in blue) carries the pair 00, $s_1(t)$ (in red) carries the pair 01, $s_2(t)$ (in green) carries the pair 10, and $s_3(t)$ (in cyan) transports the pair 11. Now, as it is a constellation of $M = 4$ symbols, each symbol carries $m = \log_2(M) = 2$ bits of information.

Finally, the fourth example uses the same four-symbol constellation as the previous example, with the same binary assignment, but with another modulator of dimension $N = 2$, as shown in Figure 3.88

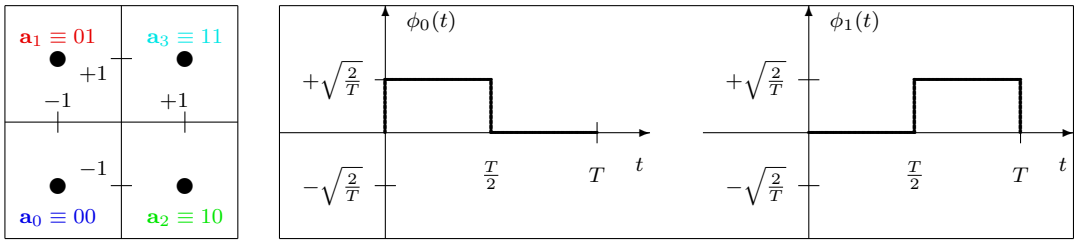


Figure 3.88: Constellation (ENCODER) and orthonormal basis (MODULATOR) of a communications system. Example D.

The analytical expression of the four signals based on the elements of the basis, as the constellation does not change, is the same as in the previous example

$$s_0(t) = -1 \times \phi_0(t) - 1 \times \phi_1(t),$$

$$s_1(t) = -1 \times \phi_0(t) + 1 \times \phi_1(t),$$

$$s_2(t) = +1 \times \phi_0(t) - 1 \times \phi_1(t),$$

y

$$s_3(t) = +1 \times \phi_0(t) + 1 \times \phi_1(t),$$

but with different values for the two functions that form the orthonormal basis, so the four resulting signals are now those shown in Figure 3.89

If the same binary sequence of information is transmitted as in the previous example

$$B_b[\ell] = 00\ 10\ 11\ 01\ 10\ 00\ 11\ 00\ 01\ 00\ 10\ \cdots,$$

the corresponding modulated signal is the one shown in Figure 3.90

Again, in each symbol interval two bits are transmitted, which are carried in one of the four waveforms, $s_0(t)$ (in blue) carries the pair 00, $s_1(t)$ (in red) carries the pair 01, $s_2(t)$ (in green) carries the pair 10, and $s_3(t)$ (in cyan) carries the pair 11. Being the same constellation of $M = 4$

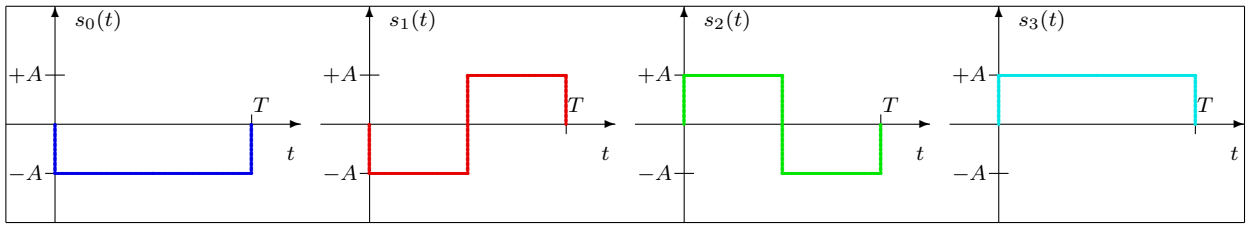


Figure 3.89: Signals associated with each symbol in Example D.

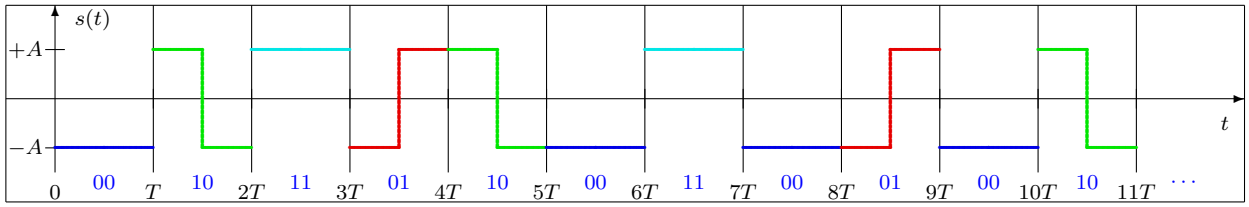
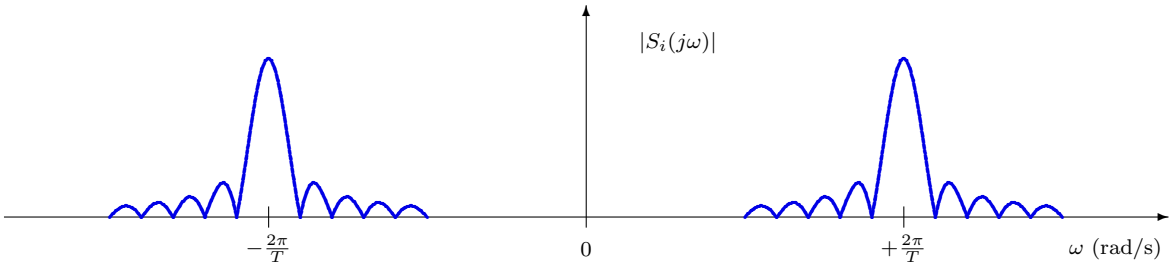


Figure 3.90: Modulated signal for the information sequence transmitted in Example D.

symbols, each symbol again carries $m = \log_2(M) = 2$ bits of information. What changes is the waveform that each pair of bits carries, as the orthonormal basis that defines the modulator has changed.

If we compare the systems of the last two examples, which share an encoder, the performance if it is transmitted over a Gaussian channel are identical, and the energies of the four symbols are also identical, since these are determined by the choice of the encoder (which contains the discrete representation of the signals). What changes from the third example to the fourth is the waveform of the signals and therefore their frequency response. In the fourth example we have square signals, whose frequency response is baseband (centered at 0 Hz), while in the third example we have signals whose frequency response is centered at the frequency of the sinusoids that form the basis, so which is a bandpass response, as shown in Fig. 3.91. The choice of one orthornormal basis (modulator) and another will depend on the characteristics of the channel through which we want to transmit, and in particular in this case, on the band in which the transmission is to be carried out. If the channel has a low pass response, it will be better to use the modulator from the fourth example, which generates adequate signals for this type of channel. If the channel has a bandpass response, it will be better to use the modulator of the third example, which generates signals more suitable for this type of response.

- Example C



- Example D

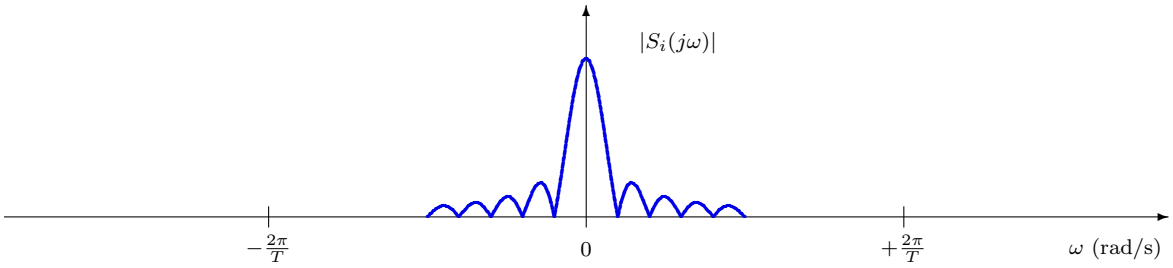


Figure 3.91: Frequency response of the signals of examples C and D.

Chapter 4

Fundamental limits in digital communications systems

The main objective of a communications system is the reliable transmission of information. A source produces the information, and the purpose of the communication system is to transmit the output of the source to the destination of the information, as illustrated in Figure 4.1.

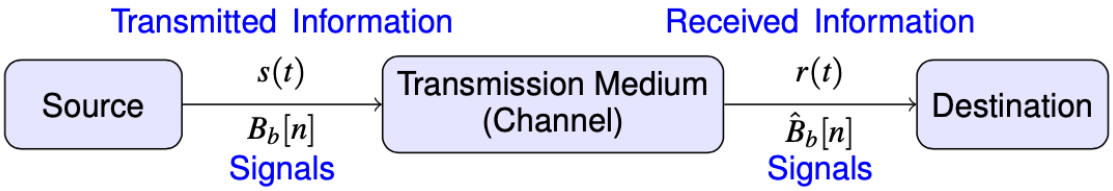


Figure 4.1: Simplified functional diagram of a communications system.

There is a wide variety of information sources, and each of them produces information of a different nature. Some types of information sources, with the corresponding type of information they generate, could be, for example:

- Radio broadcasting: voice or audio source.
- TV broadcasting: video source.
- FAX transmission: still image.
- Communication between PCs: binary or ASCII source.
- Data storage: binary source.

When analyzing a communications system and measuring its performance or its limits, it is possible to consider quantifying the amount of transmitted information, or the reliability in the transmission of information.

Two types of communication systems have been studied in previous chapters: analog communication systems and digital communication systems. Regarding analog communications systems, the compromise between performance and resource consumption of the different modulation variants has already been studied, especially power and bandwidth. In this chapter, only digital

communications systems will be considered, and the objective will be to study the fundamental limits that can be reached in the reliable transmission of information with this type of systems. It should be remembered that the use of a digital communications system does not imply the exclusion of analog sources. Digital transmission generally allows greater immunity against noise, greater flexibility, the application of encryption and makes easier the implementation of equipment, which has made digital systems predominate over analog ones. But it is possible to transmit information of an analog nature through a digital system, performing an analog-to-digital conversion on the transmitter side and the corresponding digital-to-analog conversion on the receiver side.

When studying the performance of digital communications systems in the previous chapter, it has been seen that by increasing the transmission rate by increasing the number of bits per symbol of the constellation, for a certain average energy per fixed symbol, the probability of error increases and therefore the performance is degraded. This fact was considered unavoidable until Claude Shannon demonstrated in the 1940s that it is possible to transmit with as low a probability of error as desired at an arbitrary bit rate as long as that bit rate is below the so-called *channel capacity*. This demonstration is considered the beginning of the so-called *information theory*, and sets a limit to communication systems as to the maximum amount of information that can be reliably transmitted. It should be noted that the proof of this result establishes the limit, but does not specify how the limit can be reached. At the moment, information theory has no answer to this question.

This chapter intends to analyze what is the maximum amount of information that can be transmitted reliably using a digital communications system, and the factors that this limit depends on. Although almost everyone has an intuitive notion of what information is, to carry out the analysis of a communications system, this intuitive notion is not enough, rather it is necessary to have quantitative measures of information. Hartley, Nyquist, or Shannon were pioneers in developing definitions of measures for information that are useful for the purpose of this chapter. These measures of information are related to the probability distributions of the elements whose information is to be quantified. In order to make use of them in the analysis of a digital communications system, it is necessary to first define probabilistic models that can be used to represent the behavior of digital information sources. Since the information will be transmitted through a channel using the communications system, it will also be necessary to have probabilistic models that allow representing the behavior of the channel and the behavior of the communications system at different levels. From these probabilistic models, and using quantitative measures of information, it will finally be possible to obtain the limits that a certain communications system can reach.

This chapter is divided into four parts:

- In the first part, probabilistic models will be presented to represent the behavior of information sources.
- Next, probabilistic models will be defined to represent the behavior of the communications channel and the elements of the communications system at different levels. All these models will be referred to generically as channel models, since the different elements of the communications system can be considered in some way as part of the channel through which the information is transmitted.
- The third part will present different quantitative measures of information.
- Finally, in the last part, making use of the probabilistic models for sources and channels, and of the quantitative measures of information presented in the third part, the fundamental

limits achievable by a digital communication system will be obtained.

4.1 Modeling of information sources

A source of information produces something as output, which will be generically called information, which is of interest to a potential receiver, who does not know it in advance. The mission of the communications system is to ensure that the information is transmitted correctly.

As the output of the information source is a time varying function that is unpredictable (if it is predictable, there is not much interest in transmitting it). This output can be modeled by means of a random process. The characteristics of this random process will depend on the characteristics and nature of the specific source. For example, it could be a continuous time or discrete time random process depending on whether the source generates information of analog or digital nature, as shown in Figure 4.2.

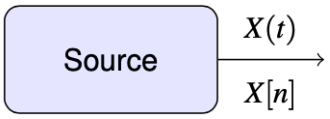


Figure 4.2: Statistical model of an information source: a random process.

Although this chapter will focus on the study of digital systems, and the most relevant models will therefore be the models for digital sources (analog sources will be converted to digital format before transmission), the models used for analog sources will also be briefly studied for completeness.

4.1.1 Analog sources

The model used to characterize an analog source is usually continuous time random process, $X(t)$, whose statistical properties depend on the nature of the source. The modeling of a voice source can be taken as an example. It is known that the voice signal has most of its power essentially distributed in the frequency band between 300 and 4000 Hz. Therefore, this source can be modeled by a random process whose power spectral density adjusts to these characteristics, as illustrated in Figure 4.3.

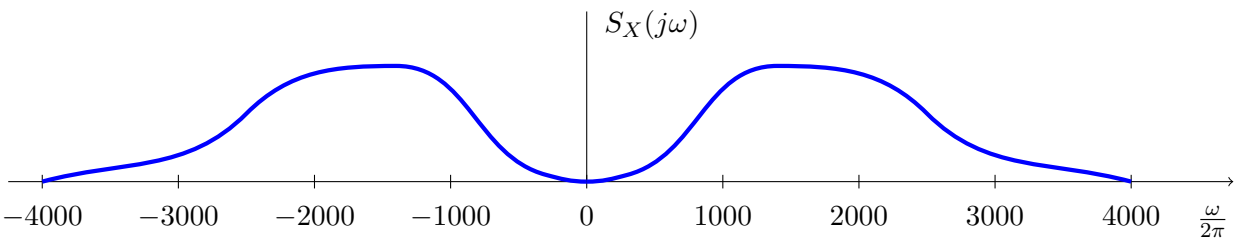


Figure 4.3: Example of a power spectral density that represents the typical frequency components of a voice signal.

It is usually considered that the mean behavior does not change over time, which allows us to assume that the process is stationary. As the mean of the signal is also zero, the random process used to model the voice signal could be a stationary random process, with zero mean, and a power spectral density function like the one in the figure. Being a stationary process, its autocorrelation function will be given by the inverse Fourier transform of this power spectral density.

The same procedure is applicable to different analog sources. For example, for TV signals, depending on the system (PAL, SECAM or NTSC), the band is between 0-6.5 MHz or between 0-4.5 MHz. A stationary random process would be used to model them, whose power spectral density represents the mean behavior of the frequency response of the squared signal.

Although each analog signal will have different spectral characteristics, there are some common aspects in most of the models:

1. Band-limited processes are considered.
2. This allows them to be sampled following the Nyquist criteria and can later be reconstructed.

4.1.2 Digital sources

In the case of a digital source, its output can be modeled by a discrete time random process. The source is modeled as a discrete time random process, $X[n]$. The source alphabet can be:

- Discrete: For example to model a digital data source or sampled and quantized analog signals.
- Continuous: For example to represent sampled analog sources (such as a speech signal) before quantization.

Depending on the type of source, the statistical parameters of the random process will be different. In this section we are going to study the simplest source model, which allows us to carry out all the subsequent development of the chapter. This model is the *discrete memoryless model*.

Discrete memoryless source model

The discrete memoryless source model, or DMS, of an information source is a discrete time random process. All the random variables that make up the random process $X[n]$ are independent and have the same distribution (i.i.d.: independent and identically distributed). Thus, a DMS source generates a sequence of i.i.d. random variables which take values from a discrete alphabet. To fully describe this type of font, it is enough to know its alphabet and its distribution, that is,

1. $\mathcal{A}_X = \{x_0, x_1, \cdots, x_{M-1}\}$.
2. $p_X(x_i) = P(X = x_i)$ for $i = 0, 1, \cdots, M-1$.

An example of a DMS is provided below.

Example

A binary information source uses as a model a DMS that is described by the alphabet

$$\mathcal{A}_X = \{0, 1\},$$

and the following probabilities for the binary symbols

$$P(X = 1) = p \text{ and } P(X = 0) = 1 - p.$$

In the particular case where $p = 1/2$, this type of source is called *binary symmetric source* or BSS.

4.2 Probabilistic channel models

In Chapter 3 the general model of a digital communication system has been presented, whose functional elements are shown in Figure 4.4.

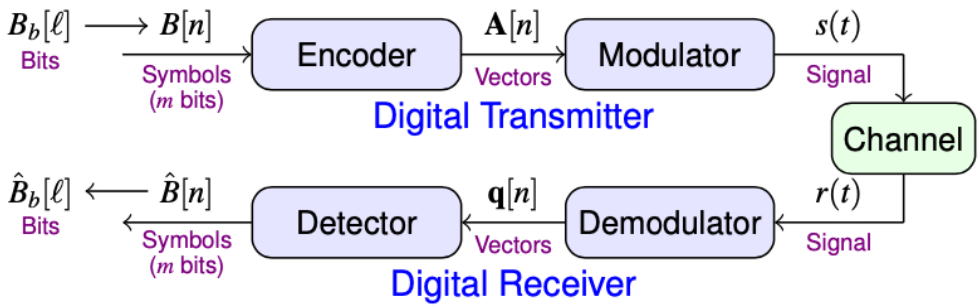


Figure 4.4: General model of a digital communication system.

In this scheme, the abstraction of the physical transmission medium was called channel, and a Gaussian additive channel model was used as a simplified model to represent it. In this section the term channel will have a broader meaning than the one considered in the diagram of Figure 4.4. Several probabilistic models will be defined, which will be generically called channel models. These models will establish the probabilistic relationship between the received information and the transmitted information at different levels of the communication system, which can be understood as the definition of various channels on the system, each one at a different level of abstraction.

The characterization of these models will be given by the conditional distribution of the output given the input. In all cases, conditional independence will be considered between inputs and outputs for different time instants, so that the time dependency can be eliminated. This means that the input value at a certain instant will be modeled by a random variable X, and the corresponding output value at the same instant will be modeled by another random variable, Y. In this case, the probabilistic model will be characterized by the conditional distribution

$$f_{Y|X}(y|x).$$

The difference between the models that are going to be presented is in the definition of what in each case is considered input and output for each *channel*. Four different probabilistic models (or *channels*) will be presented, each one considering what the channel is with a different level of abstraction on the general model of the communications system. Specifically, the four models are those shown in Figure 4.5, and which are listed below:

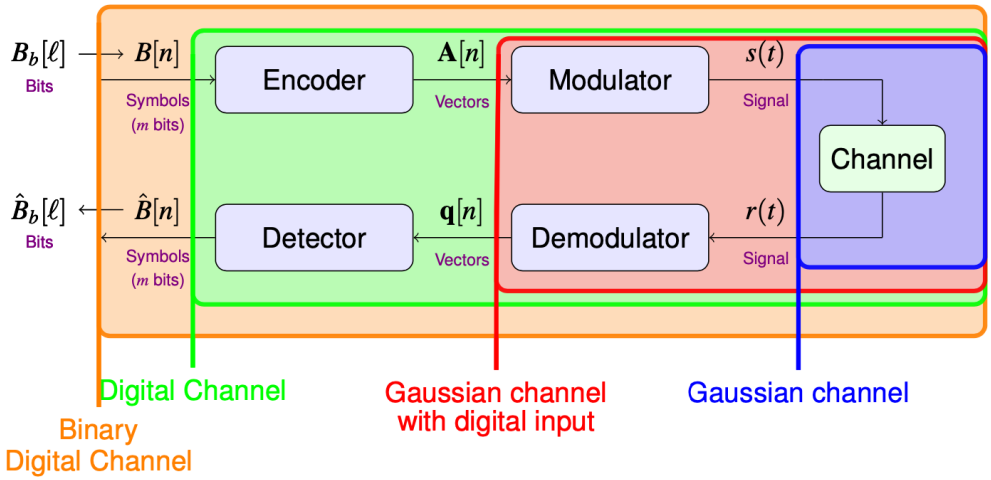


Figure 4.5: Definition of the different channel models on a digital communications system.

1. **Gaussian channel**. It is the model for representing the physical channel itself, which transforms the transmitted signal by adding additive noise, modeled by a stationary, ergodic, white, Gaussian random process, with zero mean and power spectral density $N_0/2$, so that

$$r(t) = s(t) + n(t).$$

In this model, the values of the transmitted signal and the received signal, at the same time instant, are considered as the input and the output, respectively

$$X \equiv s(t),\ Y \equiv r(t).$$

In both cases, the random variables are continuous random variables, since the transmitted and received signals can take on values over a continuous range of amplitudes.

2. **Gaussian channel with digital input**. This model coincides with what in the previous chapter was called equivalent discrete channel, which presents as input a sequence of symbols from a discrete alphabet (constellation) of M symbols. As output, it has an observation with a continuous domain, the output of the demodulator. In the previous chapter it was seen that the relationship between both was an additive relationship with the noise term at the demodulator output.

$$\mathbf{q}[n] = \mathbf{A}[n] + \mathbf{z}[n].$$

The statistical characteristics of the noise vector $\mathbf{z}[n]$ were studied in the previous chapter. Therefore, in this model the input is the transmitted symbol at a discrete instant n, and the output is the observation at the output of the demodulator at that same instant. Therefore, they are symbols and vector observations of dimension N, so multidimensional random variables (vectors) of the same dimension will be used to represent them.

$$\mathbf{X} \equiv \mathbf{A}[n],\ \mathbf{Y} \equiv \mathbf{q}[n].$$

3. **Digital channel**. This model considers the set formed by the encoder, modulator, channel, demodulator and detector. The input is the transmitted symbol, from an M-ary alphabet, and the output is the estimated symbol, which has the same alphabet: $B[n]$ and $\hat{B}[n]$ (or $\mathbf{A}[n]$ and $\hat{\mathbf{A}}[n]$, since there is a one-to-one equivalence between symbols as blocks of M bits and N-dimensional vectors of the constellation that carry them). Therefore, in this model the input and output will be defined as

$$X \equiv B[n],\ Y \equiv \hat{B}[n] \text{ or equivalently } X \equiv \mathbf{A}[n],\ Y \equiv \hat{\mathbf{A}}[n].$$

4. **Binary digital channel**. This is the last level of abstraction, in which the entire communications system is considered as a channel, whose inputs and outputs are binary symbols. It represents the highest level of abstraction of a communications system: the complete system seen as a vehicle or channel for the transmission of bits. In this case, therefore, the input and output will be defined as the transmitted and received bit, at the same time instant

$$X \equiv B_b[\ell],\ Y \equiv \hat{B}_b[\ell].$$

Once these four channels have been presented, the probabilistic model (conditional distribution of the output given the input) that will be used to represent each one of them will be obtained.

4.2.1 Gaussian channel

The input-output relationship of the Gaussian channel is

$$r(t) = s(t) + n(t),$$

where the noise term $n(t)$ is modeled as a stationary, ergodic, white, Gaussian, random process with zero mean and power spectral density $S_n(j\omega) = N_0/2$. The autocorrelation function of the noise is therefore

$$R_n(\tau) = \frac{N_0}{2}\,\delta(\tau).$$

To obtain the probabilistic model given by the conditional probability of the output given the input, the first thing to take into account is that the power of the noise process, being white, is strictly infinite. This implies that in practice, to minimize the effect of noise on the receiver, a frequency selective filter must be used. Ideally, this filter will not introduce any distortion into the signal $s(t)$, which will be considered band limited, with bandwidth B Hz, and at the same time should minimize the noise power at its output. The filter that meets these conditions is an ideal low-pass filter, with a bandwidth equal to that of the received signal. In this way the Gaussian channel will actually model the relationship between the transmitted signal and the received signal after this filtering, as illustrated in Figure 4.6, where the response $h_n(t)$ (or its frequency equivalent $H_n(j\omega)$) is that of the ideal filter used to limit the noise power.

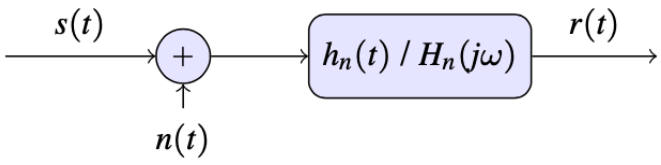


Figure 4.6: Gaussian channel with receiver filtering to limit noise power.

If the signal has a bandwidth of B Hz, the filter's bandwidth will be B Hz, so its frequency response is

$$H_n(j\omega) = \Pi\left(\frac{\omega}{2W}\right),$$

where $W = 2\pi B$ denotes the bandwidth in radians per second, and its impulse response

$$h_n(t) = 2B\ \mathrm{sinc}(2Bt).$$

The power of the noise term at the output of this filter, as seen in Chapter 2, is obtained by integrating the power spectral density of the filtered noise process, which for ideal filters involves multiplying N_0 by the bandwidth in Hz of the filter

$$\sigma^2 = \int_{-\infty}^{\infty} S_n(j\omega)|H_n(j\omega)|^2 \, d\omega = \frac{N_0}{2} \times 2B = N_0 B.$$

Taking this into account, the *Gaussian probabilistic channel* is defined as the one that relates two random variables X and Y with continuous probability density functions on $\mathbb{R}$ that represent the value of $s(t)$ and $r(t)$ at a certain instant of time. Given that the value of the output at an instant will be that of the input plus the value of the filtered noise at that instant, and that this noise has a Gaussian distribution with power N_0B, this probabilistic model is characterized by the conditional probability density function

$$f_{Y|X}(y|x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(y-x)^2}{2\sigma^2}},$$

where $\sigma^2 = N_0 B$.

4.2.2 Gaussian channel with digital input

Figure 4.7 conceptually illustrates the Gaussian channel with digital input, which models the relationship between the constellation symbols being transmitted and the observation at the demodulator output at a given time.

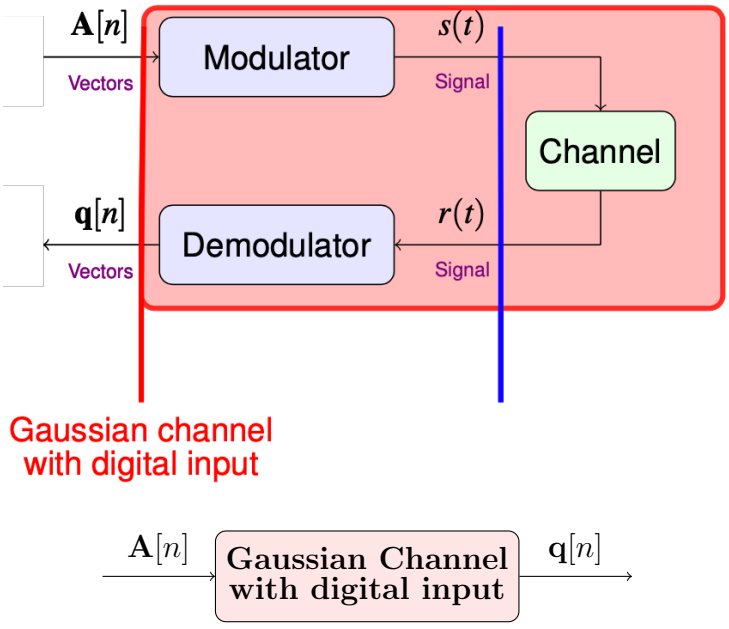


Figure 4.7: Gaussian channel with digital input.

As seen in Chapter 3, the observation $\mathbf{q}[n]$ (the output of this channel) cab be written as

$$\mathbf{q}[n] = \mathbf{A}[n] + \mathbf{z}[n],$$

where $\mathbf{z}[n]$ is the noise component of the observation noise, which is a discrete-time, multidimensional stochastic process (of the dimension of the system signal space, N), made up of N jointly

Gaussian random variables, and independent of $\mathbf{A}[n]$. In addition, its components are in turn statistically independent from each other. This makes it possible to remove the time dependency and use the representation

$$\mathbf{q} = \mathbf{A} + \mathbf{z},$$

where $\mathbf{z}$ has a Gaussian probability density function, N-dimensional, with zero mean and variance $N_0/2$ in all directions of space

$$f_{\mathbf{z}}(\mathbf{z}) = \mathcal{N}^N\left(\mathbf{0}, \frac{N_0}{2}\right) = \frac{1}{(\pi N_0)^{N/2}} e^{-\frac{||\mathbf{z}||^2}{N_0}}.$$

From here, obtaining the probabilistic model is simple. It is formally defined as the one that relates the N-dimensional random variables $\mathbf{X}$ and $\mathbf{Y}$, the first with a discrete alphabet $\{\mathbf{x}_i\}$, with $i = 0, \cdots, M-1$, where each value of the alphabet will be identified with one of the N-dimensional vectors that form the constellation of the system, and the second with an N-dimensional continuous probability density function over $\mathbb{R}$. Under this premise, the probabilistic model is given by the conditional probability density function

$$f_{\mathbf{Y}|\mathbf{X}}(\mathbf{y}|\mathbf{x}_i) = \frac{1}{(\pi N_0)^{N/2}} e^{-\frac{||\mathbf{y}-\mathbf{x}_i||^2}{N_0}},$$

that is, an N-dimensional jointly Gaussian distribution with mean equal to the transmitted symbol and variance $\sigma^2 = N_0/2$ in each direction of space. As seen, it agrees with the probabilistic model that defines the equivalent discrete channel, now using a slightly different notation in terms of the random variables $\mathbf{X}$ and $\mathbf{Y}$ to denote, respectively, $\mathbf{A}$ and $\mathbf{q}$.

4.2.3 Digital channel

Figure 4.8 represents the digital channel. Usually, we work with the vectors of the constellation $\mathbf{A}[n]$ instead of with the symbols $B[n]$, although given the one-to-one relationship, it can be applied to the latter in the same way.

In this model, each symbol in the sequence $B[n]$ is considered to be statistically independent of the rest of the symbols in the sequence. On the other hand, it has already been seen that given that the signals only occupy the intervals of duration T dedicated to a symbol and that the noise is independent at each moment, the reception of each symbol is independent of the rest of the symbols. Under these conditions, the probability of having a given symbol at the output of the digital channel, $\hat{B}[n]$, depends solely on the symbol $B[n]$ that is transmitted at that same instant. Therefore, it is possible to eliminate the time dependency and only analyze the case of the transmission of an isolated symbol, understanding that every time the channel is used to transmit a symbol, the channel will not modify its behavior.

When the symbol b_i is transmitted, at the output of the channel symbol b_j is decided with a specific probability: this is the conditional probability

$$p_{\hat{B}|B}(b_j|b_i).$$

Due to the one-to-one assignment between a symbol and the vector representation of the signal that transmits it, this probability satisfies that

$$p_{\hat{B}|B}(b_j|b_i) = p_{\hat{\mathbf{A}}|\mathbf{A}}(\mathbf{a}_j|\mathbf{a}_i).$$

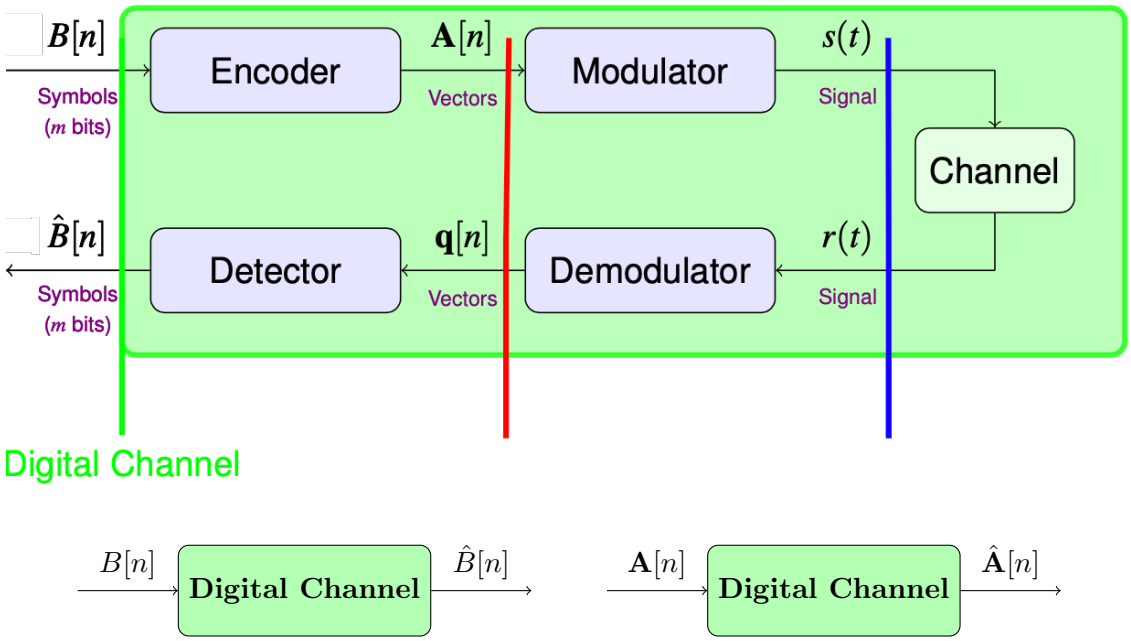


Figure 4.8: Conceptual representation of the digital channel.

In the previous chapter, we studied how these probabilities are calculated by integrating the conditional distribution of the observation for the transmitted symbol in the corresponding decision region I_j. Since in this model the input and output were $X \equiv B$ and $Y \equiv \hat{B}$, if these probabilities are known for all possible combinations of transmitted and received symbols, the channel will be fully characterized.

There is a widely used probabilistic model that contemplates the digital channel as a particular case and is called *discrete memoryless channel* or DMC.The DMC is a statistical model that relates a random variable X with a discrete probability density function that we call input and another random variable Y with a discrete probability function that we call output. In the particular case of the DMC, the input and output alphabets may be different. In our application they share the same alphabet.

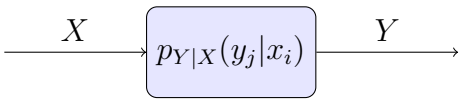


Figure 4.9: Probabilistic model: discrete memoryless channel (DMC).

The DMC can be represented by the diagram in Figure 4.9. In the block that represents the channel, neither impulse responses nor frequency responses appear, but rather the conditional probabilities of the output given the input. The *discrete* term comes from the nature of X and Y, which have a discrete alphabet. The term *memoryless* comes from the probabilistic model of the input and output, random variables instead of random processes (there is no time dependence of the statistics). Formally, a discrete memoryless channel is defined through the following elements:

1. The *input alphabet* (of M_X possible values)

$$\mathcal{A}_X = \{x_i \mid i = 0, \cdots, M_X - 1\}.$$

2. The *output alphabet* (of M_Y possible values)

$$\mathcal{A}_X = \{y_i \mid i = 0, \cdots, M_Y - 1\}.$$

3. The set of conditional probabilities

$$p_{Y|X}(y_j|x_i).$$

These probabilities are called *transition probabilities*, and they are usually grouped in the so-called *channel matrix*, which is a matrix of M_X rows and M_Y columns that arranges the transition probabilities as follows

$$\mathbf{P} = \begin{bmatrix} p_{Y|X}(y_0|x_0) & p_{Y|X}(y_1|x_0) & \cdots & p_{Y|X}(y_{M_Y-1}|x_0) \\ p_{Y|X}(y_0|x_1) & p_{Y|X}(y_1|x_1) & \cdots & p_{Y|X}(y_{M_Y-1}|x_1) \\ \vdots & \vdots & \ddots & \vdots \\ p_{Y|X}(y_0|x_{M_X-1}) & p_{Y|X}(y_1|x_{M_X-1}) & \cdots & p_{Y|X}(y_{M_Y-1}|x_{M_X-1}) \end{bmatrix}$$

In this matrix, a row is associated with a certain input, while a column is associated with a certain output. Therefore, the sum of the elements of a row results in 1 (we have the sum of the conditional probability distribution over the entire observation space). Furthermore, if two DMCs are concatenated, the channel matrix of the concatenation will be obtained by the product of the channel matrices of each of the channels:

$$\mathbf{P}Z|X = \mathbf{P}Y|X \times \mathbf{P}Z|Y.$$

Sometimes, instead of the channel matrix, a graphical representation is used to specify the transition probabilities, using an arrow diagram or trellis diagram, such as the one shown in Figure 4.10. In this case, the transition probabilities are included in the weights associated with the different arrows that form the diagram joining elements of the input with the output. Given the definition of these probabilities, the probabilities associated with the arrows coming out of the same node add up to unity.

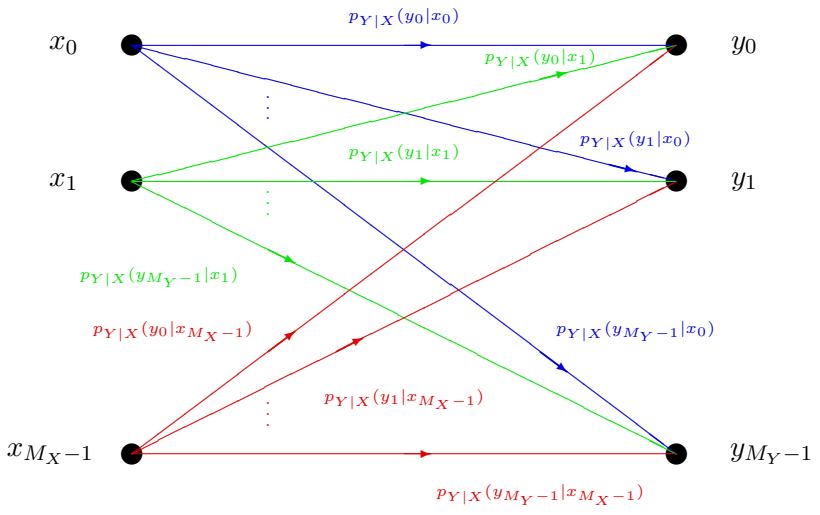


Figure 4.10: Representation of a DMC using an arrow diagram.

It should be noted that in the definition of a discrete memoryless channel, the input and output alphabets appear, but not the probability distributions of the input, $p_X(x_i)$, nor of the output, $p_Y(y_j)$, since these probabilities are not part of the nature of the channel.

The following shows how to obtain a DMC that represents the digital channel for a certain communication system. The model is obtained from the symbol error probabilities defined in the previous chapter. In the first place, for a DMC that models a communications system, both the input and output alphabets correspond to the alphabet of symbols of the system, either $\mathbf{A}[n]$ or $B[n]$, given their equivalence. For convenience we will use the vector representation of the symbols, that is

$$x_i \equiv \mathbf{a}_i,$$

$$y_j \equiv \mathbf{a}_j,$$

and now $M_X = M_Y = M$. In this way, the association of the alphabet with the symbols of the constellation is implicit in the subscripts. Regarding the transition probabilities, it is evident that

$$p_{Y|X}(y_j|x_i) \equiv p_{\hat{\mathbf{A}}|\mathbf{A}}(\mathbf{a}_j|\mathbf{a}_i).$$

That is, the transition probability $p_{Y|X}(y_j|x_i)$ indicates the probability of receiving the symbol $\mathbf{a}_j$ when the symbol $\mathbf{a}_i$ has been transmitted. In this case, the elements of the main diagonal of the channel matrix, for which $j = i$, correspond to the conditional accuracies (probabilities of a correct decision) for each symbol

$$p_{Y|X}(y_i|x_i) = p_{\hat{\mathbf{A}}|\mathbf{A}}(\mathbf{a}_i|\mathbf{a}_i) = P_{a|\mathbf{a}_i} = 1 - P_{e|\mathbf{a}_i}.$$

The elements out of the diagonal, on the other hand, correspond to the error probabilities between different symbols.

$$p_{Y|X}(y_j|x_i) = p_{\hat{\mathbf{A}}|\mathbf{A}}(\mathbf{a}_j|\mathbf{a}_i) = P_{e|\mathbf{a}_i \to \mathbf{a}_j}.$$
$$\underset{j\neq i}{} \qquad \underset{j\neq i}{}$$

Consequently, the sum of the elements of each row outside the main diagonal is equal to the conditional error probability for the symbol associated with that row.

$$\sum_{\substack{j=0\\ j\neq i}}^{M-1} p_{Y|X}(y_j|x_i) = \sum_{\substack{j=0\\ j\neq i}}^{M-1} P_{e|\mathbf{a}_i \to \mathbf{a}_j} = P_{e|a_i}.$$

This means that in an ideal system, the channel matrix or trellis diagram should be as shown in Figure 4.11: a diagonal matrix, or a trellis diagram with a single arrow from each input symbol to the same output symbol.

$$\mathbf{P} = \mathbf{I} = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & 1 & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix}$$

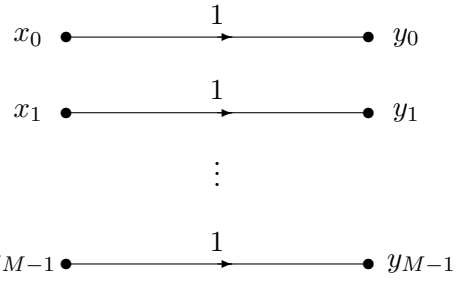


Figure 4.11: Ideal values of a DMC that models a Gaussian channel.

To illustrate the procedure for obtaining the channel matrix for a communications system, a system with a constellation of four symbols, $M = 4$, in a one-dimensional space, $N = 1$, will be used as an example, with coordinates $\mathbf{a}_0 = -3,\ \mathbf{a}_1 = -1,\ \mathbf{a}_2 = +1,\ \mathbf{a}_3 = +3$ and equiprobable, with which the decision regions are given by the thresholds $q_{u1} = -2,\ q_{u2} = 0,\ q_{u3} = +2$

$$I_0 = (-\infty, -2],\ I_1 = (-2, 0],\ I_2 = (0, +2],\ I_3 = (+2, +\infty)$$

as shown in Figure 4.12.

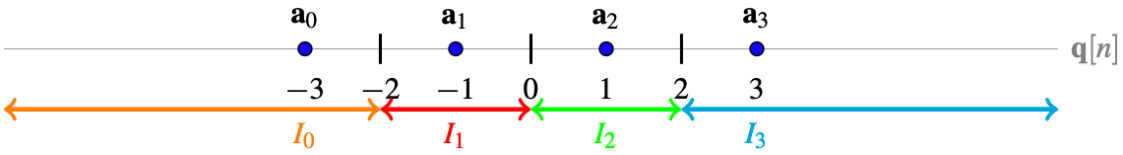


Figure 4.12: One-dimensional constellation of four equiprobable symbols and their corresponding decision regions.

In this case we have the following association

$$X \equiv \mathbf{A}[n],\ Y \equiv \hat{\mathbf{A}}[n],$$

so the input and output alphabets match

$$x_i \equiv \mathbf{a}_i,\ y_j \equiv \mathbf{a}_j \text{ for } i, j \in \{0, 1, \cdots, M-1\}.$$

The transition probabilities $p_{Y|X}(y_j|x_i)$ probabilistically defining the system correspond in this case to the probability of receiving the index symbol j when the index symbol i has been transmitted. These values define the probability of succeeding in the transmission of a symbol, if $j = i$, or the error probability between two symbols, if $j \neq i$. Using the notation from the previous chapter, we have

$$p_{Y|X}(y_i|x_i) = P_{a|\mathbf{a}_i} = 1 - P_{e|\mathbf{a}_i},$$

and

$$p_{Y|X}(y_j|x_i) = P_{e|\mathbf{a}_i \to \mathbf{a}_j} \text{ for } j \neq i.$$

In the previous chapter it was explained how these values were obtained, and will be obtained again for this constellation.

We start with the transition probabilities appearing in the first row of the channel matrix, $p_{Y|X}(y_j|x_0)$, $\forall j$. In this case, they are the probabilities of receiving each of the 4 symbols when $\mathbf{a}_0$ (the symbol associated with x_0) is transmitted. The distribution of the observation when the symbol $\mathbf{a}_0$ is transmitted is Gaussian, with mean $\mathbf{a}_0$ and variance $N_0/2$. So to get the conditional probabilities, you just have to integrate that Gaussian distribution in each of the 4 decision regions, as illustrated in Figure 4.13.

These probabilities are then calculated:

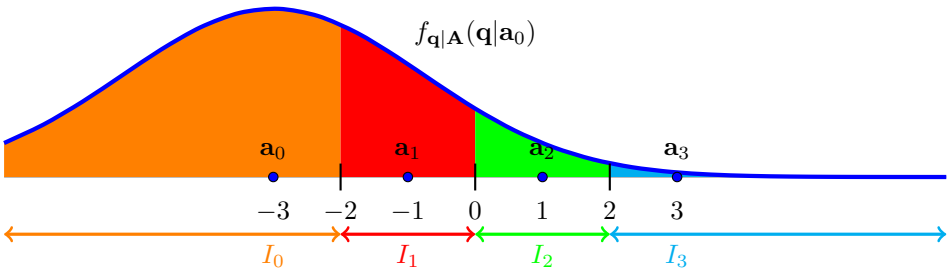


Figure 4.13: Calculation of the transition probabilities associated with the first row of the channel matrix.

- Distribution $f_{\mathbf{q}|\mathbf{A}}(\mathbf{q}|\mathbf{a}_0)$: Gaussian, mean $\mathbf{a}_0 = -3$ and variance $N_0/2$

$$
\begin{aligned}
p_{Y|X}(y_0|x_0) =& 1 - P_{e|\mathbf{a}_0} = 1 - Q\left(\frac{1}{\sqrt{N_0/2}}\right) \\
p_{Y|X}(y_1|x_0) =& P_{e|\mathbf{a}_0 \to \mathbf{a}_1} = Q\left(\frac{1}{\sqrt{N_0/2}}\right) - Q\left(\frac{3}{\sqrt{N_0/2}}\right) \\
p_{Y|X}(y_2|x_0) =& P_{e|\mathbf{a}_0 \to \mathbf{a}_2} = Q\left(\frac{3}{\sqrt{N_0/2}}\right) - Q\left(\frac{5}{\sqrt{N_0/2}}\right) \\
p_{Y|X}(y_3|x_0) =& P_{e|\mathbf{a}_0 \to \mathbf{a}_3} = Q\left(\frac{5}{\sqrt{N_0/2}}\right)
\end{aligned}
$$

It can be seen that these four probabilities add up to unity, as expected.

Next, the transition probabilities in the second row of the channel matrix, $p_{Y|X}(y_j|x_1)$, $\forall j$. In this case there are the probabilities of receiving each of the 4 symbols when $\mathbf{a}_1$ (the symbol associated with x_1) is transmitted. The distribution of the observation given that $\mathbf{a}_1$ is transmitted is Gaussian, with mean $\mathbf{a}_1$ and variance $N_0/2$. The transition probabilities are the integrals in the 4 decision regions, as shown in Figure 4.14.

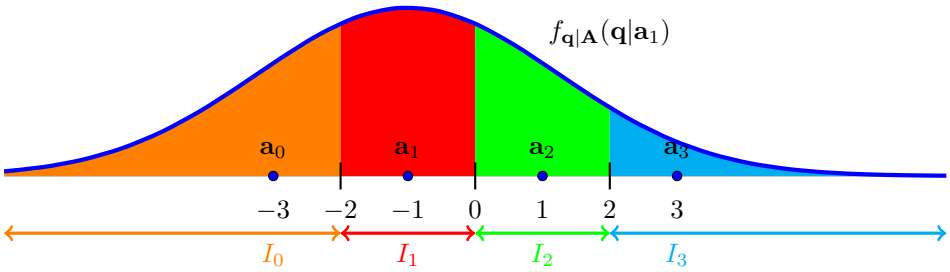


Figure 4.14: Calculation of the transition probabilities associated to the first second of the channel matrix.

These probabilities are then obtained:

- Distribuiton $f_{\mathbf{q}|\mathbf{A}}(\mathbf{q}|\mathbf{a}_1)$: Gaussian, mean $\mathbf{a}_1$ and variance $N_0/2$

$$
\begin{aligned}
p_{Y|X}(y_0|x_1) =& P_{e|\mathbf{a}_1\to\mathbf{a}_0} = Q\left(\frac{1}{\sqrt{N_0/2}}\right)\\
p_{Y|X}(y_1|x_1) =& 1 - P_{e|\mathbf{a}_1} = 1 - 2Q\left(\frac{1}{\sqrt{N_0/2}}\right)\\
p_{Y|X}(y_2|x_1) =& P_{e|\mathbf{a}_1\to\mathbf{a}_2} = Q\left(\frac{1}{\sqrt{N_0/2}}\right) - Q\left(\frac{3}{\sqrt{N_0/2}}\right)\\
p_{Y|X}(y_3|x_1) =& P_{e|\mathbf{a}_1\to\mathbf{a}_3} = Q\left(\frac{3}{\sqrt{N_0/2}}\right)
\end{aligned}
$$

In the third row of the channel matrix we have the transition probabilities $p_{Y|X}(y_j|x_2)$, $\forall j$. The distribution of the observation given that $\mathbf{a}_2$ is transmitted is in this case Gaussian with mean $\mathbf{a}_2 = +1$ and variance $N_0/2$. As in the previous cases, to obtain the conditional probabilities, this Gaussian distribution must be integrated in each of the 4 decision regions, as illustrated in Figure 4.15.

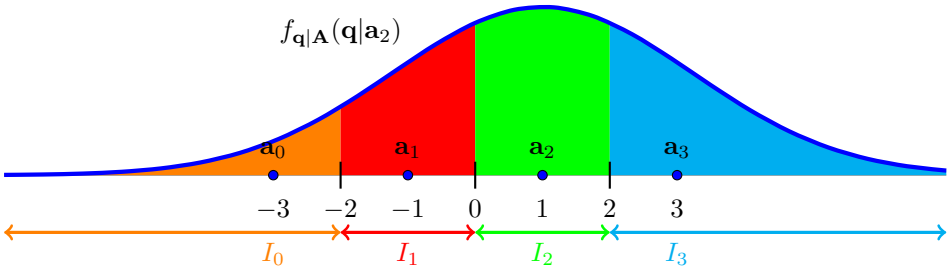


Figure 4.15: Calculation of the transition probabilities associated with the third row of the channel matrix.

Now:

- Distribution $f_{\mathbf{q}|\mathbf{A}}(\mathbf{q}|\mathbf{a}_2)$: Gaussian, mean $\mathbf{a}_2$ and variance $N_0/2$

$$
\begin{aligned}
p_{Y|X}(y_0|x_2) =& P_{e|\mathbf{a}_2\to\mathbf{a}_0} = Q\left(\frac{3}{\sqrt{N_0/2}}\right)\\
p_{Y|X}(y_1|x_2) =& P_{e|\mathbf{a}_2\to\mathbf{a}_1} = Q\left(\frac{1}{\sqrt{N_0/2}}\right) - Q\left(\frac{3}{\sqrt{N_0/2}}\right)\\
p_{Y|X}(y_2|x_2) =& 1 - P_{e|\mathbf{a}_2} = 1 - 2Q\left(\frac{1}{\sqrt{N_0/2}}\right)\\
p_{Y|X}(y_3|x_2) =& P_{e|\mathbf{a}_2\to\mathbf{a}_3} = Q\left(\frac{1}{\sqrt{N_0/2}}\right)
\end{aligned}
$$

Finally, in the fourth and last row of the channel matrix we have the transition probabilities $p_{Y|X}(y_j|x_2)$, $\forall j$. The conditional distribution of the observation when $\mathbf{a}_3$ is transmitted is in this case Gaussian with mean $\mathbf{a}_3 = +3$ and variance $N_0/2$. The transition probabilities are as represented graphically in Figure 4.16.

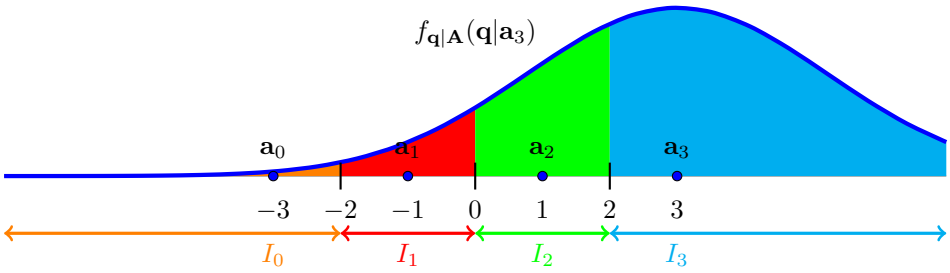


Figure 4.16: Calculation of the transition probabilities associated with the fourth row of the channel matrix.

These probabilities are:

- Distribution $f_{\mathbf{q}|\mathbf{A}}(\mathbf{q}|\mathbf{a}_3)$: Gaussian, mean $\mathbf{a}_3$ and variance $N_0/2$

$$
\begin{aligned}
p_{Y|X}(y_0|x_3) &= P_{e|\mathbf{a}_3\to\mathbf{a}_0} = Q\left(\frac{5}{\sqrt{N_0/2}}\right)\\
p_{Y|X}(y_1|x_3) &= P_{e|\mathbf{a}_3\to\mathbf{a}_1} = Q\left(\frac{3}{\sqrt{N_0/2}}\right) - Q\left(\frac{5}{\sqrt{N_0/2}}\right)\\
p_{Y|X}(y_2|x_3) &= P_{e|\mathbf{a}_3\to\mathbf{a}_2} = Q\left(\frac{1}{\sqrt{N_0/2}}\right) - Q\left(\frac{3}{\sqrt{N_0/2}}\right)\\
p_{Y|X}(y_3|x_3) &= 1 - P_{e|\mathbf{a}_3} = 1 - Q\left(\frac{1}{\sqrt{N_0/2}}\right)
\end{aligned}
$$

For this example, given the symmetry of the constellation, once the values of the transition probabilities for the first two rows are obtained, those of the next two rows can be obtained immediately. In any case, grouping all the transition probabilities, the DMC that represents the communication system that uses the 4-symbol constellation of the example has the following channel matrix:

$$
\mathbf{P} = \begin{bmatrix}
1 - Q\left(\frac{1}{\sqrt{N_o/2}}\right) & Q\left(\frac{1}{\sqrt{N_o/2}}\right) - Q\left(\frac{3}{\sqrt{N_o/2}}\right) & Q\left(\frac{3}{\sqrt{N_o/2}}\right) - Q\left(\frac{5}{\sqrt{N_o/2}}\right) & Q\left(\frac{5}{\sqrt{N_o/2}}\right) \\
Q\left(\frac{1}{\sqrt{N_o/2}}\right) & 1 - 2Q\left(\frac{1}{\sqrt{N_o/2}}\right) & Q\left(\frac{1}{\sqrt{N_o/2}}\right) - Q\left(\frac{3}{\sqrt{N_o/2}}\right) & Q\left(\frac{3}{\sqrt{N_o/2}}\right) \\
Q\left(\frac{3}{\sqrt{N_o/2}}\right) & Q\left(\frac{1}{\sqrt{N_o/2}}\right) - Q\left(\frac{3}{\sqrt{N_o/2}}\right) & 1 - 2Q\left(\frac{1}{\sqrt{N_o/2}}\right) & Q\left(\frac{1}{\sqrt{N_o/2}}\right) \\
Q\left(\frac{5}{\sqrt{N_o/2}}\right) & Q\left(\frac{3}{\sqrt{N_o/2}}\right) - Q\left(\frac{5}{\sqrt{N_o/2}}\right) & Q\left(\frac{1}{\sqrt{N_o/2}}\right) - Q\left(\frac{3}{\sqrt{N_o/2}}\right) & 1 - Q\left(\frac{1}{\sqrt{N_o/2}}\right)
\end{bmatrix}.
$$

For different values of N_0, the array will have different values. As N_0 decreases, the matrix approaches the identity matrix, which is the ideal channel matrix.

4.2.4 Binary digital channel

The binary digital channel is the probabilistic model that involves a greater abstraction by considering the entire communications system as a channel that transmits and receives bits.

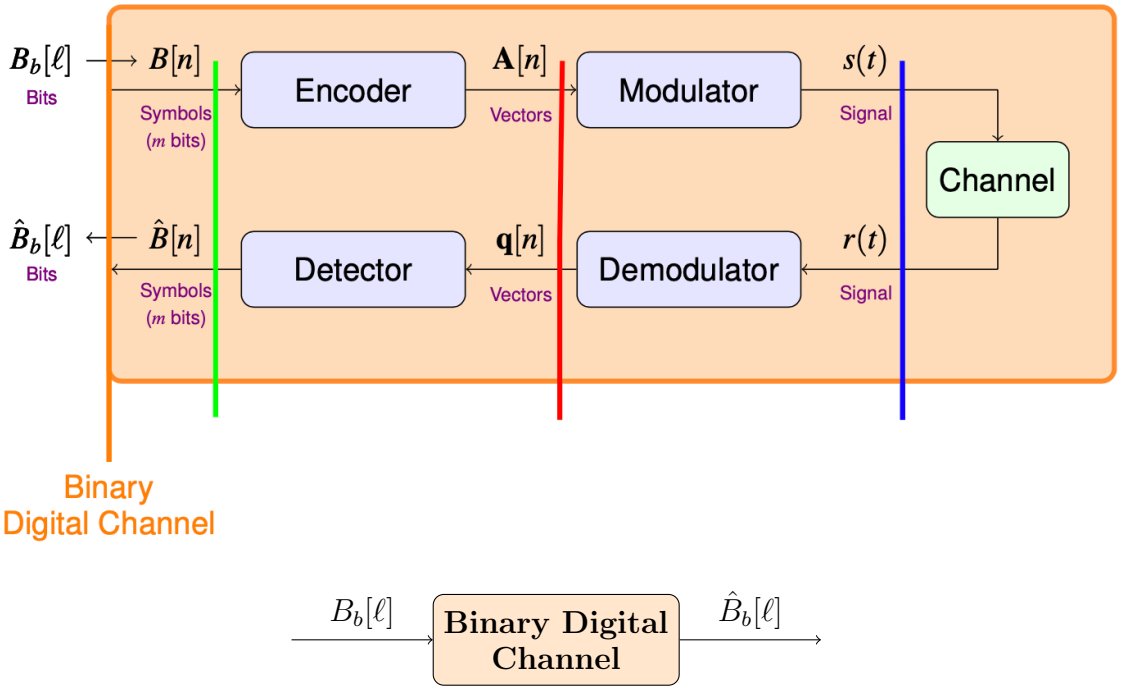


Figure 4.17: Conceptual representation of the binary digital channel.

It is therefore a model in which the probabilistic description is given by the probability of receiving each of the possible values of the binary sequence $\hat{B}_b[\ell]$ given the transmitted sequence $B_b[\ell]$. In the case in which time independence is assumed, $x_0 \equiv 0$, $x_1 \equiv 1$, $y_0 \equiv 0$ and $y_1 \equiv 1$ are used, and taking into account that by definition

$$p_{Y|X}(y_0|x_i) = 1 - p_{Y|X}(y_1|x_i),$$

the transition probabilities are reduced to only two relevant probabilities: the conditional probability of error (or, alternatively, the accuracy) for each bit. In this case, a DMC particularized for the case $M_X = M_Y = 2$ can be used as a model of the binary digital channel, which will have the form

$$\mathbf{P} = \begin{bmatrix} p_{Y|X}(y_0|x_0) & p_{Y|X}(y_1|x_0) \\ p_{Y|X}(y_0|x_1) & p_{Y|X}(y_1|x_1) \end{bmatrix} = \begin{bmatrix} 1-p_{e|0} & p_{e|0} \\ p_{e|1} & 1-p_{e|1} \end{bmatrix}.$$

The probabilities $p_{e|0}$ and $p_{e|1}$ denote the bit error probability when a zero or a one is transmitted, respectively. In most communication systems these two probabilities are equal

$$p_{e|0} = p_{e|1} = \varepsilon,$$

in which case the channel matrix is symmetric

$$\mathbf{P} = \begin{bmatrix} p_{Y|X}(y_0|x_0) & p_{Y|X}(y_1|x_0) \\ p_{Y|X}(y_0|x_1) & p_{Y|X}(y_1|x_1) \end{bmatrix} = \begin{bmatrix} 1-\varepsilon & \varepsilon \\ \varepsilon & 1-\varepsilon \end{bmatrix}.$$

This case is known as *Binary Symmetric Channel* or BSC.The trellis diagram representation for this model is shown in Figure 4.18.

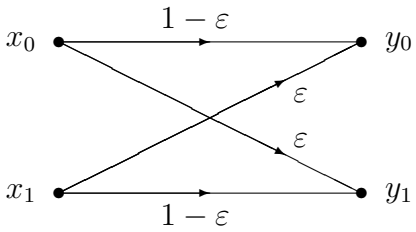


Figure 4.18: Trellis diagram representation of a binary symmetric channel or BSC.

The BSC is a probabilistic model that may be appropriate to represent the binary digital channel. However, before concluding the equivalence between binary digital channel and BSC, the following points should be discussed:

1. For binary systems, $M = 2$, a symbol of the sequence $B[n]$ carries a single bit, and if the symbols are transmitted independently in $B[n]$ that implies a bit-independent transmission in the sequence $B_b[\ell]$. In this case, the BSC model accurately represents the binary digital channel.

2. For M-ary systems with $M > 2$, the BSC represents the average behavior over time of the binary digital channel, because in the real system the transmission is carried out by blocks of $m = \log_2 M$ bits, symbol by symbol (sequence $B[n]$ and its vectorial representation $\mathbf{A}[n]$). It can therefore be said that while the digital channel does not have memory (because the transmission is carried out symbol-by-symbol, independently) and fits perfectly into the DMC model, the binary digital channel has the memory introduced by the encoder that transforms the sequence $B_b[\ell]$ into the sequence $B[n]$. This means that it cannot strictly be considered a channel without memory. From this point of view, the BSC is an approximation to the binary digital channel that represents its average behavior over time.

3. The value of the error probability ε defined for the BSC is in both cases is the bit error rate (BER) of the system.

Despite this discrepancy between the assumptions of the BSC model and the nature of the binary digital channel, the equivalence of both is usually accepted in practice, assigning to ε the BER of the system

$$\mathbf{P} = \begin{bmatrix} 1-BER & BER \\ BER & 1-BER \end{bmatrix}.$$

4.3 Quantitative measures of information

Once the probabilistic models have been established for sources and channels, in this section various types of quantitative measures of information are introduced. On the one hand, measures that can be applied to a random variable, which can be, for example, the one that models the input or output of a channel. On the other hand, measures that are applied simultaneously to two random variables, which in our problem would take into account the relationship between the input and output of a channel. Through these measurements it will be possible to subsequently calculate the maximum amount of information that can be reliably transmitted with a digital communications system.

4.3.1 Information and entropy

First of all, the information that a certain discrete random variable has is going to be quantified. In the application for the study of communication systems, this random variable can represent both the output of an information source and the input or output of a digital channel.

In order to obtain a quantitative measure of the information that a random variable contains, we will first look for a measure for the information that contains an event of that random variable, that is, the fact that the random variable takes on a certain value within its alphabet, such as $X = x_i$. Before establishing a quantitative information measure for this case, some of the basic properties that this measure must fulfill will be intuitively presented, with the aim of later finding some function that fulfills these properties.

- An intuitive notion of information indicates that the amount of information about a certain event is related to the probability with which it occurs.
- Moreover, it must be a decreasing function fo this probability: To know that an unlikely event has occurred generally provides more information than the knowledge of a more probable event.
- Small changes in the probability of the event should lead to small changes in its information, or alternatively, events with a similar probability should have similar information.
- Finally, it must be additive for independent events, if a joint event is defined as the simultaneous occurrence of two independent events. Intuitively, the information that they have as a whole should be the sum of the information of each one of them separately.

Starting from these intuitive notions about some characteristics that a measure of information about a certain event must have, we arrive at the so-called *self-information*, or *surprisal*, which is a quantitative measure of the information contained in an event of a random variable.

Self-information (surprisal)

The self-information or surprisal of the event $X = x_i$ is denoted as $I_X(x_i)$. In order to obtain the analytical expression of this function, the intuitive notions about said measure that have been commented above have been translated into mathematical notation, which gives rise to the four conditions that said function must satisfy.

1. The information measure of an event should depend on its probability, and not on the value of the event itself, that is
$$I_X(x_i) = f(p_X(x_i)).$$
2. It must also be a decreasing function of the probability of the event
$$p_X(x_i) > p_X(x_j) \text{ will imply that } I_X(x_i),$$
which means that the function $f(\cdot)$ must be a decreasing function
$$f(a) < f(b) \text{ for all } a > b.$$

3. The function $f(\cdot)$ used for self-information must also be a continuous function of its argument, so that the variation of information will be continuous over the probability of events.

4. Finally, if two random variables are independent, and a joint event $X = x_i$ and $Y = y_j$ is defined, the information of the joint event must be the sum of the information of each event

$$I_{X,Y}(x_i, y_j) = I_X(x_i) + I_Y(y_j).$$

Since for independent random variables the joint probability can be written as the product of the marginal probabilities of each variable

$$p_{X,Y}(x_i, y_j) = p_X(x_i) \times p_Y(y_j),$$

this means that the function $f(\cdot)$ chosen for the information measure must be additive over the product of the arguments, that is

$$f(a \times b) = f(a) + f(b).$$

It can be shown that the only function that satisfies these properties is the logarithmic function. Therefore, self-information is defined as

$$I_X(x_i) = -\log(p_X(x_i)).$$

Taking into account the properties of the logarithmic function, the self-information can alternatively be written as

$$I_X(x_i) = \log\left(\frac{1}{p_X(x_i)}\right).$$

The base of the logarithm does not determine the characteristics of the information measure, but defines its units. The most frequently used bases are 2 and the natural base, or Euler number e (natural logarithm). If the base is 2, the units are *bits*, and if the natural logarithm is used, the units are *nats*. Henceforth, when the base is not specified, base 2 logarithms will be assumed and therefore bits as units of information. In any case, the change of the base, and therefore of units, does not imply more than a scaling, since in general the logarithms in a certain base are related to the logarithms in the natural base through the following relation

$$\log_b x = \frac{\log_e x}{\log_e b} = \frac{\ln x}{\ln b},$$

which directly supposes a linear relation between the logarithms in two different bases.

Entropy

Self-information provides a quantitative measure of information about an isolated event. If you want to quantify the information of a random variable (for example to model a source of information), all possible events must be taken into account. A reasonable option is to average the information of each event considering its probability. The information content of a random variable thus calculated is called *entropy* and is denoted by $H(X)$. Therefore, the entropy of the random variable X is obtained by averaging the self-information of each of the events that are part of the alphabet of the random variable

$$H(X) = -\sum_{i=0}^{M_X-1} p_X(x_i) \cdot \log p_X(x_i) = \sum_{i=0}^{M_X-1} p_X(x_i) \cdot \log\left(\frac{1}{p_X(x_i)}\right).$$

The units will be bits or nats per symbol, depending on the base that is used.

For the purposes of the computation, it must be taken into account that $0\log(0) = 0$ will be considered. The entropy is a function of the probability of each event, i.e., of the discrete probability density function, and provides a number that represents the information content of that source. It should not be confused with a function of a random variable which is another random variable, as the notation may make it appear.

Entropy can be interpreted as a quantity that represents the uncertainty about the specific value that a random variable X takes, which can model, for example, the output of an information source. If X always takes the same value x_i, that is, if $p_X(x_i) = 1$, there is no uncertainty about the value of the random variable and the entropy is equal to 0. If X stops always taking the same value, then the uncertainty increases and with it the entropy.

Two important properties of the entropy of a discrete random variable are:

1. The entropy of a discrete random variable is a non-negative function, that is

$$H(X) \geq 0.$$

   This is evident since the range of values of a probability is $0 \leq p_X(x_i) \leq 1$ and $\log(x) \leq 0$ for $0 < x \leq 1$. The value $H(X) = 0$ only occurs if one of the elements of the alphabet has probability one and therefore the rest have probability zero.

2. The maximum value that the entropy of a discrete random variable can take is the logarithm of the number of elements of its alphabet

$$H(X) \leq \log(M_X).$$

   That maximum value occurs only if the symbols are equiprobable, $p_X(x_i) = 1/M_X$, which is the situation of maximum uncertainty.

These two properties establish the limits for the minimum and maximum values that the entropy of a random variable X can take, and indicate under which distributions the minimum and maximum values are obtained, respectively. To illustrate these properties, entropy is calculated below in a very simple case: a binary random variable, $M_X = 2$, where the probabilities of each symbol are parameterized with the probability of one of the two elements of the alphabet, $p_X(x_0) = p$, $p_X(x_1) = 1 - p$. In this case the entropy is equal to

$$H(X) = -p\ \log(p) - (1-p)\ \log(1-p) = p\ \log\frac{1}{p} + (1-p)\log\frac{1}{1-p} \equiv H_b(p) \equiv \Omega(p)$$

This function, denoted as $H_b(p)$, or alternatively as $\Omega(p)$, is called *binary entropy function*, and is represented in Figure 4.19 as a function of its argument. Remember that this argument represents the probability of one of the two elements of the alphabet of the binary random variable. Sometimes, the binary entropy function is also called the *horseshoe* function, because of the shape of the function, which can be seen in the figure.

If there is no uncertainty, $p = 0$ or $p = 1$, the entropy is null. Outside of these cases, the value is always greater than zero, taking the maximum value, 1 bit per symbol, when the symbols are equiprobable, which represents the situation of maximum possible uncertainty. Furthermore, since

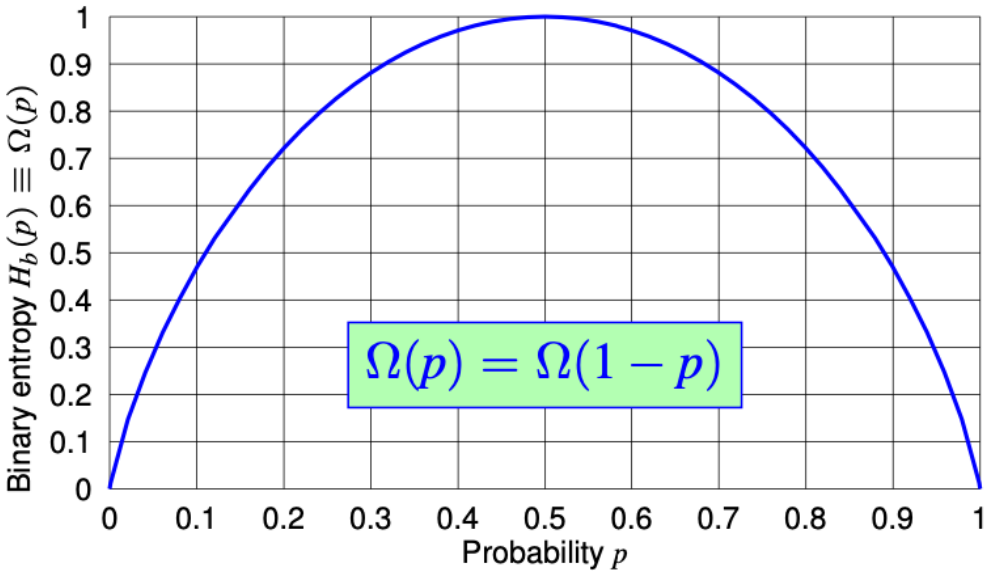


Figure 4.19: Binary entropy function, $H_b(p)$ or $\Omega(p)$, expressed in bits per symbol.

entropy depends only on the values of the probability distribution and not on the alphabet, it is a symmetric function with respect to $p = \frac{1}{2}$, which is evident since by definition

$$\Omega(p) = \Omega(1-p).$$

The binary entropy function can serve as a reference to define an information bit: a bit is the information that is obtained when two symbols are transmitted with equal probability.

Below is an example of calculating the entropy of a random variable with an alphabet of more symbols, specifically five symbols.

Example

A source can be modeled with the DMS model with an alphabet

$$\mathcal{A}_X = \{-2, -1, 0, 1, 2\},$$

and probabilities

$$p_X(-2) = \frac{1}{2},\ p_X(-1) = \frac{1}{4},\ p_X(0) = \frac{1}{8},\ p_X(1) = \frac{1}{16},\ p_X(2) = \frac{1}{16}.$$

In this case the entropy is

$$H(X) = \frac{1}{2}\log(2) + \frac{1}{4}\log(4) + \frac{1}{8}\log(8) + 2 \times \frac{1}{16}\log(16) = \frac{15}{8} \text{ bits/symbol.}$$

It can be seen that the value of entropy depends only on the probabilities of the possible elements of the alphabet, and not on the concrete values of the alphabet. For example, a font with a different alphabet

$$\mathcal{A}_X = \{0, 1, 2, 3, 4\},$$

but with the same set of values for the probabilities, although with a different assignment to each element of the alphabet

$$p_X(0) = \frac{1}{2},\ p_X(1) = \frac{1}{4},\ p_X(2) = \frac{1}{8},\ p_X(3) = \frac{1}{16},\ p_X(4) = \frac{1}{16},$$

has the same entropy as the previous one.

4.3.2 Joint entropy

The definition of entropy can be extended to more than one random variable, which would be applicable, for example, to measure the joint entropy of the input and output of a communication system. The *joint entropy* of two random variables X and Y, generally with different alphabets and probabilities, $\mathcal{A}_X = \{x_i\}_{i=0}^{M_X-1}$, $p_X(x_i)$, and $\mathcal{A}_Y = \{y_i\}_{i=0}^{M_Y-1}$, $p_Y(y_i)$, is defined as a trivial extension of the entropy of a random variable, considering in this case that there are as many events as joint cases and that the probability of each of them is given by the joint probability, which leads to the expression

$$H(X,Y) = \sum_{i=0}^{M_X-1} \sum_{j=0}^{M_Y-1} p_{X,Y}(x_i, y_j) \ \log\left(\frac{1}{p_{X,Y}(x_i, y_j)}\right).$$

As for the entropy of a random variable, given the properties of the logarithm function, a change in the sign and the inversion of the argument of the logarithm provide the same value

$$H(X,Y) = -\sum_{i=0}^{M_X-1} \sum_{j=0}^{M_Y-1} p_{X,Y}(x_i, y_j) \ \log\left(p_{X,Y}(x_i, y_j)\right).$$

Like entropy, it is also measured in bits or nats per symbol. The concept can be extended to N random variables. In this case

$$\mathbf{X} = (X_1, X_1, \cdots, X_N),$$

and

$$H(\mathbf{X}) = -\sum_{x_1,x_2,\cdots,x_N} p_{\mathbf{X}}(x_1, x_2, \cdots, x_N) \ \log(p_{\mathbf{X}}(x_1, x_2, \cdots, x_N)).$$

In this notation, the sum indicates the N sums contemplating all the possible combinations of the alphabets of each random variable.

The interpretation of joint entropy does not differ from that of entropy for a random variable. After all, a pair of variables X and Y can be thought of as a single vector random variable with an alphabet of $M_X \times M_Y$ symbols.

If the random variables X and Y are independent, the joint probability is

$$p_{X,Y}(x_i, y_j) = p_X(x_i) \times p_Y(y_j).$$

In this case, their joint entropy is the sum of the individual entropies. This had been pointed out in the definition of the conditions that the information measure had to meet, and it is demonstrated in a very simple way, as can be seen in the following development

$$\begin{aligned}
H(X,Y) &= \sum_{i=0}^{M_X-1} \sum_{j=0}^{M_Y-1} p_X(x_i)\, p_Y(y_j) \ \log \frac{1}{p_X(x_i)\, p_Y(y_j)} \\
&= \sum_{i=0}^{M_X-1} \sum_{j=0}^{M_Y-1} p_X(x_i)\, p_Y(y_j) \ \log \frac{1}{p_X(x_i)} + \sum_{i=0}^{M_X-1} \sum_{j=0}^{M_Y-1} p_X(x_i)\, p_Y(y_j) \ \log \frac{1}{p_Y(y_j)} \\
&= \sum_{i=0}^{M_X-1} p_X(x_i) \ \log \frac{1}{p_X(x_i)} + \sum_{j=0}^{M_Y-1} p_Y(y_j) \ \log \frac{1}{p_Y(y_j)} \\
&= H(X) + H(Y).
\end{aligned}$$

As we will see in more detail in the next section, it should be noted that this relationship is only fulfilled under the hypothesis of independence between the random variables.

4.3.3 Conditional entropy

Independent random variables produces the greatest entropy if they are independent, since if both variables were not independent, knowledge of the value of one of them would eliminate uncertainty about the value of the other. To measure this uncertainty, the *conditional entropy* of two random variables X and Y, $H(X|Y)$ is used, which averages the value of the conditional entropy of X given Y over all values of the alphabet of Y

$$H(X|Y) = \sum_{j=0}^{M_Y-1} p_Y(y_j)\, H(X|Y=y_j).$$

Considering the definition for the entropy $H(X|Y=y_j)$ from the conditional distribution of X given Y, $p_{X|Y}(x_i|y_j)$, the following equivalent expression is obtained

$$\begin{aligned} H(X|Y) &= \sum_{j=0}^{M_Y-1} p_Y(y_j) \sum_{i=0}^{M_X-1} p_{X|Y}(x_i|y_j)\ \log \frac{1}{p_{X|Y}(x_i|y_j)} \\ &= \sum_{i=0}^{M_X-1} \sum_{j=0}^{M_Y-1} p_{X,Y}(x_i,y_j)\ \log \frac{1}{p_{X|Y}(x_i|y_j)}. \end{aligned}$$

According to Bayes' rule, this probability satisfies the relation

$$p_{X|Y}(x_i|y_j)\, p_Y(y_j) = p_{X,Y}(x_i,y_j).$$

In general, this definition can be naturally extended when the conditioning is with respect to several random variables.

$$\begin{aligned} H(X_N|X_1,X_2,\cdots,X_{N-1}) = \\ - \sum_{x_1,x_2,\cdots,x_N} p_{\mathbf{X}}(x_1,x_2,\cdots,x_N) \log p_{X_N|X_1,X_2,\cdots,X_{N-1}}(x_N|x_1,x_2,\cdots,x_{N-1}). \end{aligned}$$

Conditional entropy can be interpreted as a measure of the uncertainty of a random variable, X, when the value of another random variable, Y, is known. Or in another way, the comparison between $H(X)$ and $H(X|Y)$ quantifies the information that the knowledge of Y gives about X. When the random variables X and Y are independent, knowing the value of one of them does not provide knowledge about the other and therefore does not eliminate uncertainty about its value. Therefore, in this case

$$H(X|Y) = H(X).$$

Conversely, if the knowledge of Y completely determines the value of X, knowing the value of Y there is no uncertainty about the value of X, and the conditional entropy would be $H(X|Y) = 0$.

The joint entropy is related to the entropy and to the conditional entropy through the following

expression

$$\begin{aligned}
H(X,Y) &= \sum_{i=0}^{M_X-1}\sum_{j=0}^{M_Y-1} p_{X,Y}(x_i,y_j)\ \log\frac{1}{p_{X,Y}(x_i,y_j)} \\
&= \sum_{i=0}^{M_X-1}\sum_{j=0}^{M_Y-1} p_{X,Y}(x_i,y_j)\ \log\frac{1}{p_X(x_i)\ p_{Y|X}(y_j|x_i)} \\
&= \sum_{i=0}^{M_X-1}\sum_{j=0}^{M_Y-1} p_{X,Y}(x_i,y_j)\ \log\frac{1}{p_X(x_i)} + \sum_{i=0}^{M_X-1}\sum_{j=0}^{M_Y-1} p_{X,Y}(x_i,y_j)\ \log\frac{1}{p_{Y|X}(y_j|x_i)} \\
&= \sum_{i=0}^{M_X-1} p_X(x_i)\ \log\frac{1}{p_X(x_i)} + \sum_{i=0}^{M_X-1}\sum_{j=0}^{M_Y-1} p_{X,Y}(x_i,y_j)\ \log\frac{1}{p_{Y|X}(y_j|x_i)} \\
&= H(X) + H(Y|X).
\end{aligned}$$

With an equivalent expansion, it is easy to show that the following relation also holds

$$H(X,Y) = H(Y) + H(X|Y).$$

This does not mean that $H(X|Y)$ is equal to $H(Y|X)$, in general. This relationship only holds when $H(X) = H(Y)$.

The joint entropy is obtained as the sum of the entropy of a random variable plus that of the other conditioned on the first. Therefore, the uncertainty of one of the random variables is added to that of the other when the first is known. This means, as we have seen previously, that the joint entropy will only be equal to the sum of the entropy of each of the random variables when the random variables are independent.

In general, this relationship can be extended to the case of a larger number of random variables, in which case applying the chain rule we have the following general relationship

$$H(\mathbf{X}) = H(X_1) + H(X_2|X_1) + H(X_3|X_1,X_2) + \cdots + H(X_N|X_1,X_2,\cdots,X_{N-1}).$$

When $(X_1, X_2, \cdots, X_N)$ are independent random variables, the joint entropy of all of them will be the sum of the entropy of each of the random variables

$$H(\mathbf{X}) = \sum_{i=1}^{N} H(X_i).$$

4.3.4 Mutual Information

Entropy represents a measure of uncertainty about the value of one or more random variables. Another concept that we could define as "complementary" is the so-called *mutual information* between two random variables X and Y, which is denoted $I(X,Y)$. Mutual information represents the information provided by Y about the knowledge of X. The formal definition from the marginal distributions of each random variable and their joint distribution is

$$I(X,Y) = \sum_{i=0}^{M_X-1}\sum_{j=0}^{M_Y-1} p_{X,Y}(x_i,y_j)\ \log\frac{p_{X,Y}(x_i,y_j)}{p_X(x_i)\ p_Y(y_j)},$$

and is measured in bits.

Mutual information is a non-negative measure, $I(X,Y) \geq 0$, which can be expressed in terms of entropy and conditional entropy, since the following relationship holds:

$$\begin{aligned}
I(X,Y) &= \sum_{i=0}^{M_X-1}\sum_{j=0}^{M_Y-1} p_{X,Y}(x_i,y_j)\ \log\frac{p_{X,Y}(x_i,y_j)}{p_X(x_i)\ p_Y(y_j)} \\
&= \sum_{i=0}^{M_X-1}\sum_{j=0}^{M_Y-1} p_{X,Y}(x_i,y_j)\ \log\frac{p_{X|Y}(x_i,y_j)}{p_X(x_i)} \\
&= \sum_{i=0}^{M_X-1}\sum_{j=0}^{M_Y-1} p_{X,Y}(x_i,y_j)\ \log\frac{1}{p_X(x_i)} + \sum_{i=0}^{M_X-1}\sum_{j=0}^{M_Y-1} p_{X,Y}(x_i,y_j)\ \log(p_{X|Y}(x_i,y_j)) \\
&= \sum_{i=0}^{M_X-1} p_X(x_i)\ \log\frac{1}{p_X(x_i)} - \sum_{i=0}^{M_X-1}\sum_{j=0}^{M_Y-1} p_{X,Y}(x_i,y_j)\ \log\frac{1}{p_{X|Y}(x_i,y_j)} \\
&= H(X) - H(X|Y).
\end{aligned}$$

Equivalently

$$I(X,Y) = H(Y) - H(Y|X).$$

On the other hand, from the very definition of mutual information

$$I(X,Y) = I(Y,X),$$

and taking into account the relationship between marginal, conditional and joint entropies

$$H(X,Y) = H(Y) + H(X|Y),$$

it is straightforward to see that the mutual information can also be obtained as

$$I(X,Y) = H(X) + H(Y) - H(X,Y).$$

These relationships between mutual information and the different entropies are usually represented graphically by a *Venn diagram*, like the one shown in Figure 4.20.

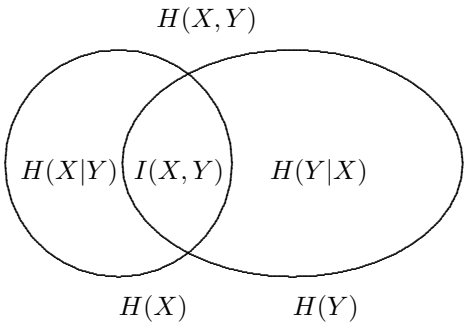


Figure 4.20: Venn diagram illustrating the relationships between entropies and mutual information.

The area of a circle represents the entropy of a random variable, $H(X)$ or $H(Y)$, and the intersection between them is the mutual information $I(X,Y)$, while the area covered by both

circles is the joint entropy, $H(X,Y)$. The difference between a circle and the intersection represents the conditional entropy. If the random variables are independent, the circles would have zero intersection. If both variables are equal, the intersection is complete and the conditional entropy is zero. Figure 4.21 shows more clearly how each of the measurements is identified on the Venn diagram with a simple color code.

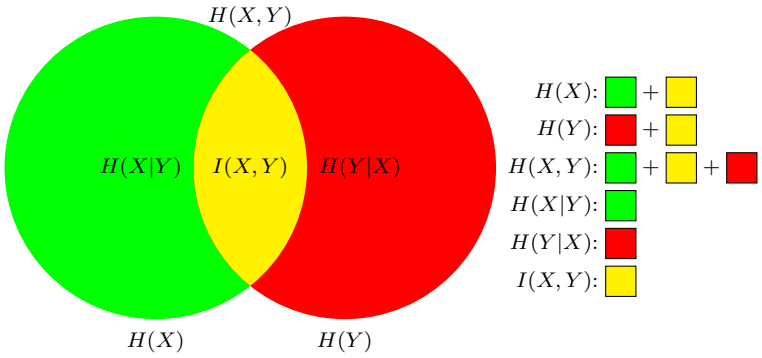


Figure 4.21: Identification with a simple color code of the different entropies and mutual information on the Venn diagram.

Below is an example in which different entropies and mutual information are calculated for two simple random variables..

Example

There are two binary random variables, X and Y, with the same alphabet $x_0 = y_0 = 0$, $x_1 = y_1 = 1$, and with the following joint distribution

$$p_{X,Y}(0,0) = \frac{1}{3},\ p_{X,Y}(0,1) = \frac{1}{3},\ p_{X,Y}(1,0) = \frac{1}{3},\ p_{X,Y}(1,1) = 0.$$

To calculate the entropy of each random variable, it is necessary to know the marginal distributions, which are easily obtained from the joint distribution

$$p_X(x_i) = \sum_{j=0}^{M_Y-1} p_{X,Y}(x_i,y_j) \text{ and } p_Y(y_j) = \sum_{i=0}^{M_X-1} p_{X,Y}(x_i,y_j).$$

In this case

$$p_X(0) = p_Y(0) = \frac{2}{3},\ p_X(1) = 0_Y(1) = \frac{1}{3}.$$

Therefore, as in this case the two variables have the same distribution, the entropy of both variables, parameterized through the binary entropy function (since the binary random variables)

$$H(X) = H(Y) = H_b\left(\frac{2}{3}\right) = H_b\left(\frac{1}{3}\right) = 0.919.$$

The joint entropy, applying its definition, will be given by

$$H(X,Y) = 3 \times \left(\frac{1}{3}\log_2(3)\right) + 0\log_2(0) = \log_2(3) = 1.585.$$

This result can also be interpreted as (X,Y) being a vector of two random variables with an alphabet of three events, $(0,0)$, $(0,1)$ and $(1,0)$, all of them equally likely.

From the previous results, the conditional entropy can be obtained through the relation

$$H(X|Y) = H(X,Y) - H(Y) = 1.585 - 0.919 = 0.666.$$

Similarly, mutual information could be obtained, for example, through the relationship

$$I(X,Y) = H(X) - H(X|Y) = 0.919 - 0.666 = 0.253.$$

Mutual information between discrete random variables has a number of properties that should be taken into account. Among them, the following should be highlighted:

1. It is non-negative
$$I(X,Y) = I(Y,X) \geq 0.$$
The minimum value $I(X,Y) = 0$ is obtained if X and Y are independent.

2. Its maximum value is bounded by the value of the entropy of each of the random variables, so in practice it is bounded by the minimum value of the entropy of the random variables
$$I(X,Y) \leq \min(H(X), H(Y)).$$
Mutual information can never be greater than the measure of information that each of the variables has individually.

3. Conditional mutual information can be defined as the average of the mutual information given each of the possible values of the random variable with respect to which it is conditioned
$$I(X,Y|Z) = \sum_{i=0}^{M_z-1} p_Z(z_i)\ I(X,Y|Z=z_i).$$

4. The conditional mutual information $I(X,Y|Z)$ can also be obtained through the conditional entropies as
$$I(X,Y|Z) = H(X|Z) - H(X|Y,Z).$$

5. The chain rule for mutual information is defined from
$$I((X,Y),Z) = I(X,Z) + I(Y,Z|X).$$

6. In general, the chain rule is
$$I((X_1,X_2,\cdots,X_N),Y) = I(X_1,Y) + I(X_2,Y|X_1) + \cdots + I(X_N,Y|X_1,\cdots,X_{N-1}).$$

7. From the definition of mutual information we obtain the definition of entropy as mutual information of a random variable with itself. This relationship is easily demonstrated by taking into account that the distribution of a random variable with itself takes the form
$$p_{X,X}(x_i,x_j) = \delta[i-j]\ p_X(x_i),$$
so the mutual information of the random variable X with itself is
$$\begin{aligned} I(X,X) &= \sum_{i=0}^{M_X-1}\sum_{j=0}^{M_X-1} \delta[i-j]\ p_X(x_i)\ \log\frac{\delta[i-j]\ p_X(x_i)}{p_X(x_i)\ p_X(x_j)} \\ &= \sum_{i=0}^{M_X-1} p_X(x_i)\ \log\frac{p_X(x_i)}{p_X(x_i)\ p_X(x_i)} \\ &= \sum_{i=0}^{M_X-1} p_X(x_i)\ \log\frac{1}{p_X(x_i)} \\ &= H(X). \end{aligned}$$

This is why the name *auto-information* is sometimes used to name the entropy.

4.3.5 Differential entropy and mutual information

Until now, the measurements that have been presented refer to discrete random variables, which serves to model discrete time information sources with a discrete alphabet. For these variables, the entropy, $H(X)$ and the mutual information $I(X,Y)$, as well as the conditional and joint entropies, $H(X|Y)$ and $H(X,Y)$, have been presented.

To model a source discrete in time but with a continuous alphabet, for example a sampled audio source, it is necessary to use a continuous random variable. In this case, the analog of entropy for discrete random variables is called *differential entropy*. However, this measure does not have the intuitive meaning that entropy had, which is due to several aspects, such as the fact that in a continuous random variable, by definition the probability of a particular continuous value is zero.

Formally, the differential entropy of a continuous random variable X, with a probability density function $f_X(x)$, is defined as

$$h(X) = \int_{-\infty}^{\infty} f_X(x)\ \log \frac{1}{f_X(x)}\ dx,$$

where, again, $0 \log(1/0) = 0$.

Example

Differential entropy of a uniformly distributed random variable in an interval $[0, a]$.

Using the definition of differential entropy directly, and taking into account that the probability density function of the random variable is $1/a$ between 0 and a, the entropy is equal to

$$h(X) = \int_0^a \frac{1}{a} \log(a)\ dx = \log(a).$$

From this example you can see some interesting properties:

1. For $a < 1$ we have $h(X) < 0$, which goes against the non-negativity property of the entropy of a discrete random variable.
2. For $a = 1$, we have $h(X) = 0$, and in this case X is not deterministic, so it has a certain degree of uncertainty. This also goes against the properties of entropy for discrete random variables.

Example

If X has a Gaussian probability density function with zero mean and variance σ^2

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{x^2}{2\sigma^2}}.$$

Using natural logarithms, the differential entropy for this random variable in nats is computed as

$$\begin{aligned} h(X) =& -\int_{-\infty}^{\infty} f_X(x)\ \ln \frac{1}{\sqrt{2\pi\sigma^2}}\ dx - \int_{-\infty}^{\infty} f_X(x)\ \ln\left(e^{-\frac{x^2}{2\sigma^2}}\right)\ dx \\ =& \ln(\sqrt{2\pi\sigma^2}) + \frac{\sigma^2}{2\sigma^2} = \ln(\sqrt{2\pi\sigma^2}) + \frac{1}{2} \\ =& \frac{1}{2} \ln(2\pi e\sigma^2) \text{ nats.} \end{aligned}$$

To arrive at this expression, the following properties of the Gaussian distribution have been used

$$\int_{-\infty}^{\infty} f_X(x)\, dx = 1 \text{ y } \int_{-\infty}^{\infty} x^2\, f_X(x)\, dx = \sigma^2.$$

Changing the base of the logarithm to 2, we have the differential entropy in bits

$$h(X) = \frac{1}{2}\log_2(2\pi e\sigma^2) \text{ bits.}$$

Depending on the variance, in particular when compared with $\sigma^2 = \frac{1}{2\pi e}$, this entropy can take positive, negative or zero values.

As for discrete random variables, joint and conditional entropies are also defined for continuous random variables. The *joint differential entropy* between two random variables X and Y is defined as

$$h(X,Y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{X,Y}(x,y)\log\frac{1}{f_{X,Y}(x,y)}dx\, dy.$$

As for the *conditional differential entropy*, its definition is

$$h(X|Y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{X,Y}(x,y)\log\frac{1}{f_{X|Y}(x|y)}dx\, dy.$$

The alternative but equivalent definition is often used

$$h(X|Y) = \int_{-\infty}^{\infty} f_Y(y)\int_{-\infty}^{\infty} f_{X|Y}(x|y)\log\frac{1}{f_{X|Y}(x|y)}dx\, dy.$$

It can be seen that they are the natural extensions of the definitions for discrete random variables. Therefore, the same relationships hold. Specificly

$$h(X,Y) = h(Y) + h(X|Y).$$

Similarly, the *mutual information* for continuous random variables can be defined as

$$I(X,Y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{X,Y}(x,y)\log\frac{f_{X,Y}(x,y)}{f_X(x)f_Y(y)}dx\, dy,$$

which is also measured in bits, and where to avoid ambiguities we define $0\log\frac{0}{0} = 0$.

As in the case of discrete random variables, mutual information can be expressed in terms of entropies

$$I(X,Y) = h(Y) - h(Y|X) = h(X) - h(X|Y) = h(X) + h(Y) - h(X,Y).$$

Unlike for differential entropy for continuous random variables, where the intuitive interpretation of entropy for discrete random variables as a measure of uncertainty or information is not maintained, for mutual information it is maintained and has the same meaning. The mutual information indicates the knowledge that one variable contributes about the other. Furthermore, most of the basic properties of mutual information for discrete random variables hold. In particular, given its definition, the following properties are satisfied:

1. $I(X,Y) \geq 0$, that is, it is a non-negative function.
2. $I(X,Y) = 0$ only if the variables X and Y are independent.
3. $I(Y,X) = I(X,Y)$.

4.4 Channel capacity

Once the concepts of entropy and mutual information have been defined, in this section we will try to determine the maximum amount of information that can be transmitted through a channel using these information measures. Firstly, the concept of channel coding will be introduced as a mechanism to achieve reliable communication through a certain unreliable channel, to then define the capacity of a channel and study how this value is obtained for two types of channels: the digital channel modeled by means of a DMC and the Gaussian channel.

4.4.1 Channel coding for reliable transmission

The main objective when transmitting information over any communication channel is *reliability*. This reliability in digital communication systems is measured by the error probability at the receiver. As in any communication channel, apart from possible distortions, noise is introduced. At first glance, it may seem that the error probability will always be bounded by a non-zero value that will depend on the noise level, that is,

$$P_e \geq K = f(\sigma^2),$$

where σ^2 is the noise power at the receiver input. However, as already outlined in the introduction, a fundamental result of information theory says that reliable transmission, meaning reliable transmission as one in which there is a probability of error below any fixed limit, is possible even on noisy channels as long as the transmission rate is below a certain value called *channel capacity*. This result was presented by Shannon in 1948 and is known as the *noisy channel-coding theorem*. As a summary, what the channel coding theorem says is that *the basic limitation introduced by noise in a communications channel is not in the reliability of the communication, but in its transmission rate*. Thus, it will be possible to obtain a communication with an arbitrarily low probability of error as long as information is transmitted at a rate below a limit value that will depend on the characteristics of the channel.

In the first place, we will see how to obtain this limit in an intuitive way, following the same line of argument that is presented in [Proakis and Salehi, 2002], to later formulate it in terms of information theory.

Figure 4.22 shows a discrete memoryless channel or DMC, with an input alphabet consisting of four elements, $\mathcal{A}_X = \{a, b, c, d\}$ and the same output alphabet $\mathcal{A}_Y = \mathcal{A}_X$.

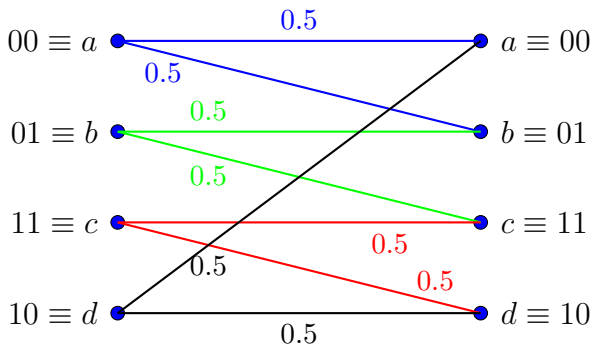


Figure 4.22: Example of a discrete memoryless channel.

Since this is not an ideal channel, if this channel is used for transmission in the usual way, the receiver will not be able to identify with zero error probability the transmitted symbol by looking at the received symbol. When the value a is present at the output, the receiver cannot discern whether a or b has been transmitted, since the transmission of both can produce the symbol a at the output of the channel. The same is true for the other possible output values. If b is observed, it is not possible to discern without error whether it is due to the transmission of a or b, and so on. Therefore, there is an error probability that is fixed by the characteristics of the channel, in this case by the transition probabilities that define it.

However, given the special characteristics of this channel, it will be possible to transmit information without errors. It is evident that the impossibility of discerning without error the transmitted symbol by looking at the output value is due to the fact that the sets that form the possible output values associated with the transmission of each symbol "*overlap*". Thus, there is an overlap between the outputs when a is transmitted (which can be a and b) and when b is transmitted (which can be b and c). So when b is observed in the output it is not possible to know with absolute certainty which symbol has been transmitted.

But for this channel it is possible to choose a subset of the input alphabet whose outputs do not overlap. As it can be seen in Figure 4.23, if only the symbols a and c are transmitted, in view of the output there will be no possible ambiguity about the transmitted symbol: if the output contains a or b, it is known with certainty that the transmitted symbol is a; the same thing happens if we have c or d in the output, in which case it is certain that c has been transmitted. Therefore, by transmitting this subset of possible values of X, the error probability is zero. The price to pay for this reliability is the transmission rate. By transmitting the 4 symbols that are part of the alphabet of X, two bits of information are transmitted per channel use. However, if only 2 symbols are transmitted, only a single bit of information will be transmitted per channel use (the number of information bits per transmitted symbol is $\log_2 M$, where M is the number of elements of the alphabet of symbols that are transmitted).

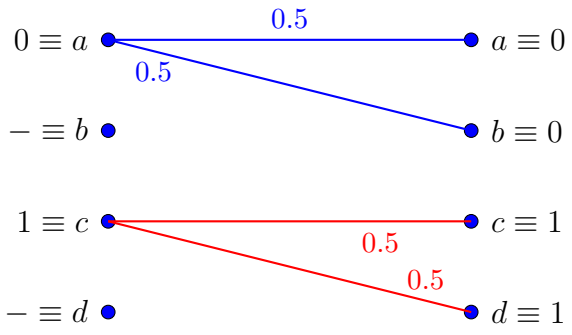


Figure 4.23: Reliable transmission over an ureliable channel.

The mechanism used in this example to be able to transmit with zero probability of error illustrates the fundamental idea underlying the channel coding theorem for a reliable transmission: to use in transmission only those symbols whose corresponding outputs are disjoint. Here it is necessary to clarify that the objective of channel coding is not really to achieve a transmission with zero error probability, but with an error probability below a certain value, which can be arbitrarily low. With this purpose, the subset of transmitted symbols can have outputs that overlap with a sufficiently low probability.

A problem that arises with this idea for transmission with arbitrarily low probability of error, is that in practice real channels do not behave like in Figure 4.22, where there are input symbols

whose outputs do not overlap. In the vast majority of cases there is not even a subset of symbols whose outputs overlap with a sufficiently low probability. However, channel coding theory proposes a simple mechanism to artificially generate a situation similar to this. Although this mechanism can be used for digital channels with M-ary alphabets, for simplicity this mechanism will be illustrated using a binary case as an example. Specifically, the binary symmetric channel model or BSC, in which the bit error probability (BER) will be denoted by ε, as shown in the trellis diagram on the left side of Figure 4.24.

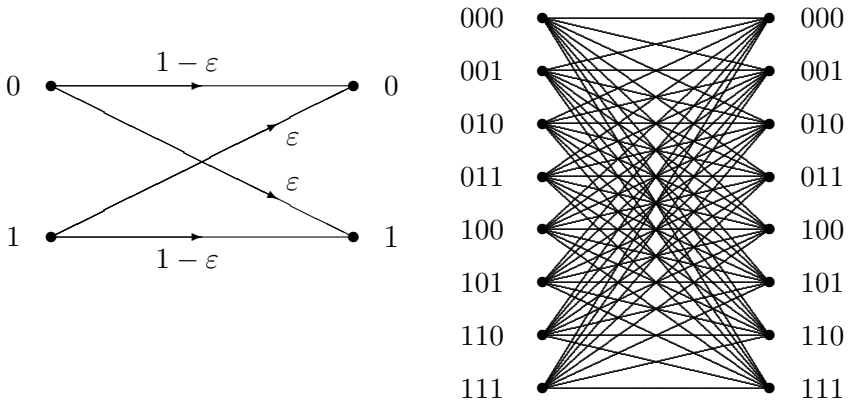


Figure 4.24: Binary Symmetric Channel (BSC) and its corresponding extended channel of order $n = 3$.

If the BSC channel is observed and one tries to apply the procedure applied to the channel in Figure 4.22, it can be seen that it is not directly possible, since the outputs of the two symbols overlap completely with an arbitrary probability (given by ε) and also because there are only two symbols (if only one of them is transmitted, there will not really be any information in the transmission). In general, it's not going to be possible to apply that procedure directly to almost any real DMC channel. To use this idea, what is done is to apply it not directly on the channel but on the so-called *extended channel* of order n. The extended channel of order n is defined as a channel in which blocks of n symbols are grouped to form extended symbols (of order n), giving rise to input alphabets $\mathcal{A}_X^n$ and $\mathcal{A}_Y^n$. The idea is to transmit the information not in each individual use of the channel, but jointly in n uses of the channel. The diagram on the right of Figure 4.24 illustrates the idea of order extension $n = 3$ for the BSC. The information will be transmitted by blocks making 3 uses of the channel, so that the alphabet of this extended channel is now made up of 8 possible extended symbols, corresponding to the 8 possible values that the symbols of the original channel can take on in the 3 uses that define the extended channel. Therefore, we have gone from working with a system with an alphabet of two symbols

$$\mathcal{A}_X = \mathcal{A}_Y = \{0, 1\},$$

to an extended system with alphabet of 2^n symbols, in this case $2^3 = 8$ symbols

$$\mathcal{A}_X^3 = \mathcal{A}_Y^3 = \{000, 001, 010, 011, 100, 101, 110, 111\}.$$

The transition probabilities on this extended channel are obtained as products of the n transition probabilities through the original channel associated with each case on the extended channel. If the extended symbols are defined as vectors of n elements, statistically represented by the vector random variables $\mathbf{X}$ and $\mathbf{Y}$, with alphabets

$$\mathbf{X} \in \{\mathbf{x}_i \mid i = 0, 1, \cdots, M_X^n - 1\},\ \mathbf{Y} \in \{\mathbf{y}_j \mid j = 0, 1, \cdots, M_Y^n - 1\}$$

with

$$\mathbf{x}_i = [x_i[0], x_i[1], \cdots, x_i[n-1]]\,,\ \mathbf{y}_j = [y_j[0], y_j[1], \cdots, y_j[n-1]]$$

the transition probabilities on the extended symbols are

$$p_{\mathbf{Y},\mathbf{X}}(\mathbf{y}_j|\mathbf{x}_i) = \prod_{\ell=0}^{n-1} p_{Y|X}(y_j[\ell]|x_i[\ell]).$$

In the figure, the branches with the transition probabilities have not been labeled due to lack of space, but these probabilities are obtained very easily, with 4 possible values:

- If the input and output extended symbols coincide, the associated transition probability is

$$p_{\mathbf{Y},\mathbf{X}}(\mathbf{y}_i|\mathbf{x}_i) = (1-\varepsilon)^3.$$

- If the Hamming distance between the input extended symbol and the output symbol is 1 (only one bit changes), the associated transition probability is

$$p_{\mathbf{Y},\mathbf{X}}(\mathbf{y}_j|\mathbf{x}_i) = \varepsilon\ (1-\varepsilon)^2.$$

- If the Hamming distance between the input extended symbol and the output symbol is 2 (two bits change), the associated transition probability is

$$p_{\mathbf{Y},\mathbf{X}}(\mathbf{y}_j|\mathbf{x}_i) = \varepsilon^2\ (1-\varepsilon).$$

- Finally, if the Hamming distance between the input extended symbol and the output symbol is 3 (the three bits change), the associated transition probability is

$$p_{\mathbf{Y},\mathbf{X}}(\mathbf{y}_j|\mathbf{x}_i) = \varepsilon^3.$$

Basically, the probability that a bit coincides between input and output is $1-\varepsilon$, and the probability that it is different is ε, from which these four probabilities are directly obtained.

Clearly, for low bit error probabilities ε, the first two probabilities are much higher than the last two (it is much more likely to have zero or one error over all three bits than two or three errors). Therefore, transitions could be divided into high-probability and low-probability transitions. This division is shown in Figure 4.25 by means of a color code: high probability transitions are represented in black and low probability transitions in green.

Having made this distinction, it is now possible to search for a subset of input extended symbols whose outputs are disjoint in terms of high probability transitions, which happens for example with the extended symbols 000 and 111, as illustrated in the figure. Now, when one of the two extended symbols is transmitted, if only the most likely transitions occur, the transmitted extended symbol will be correctly identified from the output. Obviously, some low probability transition may also occur, in which case an error will occur in identifying the transmitted extended symbol. In this case, the error probability will be given by the probability that any of the unlikely transitions will occur, which for this example will be

$$P_e = 1 \times (1-\varepsilon)^3 + 3 \times \varepsilon\ (1-\varepsilon)^2.$$

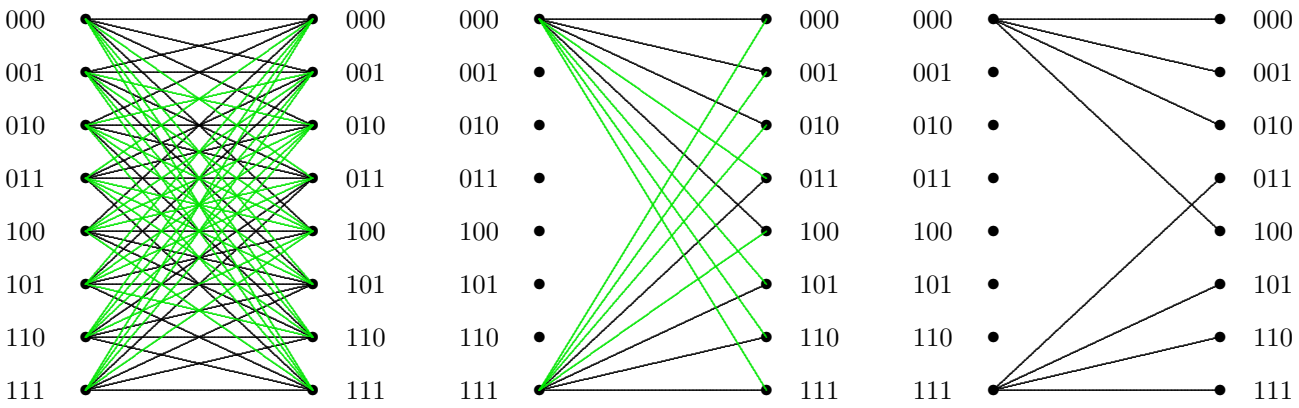


Figure 4.25: Distinguishing between high (black) and low (green) probability transitions and selecting a subset of non-overlapping extended symbols on high probability links.

This value is obtained taking into account that low probability transitions are those corresponding to two or three bit errors in the transmission, and that there is only one pattern of three errors and three of two errors for each block of three bits transmitted.

As a numerical example to illustrate the benefit obtained, if a binary communications system has a BER $\varepsilon = 0.1$, one out of every 10 bits will be in error (10% percentage). Using this system as a base and an extension like the one in the order example $n = 3$, the probability of error will become $P_e = 0.028$; that is, a bit error rate of 2.8% can be achieved using a system with a 10% error rate as a base for transmission. For the case $\varepsilon = 0.01$ the percentage of error bits of the system will be 1%. In that case $P_e = 2.98 \times 10^{-4}$, which means an error rate of approximately 0.03%. If $\varepsilon = 10^3$ (one error out of a thousand), $P_e = 2.998 \times 10^{-6}$ (three errors out of a million).

This technique makes it possible to reduce the probability of a system error, but naturally it does so at the cost of the information transmission rate. If the system is used without the extension, each time the channel is used, one bit of information is transmitted. If the extended system is used, since only 2 of the 8 possible extended symbols are transmitted, each of them will carry one bit of real information, e.g., 000 will carry "0" and 111 will carry "1". In this way, a single bit of information will be sent per three channel uses, which means that the effective rate is 1/3 (1 bit of information for every 3 uses of the channel), has decreased with respect to direct transmission without the extension.

The choice of the non-overlapping subset of elements in low-probability transitions need not be restricted to two extended symbols. For example, it would be possible to make an extension of order $n = 5$, and choose 4 extended symbols whose high probability transitions (defined as those in which zero errors or one bit error occur over the five bits transmitted) do not overlap; For example, they could be the symbols 00000, 10101, 01110 and 11011. In this case, since there are 4 extended symbols, every 5 uses of the channel will send 2 bits of real information, so the effective information transmission rate will be 2/5.

Channel coding

In general, this technique based on the definition of extended symbols of order n that is used to carry out a transmission with a sufficiently low error probability on a system that inherently has a higher error probability is called *channel coding*. The technique is not limited to its use on binary systems, as in the examples that have been used for its presentation, but it can also be used on

M-ary systems. Figure 4.26 illustrates the operation of a system that uses channel coding as a mechanism to control the probability of system error.

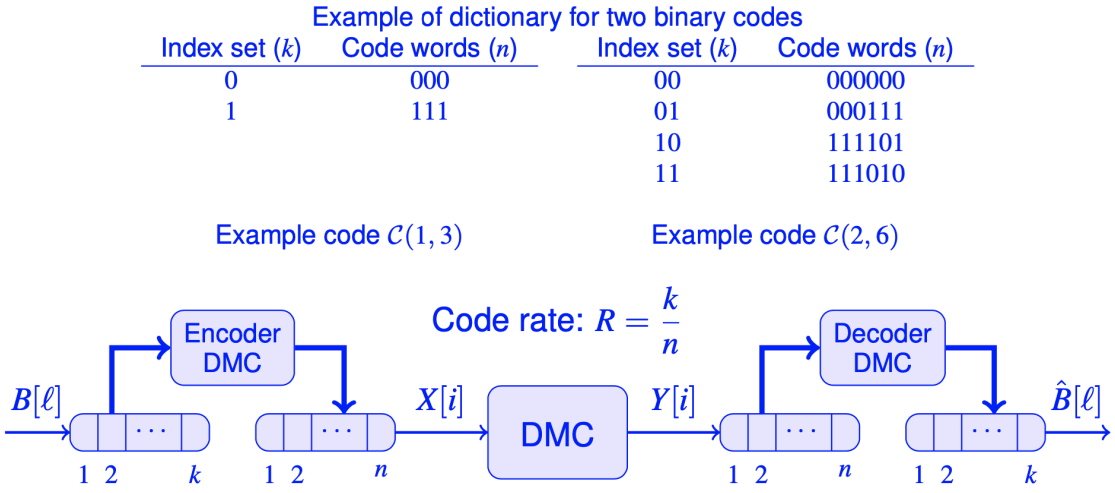


Figure 4.26: Channel encoder and decoder for transmission on a digital channel modeled by a DMC, and two examples of dictionaries for the code.

A binary channel coding system is composed of an encoder on the transmitter side, which performs *channel coding*, and a decoder on the receiver side, which performs *channel decoding*. The encoder takes as input a vector of k bits, which defines an input alphabet, formally called *index set*, of 2^k elements. In the above examples, this set of indices would be composed of the effective information blocks that are going to be transmitted in each n uses of the channel. For each input, the encoder generates as output a vector of n bits that are sent over the DMC, making each of the combinations of k bits that make up the set of indices correspond to a combination of n DMC symbols or *codewords*. In the above examples, these codewords would be the subset of extended symbols whose outputs overlap with low probability. Under these conditions, this is said to be a code $\mathcal{C}(k, n)$, where the value n is often called the length of the code. The set of all code words (2^k), is called *dictionary*. The decoder works reciprocally. Formally, a code is defined by the set of indices and the encoding and decoding functions.

A very important parameter in any channel code is ratio between the number of symbols at the encoder input, k, and the number of times the DMC is used to transmit the code word, n, since it defines the actual amount of information that is transmitted in each use of the channel. This parameter is called *transmission rate* or simply *coding rate*. It is usually denoted by R and is measured in symbols by use of the DMC

$$R = \frac{k}{n} \text{ symbols per use.}$$

For binary systems, the units of this code rate are bits per channel use.

An intuitive reasoning is that if both k and n increase, but maintaining the ratio k/n, the performance will improve (for instance, by replacing the $\mathcal{C}(1, 3)$ code with the $\mathcal{C}(2, 6)$). This intuition is correct, but it has a limit: the channel capacity.

Noisy channel coding theorem

When presenting the principle of channel coding, it has been seen that there is a compromise between the transmission rate or code rate and the obtained error rate. Some questions naturally arise regarding the performance that can be obtained with this technique. Is it possible to reduce the error probability as much as desired using this technique on any type of channel? What limitations do the channel characteristics impose on the performance that the channel capacity can offer? The answer to these questions is found in the so-called *channel coding theorem*, presented by Claude Shannon in 1948.

The channel coding theorem shows that there is a limit to the maximum transmission rate over a DMC, which is called *channel capacity*, and is formally obtained as the maximum value, over all possible distributions for the channel input alphabet, of the mutual information between the channel input and output, that is

$$C = \max_{p_X(x_i)} I(X,Y),$$

where $I(X,Y)$ is the mutual information between the input X and the output Y of the channel. The theorem also proves the following aspects:

1. If the transmission rate R is less than the capacity of the channel C, then for any arbitrarily low value $\delta > 0$ there exists a code with a sufficiently long block length n whose probability of error is less than δ.
2. If $R > C$, the error probability of any code with any block length is limited by a non-zero value that depends on the characteristics of the channel.
3. There are codes that allow reaching the channel capacity $R = C$.

It is important to clarify the third point here. The theorem, although it shows that it is possible to achieve such a capacity, does not answer the question of how such codes can be obtained in practice. In a practical problem, this capacity will not be reached, but in general codes that are below this limit are used.

In the next section we will study how to calculate the channel capacity first for a digital channel, and then for the Gaussian channel. Practical channel code design and analysis is not within the scope of this course, but will be covered in the course "*Digital Communications*".

4.4.2 Channel capacity for the digital channel

In the first place, the case of the binary digital channel is going to be studied, in which n bits are grouped to form the new extended symbols that will be grouped at the input and output of the system to implement the channel coding. For this case, the maximum amount of information that can be reliably transmitted through the channel is going to be obtained in two ways: through an intuitive explanation based on the definition of high probability transitions along the lines of the explanation used previously to explain the principle of channel coding, as presented in [Proakis and Salehi, 2002]; and by the definition presented by Shannon in the channel coding theorem.

Applying the law of large numbers, for sufficiently large values of n, when a sequence of n bits is transmitted over a binary channel with bit error probability ε, the output will have with high

probability $n \times \varepsilon$ bit errors; that is to say, the received sequence will have with high probability $n \times \varepsilon$ different bits with respect to the transmitted sequence. The number of possible sequences of n bits that differ by $n\,\varepsilon$ bits is given by the combinatorial number

$$\binom{n}{n\,\varepsilon}.$$

Considering that a combinatorial number can be obtained as

$$\binom{n}{k} = \frac{n!}{k!(n-k)!},$$

and using Stirling's approximation for the factorial numbers, which is given by

$$n! \approx n^n e^{-n}\sqrt{2\pi n}$$

that combinatorial number can be approximated as

$$\binom{n}{n\,\varepsilon} \approx 2^{n\,H_b(\varepsilon)},$$

where $H_b(\varepsilon)$ is the binary entropy function with argument ε. This means that for every possible sequence of n bits transmitted there are approximately $2^{n\,H_b(\varepsilon)}$ highly probable sequences in the output.

On the other hand, the total number of highly probable sequences of length n bits at the output depends on the uncertainty of each of the bits that make up the sequence, measured through its entropy, which in turn depends on the probability of having a one or a zero at the output. If both symbols are equiprobable, all possible sequences will have the same probability, and the number of highly probable sequences will be 2^n. But if the bits in the output are not equally likely, the probability of the output sequences will be different for each sequence. For example, if the "0" bit is less likely than the "1" bit, sequences with many zeros will be less likely. In that case, the number of highly probable bit sequences can be approximated from the entropy measure of the output, $H(Y)$, which will be $H(Y) = H_b(p_Y(0))$, using the expression

$$2^{n\,H(Y)}.$$

Therefore, the maximum number of sequences of n bits in the input without overlap between the highly probable outputs that they generate in the output, and which will be denoted as M_{no}, will be the quotient between the number of highly probable sequences in the output and the number of sequences that with high probability are generated in the output when a certain sequence is transmitted in the input, that is to say

$$M_{no} = \frac{2^{nH(Y)}}{2^{nH_b(\varepsilon)}} = 2^{n(H(Y)-H_b(\varepsilon))}.$$

The number of information bits that can be associated with those non-overlapping M_{no} sequences is

$$\log_2 M_{no} = n\,(H(Y) - H_b(\varepsilon)) \text{ bits of information.}$$

Therefore, the resulting code rate will be the quotient between the information bits and the number of uses of the channel.

$$R = \frac{\log_2 M_{no}}{n} = H(Y) - H_b(\varepsilon).$$

The maximum possible value of this rate R is the one that defines the so-called *channel capacity*, which is denoted as C. The entropy of the bits at the output of the binary channel will be maximum when the bits "0" and "1" are equally likely, in which case $H(Y) = 1$. Therefore, intuitively it has been arrived at that the channel capacity for a symmetric binary digital channel with bit error probability ε is

$$C = 1 - H_b(\varepsilon) \text{ bits per channel use.}$$

This result has been reached intuitively for the case of the BSC channel. The same result will then be obtained from information theory. To do this, the mutual information between the input and the output of the channel is calculated and through it an attempt will be made to find out what part of the information is transmitted and what part is lost as it passes through the channel.

To calculate the information between the input, X, and the output, Y, of the DMC channel, it is necessary to know the distributions of both variables. Knowing the input distribution, $p_X(x_i)$, as the transition probabilities are known, the joint distribution of the input and output is known

$$p_{X,Y}(x_i, y_j) = p_{Y|X}(y_j|x_i)\, p_X(x_i).$$

From this, the output distribution is obtained as

$$p_Y(y_j) = \sum_{i=0}^{M_X-1} p_{X,Y}(x_i, y_j) = \sum_{i=0}^{M_X-1} p_{Y|X}(y_j|x_i)\, p_X(x_i).$$

In this way, the complete input/output characterization is obtained. Note that for a BSC

$$p_{Y|X}(y_j|x_i) = \begin{cases} 1-\varepsilon & \text{si } j = i \\ \varepsilon & \text{si } j \neq i \end{cases}.$$

From these distributions the mutual information between input and output $I(X,Y)$ can be calculated, for example through the relations with the different entropies, by means of

$$I(X,Y) = H(X) + H(Y) - H(X,Y) = H(X) - H(X|Y) = H(Y) - H(Y|X).$$

The mutual information between the input and the output represents the information that the output of the channel provides about the input, or the uncertainty that is eliminated about the value of the input when the output is known: in short, the information that is transmitted by the channel.

To clarify this idea, we will use $I(X,Y) = H(X) - H(X|Y)$ and we will analyze two extreme cases of the simplest DMC, the BSC. In this case, the best possible channel is the one free of errors, that is, $\varepsilon = 0$, or alternatively $\varepsilon = 1$. Keep in mind that in a binary system, being always wrong is a way of always being right, if just the binary decision is changed. The worst possible case is the case where $\varepsilon = 1/2$ (any ε greater than $1/2$ can be assimilated, by changing the decision, to a $1-\varepsilon$ case).

For the case $\varepsilon = 0$, the joint input-output distribution is

$p_{X,Y}(x_i, y_j)$	x_0	x_1	$p_Y(y_j)$
y_0	$p_X(x_0)$	0	$P_X(x_0)$
y_1	0	$p_X(x_1)$	$P_X(x_1)$

or considering that no errors occur, $p_Y(y_i) = p_X(x_i)$ or you can simply set the equality $Y = X$. Therefore,

$$I(X,Y) = I(X,X) = H(X),$$

which means that $H(X|Y) = 0$.

On the other hand, the joint distribution when $\varepsilon = 1/2$ is

$p_{X,Y}(x_i, y_j)$	x_0	x_1	$p_Y(y_j)$
y_0	$p_X(x_0)/2$	$p_X(x_1)/2$	$1/2$
y_1	$p_X(x_0)/2$	$p_X(x_1)/2$	$1/2$

.

In this case, Y has an equiprobable distribution independently of the distribution of X, as expected. This means that X and Y are statistically independent, and the joint probability can be written as

$$p_{X,Y}(x_i, y_j) = p_X(x_i)\, p_Y(y_j).$$

As already deduced previously, if X and Y are independent, their mutual information is null,

$$I(X,Y) = 0,$$

which means that $H(X|Y) = H(X)$.

The following conclusions can be drawn from these two cases:

1. The mutual information between input and output of the channel is the amount of information that passes from the input to the output when the channel is used. In the case where the probability of error is zero, all information is passed ($I(X,Y) = H(X)$), and in the case where the input and output are statistically independent, all information is "lost" ($I(X,Y) = 0$).
2. $H(X|Y)$ can be interpreted as the information that is "lost" in the channel, and thus the information that "passes" the channel, $I(X,Y)$, is equal to the information at the input, $H(X)$, minus the information that is lost, $H(X|Y)$. When the probability of error is zero, the loss is zero, and when the input and output are statistically independent, the loss is total, that is, equal to the information at the input of the channel.

These conclusions can be extended to any DMC with input and output alphabets of M_X and M_Y symbols, respectively.

However, the mutual information between the input and the output of the channel depends on the probability distribution at the input. If we want to know what is the maximum amount of information capable of passing through a certain channel, it is necessary to consider all the possible distributions and the one that produces the greatest mutual information will be the optimal one, and the mutual information for that distribution will be the maximum information capable of passing through the channel, that is, the *channel capacity*.

Formally, the channel capacity, C, of a DMC is defined as

$$C = \max_{p_X(x_i)} I(X,Y).$$

Its units are bits (or bits per channel use).

Some properties of the channel capacity, which derive from its definition through mutual information, are those that define the bounds for its minimum and maximum value:

1. $C \geq 0$, because $I(X,Y) \geq 0$.
2. $C \leq \log M_X$, because $C = \max I(X,Y) \leq \max H(X) = \log M_X$.
3. $C \leq \log M_Y$, for the same reason.

For some simple cases, such as the binary symmetric channel, it is possible to directly calculate the channel capacity. For the BSC, mutual information is

$$\begin{aligned}
I(X,Y) =& H(Y) - H(Y|X) \\
=& H(Y) - \sum_{i=0}^{1} p_X(x_i) H(Y|X=x_i) \\
=& H(Y) - \sum_{i=0}^{1} p_X(x_i) \left(- \sum_{j=0}^{1} p_{Y|X}(y_j|x_i) \log p_{Y|X}(y_j|x_i) \right) \\
=& H(Y) - \sum_{i=0}^{1} p_X(x_i)(-\varepsilon \log(\varepsilon) - (1-\varepsilon)\log(1-\varepsilon)) \\
=& H(Y) - \sum_{i=0}^{1} p_X(x_i) H_b(\varepsilon) \\
=& H(Y) - H_b(\varepsilon).
\end{aligned}$$

This is the same previously obtained result. The maximum of this mutual information is obtained, given that the parameter ε is fixed and cannot be acted upon, when the entropy $H(Y)$ is maximum. This occurs for an equiprobable output distribution, which for the BSC is equivalent to having an equiprobable input distribution. In this case, $H(Y) = 1$ and the channel capacity is therefore

$$C = 1 - H_b(\varepsilon).$$

It is easy to represent this dependency with the error probability considering the variation form of $H_b(\varepsilon)$. In a channel with zero error probability we can transmit one bit per channel use, while in a channel with error probability 1/2 no information can be sent.

It must be taken into account that the problem of calculating the channel capacity in general can be posed as a constraint maximization problem. Maximization of mutual information, which depends on the input probabilities, with the restrictions imposed by these probabilities, since the following $2M_X + 1$ restrictions must always be fulfilled:

- $p_X(x_i) \geq 0$, for $i = 0, 1, \cdots, M_X - 1$.
- $p_X(x_i) \leq 1$, for $i = 0, 1, \cdots, M_X - 1$.
- $\sum_{i=0}^{M_X-1} p_X(x_i) = 1$.

In some simple channels it will be possible to calculate the channel capacity analytically. For other types of more complicated channels, sometimes it is not possible to easily find the solution analytically. In this case, numerical techniques are used, which sweep all the possible input distributions until finding the one that maximizes the mutual information.

Example

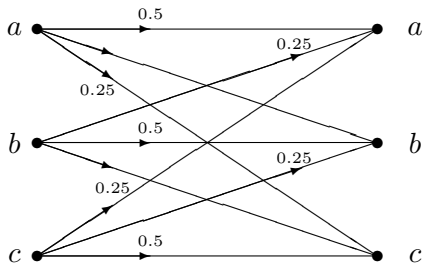


Figure 4.27: An example of DMC channel

The channel capacity of the channel shown in Figure 4.27 is obtained.

To obtain the capacity, the first thing is to obtain an expression of the mutual information. One option is through $H(Y)$ and $H(Y|X)$. From the input-output relationship, it can be seen that for the three cases, given $X = a$, $X = b$ or $X = c$, the output random variable Y has three symbols with probabilities 1/2, 1/4 and 1/4 respectively. Therefore,

$$H(Y|X = a) = H(Y|X = b) = H(Y|X = c) = \frac{1}{2}\log_2(2) + 2 \times \frac{1}{4}\log_2(4) = 1.5.$$

Since the same thing is obtained for all symbols, regardless of $P_X(x_i)$ we have that

$$H(Y|X) = 1.5.$$

Finally,

$$I(Y, X) = H(Y) - 1.5$$

To maximize $I(X, Y)$, one must maximize $H(Y)$, and this involves finding the probability distribution of the input random variable that produces equally likely outputs. In this case this happens for equally likely input symbols. In this case

$$H(Y) = \log_2(3) = 1.585.$$

Therefore, the channel capacity is

$$C = 1.585 - 1.5 = 0.085 \text{ bits/channel use.}$$

It is observed that it is relatively low, but it is normal if one takes into account that transmitting symbols through this channel fails 50% of the time.

4.4.3 Channel capacity for Gaussian channel

Having studied the digital channel and how its capacity is obtained, this section extends the study to the case of the Gaussian channel, where its output at a given instant is a continuous random variable instead of a discrete random variable. It is taken into account in this analysis that in a real communications system there are usually two important limitations on the use of resources:

- A limitation on the maximum power that can be transmitted, which is applied on the communications signal generated by the transmitter.
- A limitation on the bandwidth of the transmitted signal.

Therefore, in this section we will study what is the maximum amount of information that can be reliably transmitted in a digital communications system over a Gaussian channel when the maximum power of the transmitted signal is limited to P_X watts and the available bandwidth is B Hz. As in the previous section, the analysis will be performed in two ways. On the one hand, a more intuitive demonstration will be made in terms of the number of symbols that can be transmitted with low probability of overlap of their outputs, along the lines of the treatment followed in [Proakis and Salehi, 2002], and on the other hand, the channel capacity will also be obtained by a development based on information theory, along the lines of the analysis performed in [Artés-Rodríguez et al., 2007].

Gaussian channel capacity: intuitive proof

In a Gaussian channel the output of the channel is modeled by a random variable Y which is related to the input, modeled by a random variable X, through the additive relationship

$$Y = X + Z,$$

where Z is a random variable that models the effect of noise, with a Gaussian distribution, zero mean and variance (noise power) P_Z.

The idea to obtain the channel capacity is similar to the one followed in the digital channel. We are going to sample the transmitted and received signals at n time instants with the objective of seeing what is the maximum number of possible values of the transmitted signal that gives rise in the output to values with low probability of overlapping in the limit when n tends to infinity. The power constraint on the transmitted signal implies that if one has n realizations of the random variable X, which are grouped into a vector $\mathbf{x}$.

$$\mathbf{x} = \{x_1, x_2, \cdots, x_n\},$$

which would model the value of the transmitted signal at n instants. For a sufficiently large n value, the following is true

$$\frac{1}{n}\sum_{i=1}^{n} x_i^2 \leq P_X,$$

where P_X is the signal power. In this case x_i does not denote the alphabet of a discrete random variable, but the n realizations of the continuous random variable X that models the amplitude of the transmitted signal, $s(t)$, at n time instants.

For blocks of length n of input, output and noise values, grouped in the vectors $\mathbf{x}$, $\mathbf{y}$ and $\mathbf{z}$, respectively, one can write the vector relation

$$\mathbf{y} = \mathbf{x} + \mathbf{z}.$$

If n is sufficiently large, by the law of large numbers, the power limitation of the noise term implies the restriction

$$\frac{1}{n}\sum_{i=1}^{n} z_i^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - x_i)^2 \leq P_Z.$$

Finally, given the independence between X and Z, the power of Y will be the sum of the powers of X and Z, i.e. $P_Y = P_X + P_Z$, so that over the n samples of the output we will have the restriction

$$\frac{1}{n}\sum_{i=1}^{n} y_i^2 \leq P_Y = P_X + P_N.$$

Bearing in mind that the sum of squared coordinates of a vector gives its squared norm, the constraints imposed by the transmitted signal power and noise power can be written as follows

$$||\mathbf{x}||^2 \leq nP_X,\ ||\mathbf{y}-\mathbf{x}||^2 \leq nP_N,\ ||\mathbf{y}||^2 \leq n(P_X+P_N).$$

The geometric interpretation of the restrictions imposed by the power of the transmitted signal and the power of the noise signal leads to the following conclusions when n increases asymptotically

- The vector representation of the n samples of the transmitted signal, $\mathbf{x}$, is located in an n-dimensional hypersphere of radius $\sqrt{nP_X}$ centered on the origin.
- The vector representation of the n samples of the output signal, $\mathbf{y}$, lies in an n-dimensional hypersphere of radius $\sqrt{nP_Z}$ and centered around the vector representation of the n samples of the transmitted signal, $\mathbf{x}$.
- The vector representation of the n samples of the received signal, $\mathbf{y}$, is located in an n-dimensional hypersphere of radius $\sqrt{n(P_X+P_Z)}$ centered on the origin.

Making use of this geometric interpretation, illustrate in Figure 4.28,the calculation of the channel capacity is equivalent to finding how many different sequences of n samples of the transmitted signal $\mathbf{x}$ can be obtained in such a way that the outputs they give rise to do not overlap on the output space. Obviously, if this condition is met, then the output streams can be reliably decoded. Therefore, the question to be answered to establish the value of the channel capacity is: *How many spheres of radius* $\sqrt{nP_Z}$ *can be packed in a sphere of radius* $\sqrt{n(P_X+P_Z)}$*?*

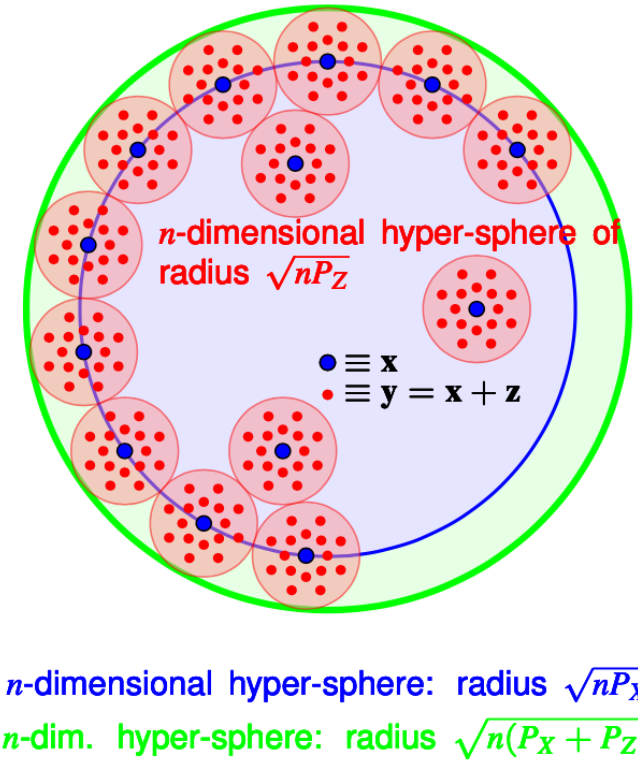


Figure 4.28: Geometric interpretation of the range of the n samples in the output when a certain value is transmitted for the n samples of the input for the calculation of the capacity of a Gaussian channel.

The answer, in an approximate way, can be obtained by means of the relation between the volumes of the two hyper-spheres. The volume of a hyper-sphere of dimension n and radius r is proportional to its radius raised to the power n,

$$V_n = K_n r^n,$$

where K_n is a radius independent constant that depends on the dimension of space. In this case, the maximum number of symbols, understood as blocks of n samples of the transmitted signal, that can be faithfully transmitted through a Gaussian channel is approximated by the quotient between the volume of the sphere in which the representation of the n samples of the channel output is contained, which has radius $\sqrt{n(P_X + P_Z)}$, and the volume of the sphere in which the representation of the output is contained when a certain symbol $\mathbf{x}$ has been transmitted, which has radius $\sqrt{nP_Z}$. This quotient is therefore

$$\begin{aligned} M_{no} =& \frac{K_n(n(P_X + P_Z))^{n/2}}{K_n(nP_Z)^{n/2}} = \left(\frac{P_X + P_Z}{P_Z}\right)^{n/2} \\ =& \left(1 + \frac{P_X}{P_Z}\right)^{n/2}. \end{aligned}$$

Since $\log_2 M_{no}$ bits of information can be encoded with M_{no} non-overlapping sequences, the capacity of the Gaussian channel with power restriction P_X and noise power P_Z is given by the quotient between this number of bits and the number of uses of the channel, which will be n, that is to say

$$\begin{aligned} C =& \frac{\log_2 M_{no}}{n} = \frac{1}{n}\frac{n}{2}\log_2\left(1 + \frac{P_X}{P_Z}\right) \\ =& \frac{1}{2}\log_2\left(1 + \frac{P_X}{P_Z}\right). \end{aligned}$$

Now we must remember that when working with a bandwidth limitation, the received signal is filtered in the receiver with an ideal filter of bandwidth B Hz to minimize the effect of the noise, so that the power of the filtered noise is

$$P_Z = N_0 B.$$

Therefore, the capacity of the Gaussian channel is

$$C = \frac{1}{2}\log_2\left(1 + \frac{P_X}{N_0 B}\right) \text{ bits/use.}$$

If this result is multiplied by the number of uses (transmissions) per second, which according to Nyquist's theorem is $2B$, the channel capacity in bits/s is obtained,

$$C = B\log_2\left(1 + \frac{P}{N_o B}\right) \text{ bits/s.}$$

This is the well-known Shannon formula for the capacity of a channel with additive white and Gaussian noise.

Gaussian channel capacity through mutual information

The same result can also be reached through a more formal derivation based on information theory. In the Gaussian channel model the relationship between input and output is given by

$$Y = X + Z,$$

where Z is a Gaussian random variable of zero mean and variance $P_Z = N_0 B$. Thus the conditional distribution of Y given X, $f_{Y|X}(y|x)$, is a Gaussian distribution of mean x and variance $\sigma^2 = P_Z$. The channel capacity will be obtained through the mutual information in the same way as for the digital channel, maximizing the mutual information between the input and output of the channel, but including the constraint on the maximum power for the transmitted signal (without this constraint, the capacity would be theoretically infinite). In other words, the channel capacity is defined as

$$C = \max_{f_X(x)|E[X^2]\leq P_X} I(X,Y),$$

where the power constraint is given by the constraint $E[X^2] \leq P_X$.

The mutual information between input and output can be calculated through its relation to the differential entropies

$$I(X,Y) = h(Y) - h(Y|X) = h(X+Z) - h(Z).$$

The differential entropy of the noise is calculated in a very simple way, since its distribution is Gaussian with zero mean and variance $\sigma^2 = P_Z$, a case that has already been considered previously, so that

$$h(Z) = \int_{-\infty}^{\infty} f_Z(z) \log_2 \frac{1}{f_Z(z)}\, dz = \frac{1}{2}\log_2 2\pi e P_Z.$$

This differential entropy depends only on the variance of the noise term, $P_Z = N_0 B$. Therefore, the mutual information can be written as follows

$$I(X,Y) = h(X+Z) - \frac{1}{2}\log_2 2\pi e P_Z.$$

To obtain its maximum, the following property on the differential entropy will be used:

- If a random variable has a fixed value of variance, the probability density function that makes its differential entropy maximum is the Gaussian distribution.

The demonstration of this property can be seen, for example, in [Artés-Rodríguez et al., 2007], Chapter 9, page 565.

Since X and Z are statistically independent and Z has zero mean, the variance of Y is the sum of the variances of X and Y, so that

$$E[Y^2] = E[(X+Z)^2] = E[X^2] + E[Z^2] \leq P_X + P_Z.$$

Since the variance of Y is bounded, the capacity is reached when Y has a Gaussian distribution, in which case the maximum value of the mutual information, and therefore the capacity, is

$$C = \frac{1}{2}\log_2 2\pi e(P_X + P_Z) - \frac{1}{2}\log_2 2\pi e\sigma^2 = \frac{1}{2}\log_2\left(1 + \frac{P_X}{P_Z}\right) \text{ bits/use.}$$

This result is the same as that obtained previously.

4.5 Limits of a digital communications system

In the previous section the channel capacity for the Gaussian channel has been obtained.

$$C = B\ \log_2\left(1 + \frac{P_X}{N_0 B}\right) \text{ bits/s.}$$

In addition to N_0, the channel capacity depends on two relevant parameters of any communication system: the power of the transmitted signal, P_X, and the available bandwidth, B Hz. In this section we will first analyze the dependence of the channel capacity on these two parameters of the communications system.

As for the transmitted signal power, the analysis is straightforward. If the transmitted power is increased, the channel capacity increases: obviously, the higher the power, the more levels (or symbols of length n samples) can be put, and the more bits/usage are possible. However, it should be noted that the increase follows a logarithmic law, so that in order to obtain a linear increase in capacity, an exponential increase in transmitted power is required. In any case, it is theoretically possible to increase the capacity to infinity by increasing the power of the transmitted signal.

The effect of bandwidth on channel capacity is different. Increasing B has two opposing effects. On the one hand, higher bandwidth increases the transmission rate, but on the other hand it increases the noise level and thus reduces performance. By taking B to the infinite limit, and applying L'Hopital's rule, we obtain the limit of the channel capacity when the bandwidth tends to infinity

$$\lim_{B\to\infty} C = \frac{P_X}{N_0}\log_2(e) = 1.44\frac{P_X}{N_0}.$$

This result means that, unlike with the power of the transmitted signal, increasing the channel bandwidth alone cannot increase the capacity to any desired value, but there is a maximum achievable limit that depends on the signal to noise ratio (P_X/N_0).

Shannon's channel coding theorem establishes a maximum limit on the information transmission rate with a digital communication system. This means that in a practical communication system, it must always be satisfied that the effective transmission rate (defined over the information bits) is below the channel capacity, $R < C$. For the case of the Gaussian channel this implies that the bit rate, R_b bits/s, has to be below the channel capacity. Defining the signal-to-noise ratio as the ratio between the transmitted signal power and N_0

$$SNR = \frac{P_X}{P_Z} = \frac{P_X}{N_0 B},$$

this means that the following relationship must be fulfilled

$$R_b < B\log\left(1+\frac{P}{N_0 B}\right) \text{ bits/s}.$$

Dividing both sides of this inequality by B, defining the spectral bit rate, or spectral efficiency, as

$$\eta = \frac{R_b}{B} \text{ bits/s/Hz},$$

the following limit for the spectral efficiency in a practical communications system is obtained

$$\eta < \log_2\left(1+\frac{P_X}{N_0 B}\right).$$

If the average energy per bit is now defined as the ratio between the power of the transmitted signal and the bit rate of transmission

$$E_b = \frac{P_X}{R_b},$$

and the corresponding relation among E_b and N_0 is,

$$\frac{E_b}{N_0} = \frac{SNR}{\eta},$$

the spectral binary rate or spectral efficiency can be rewritten in terms of this relationship as

$$\eta < \log_2\left(1 + \eta\frac{E_b}{N_0}\right).$$

This constraint relating the signal-to-noise ratio and the spectral efficiency of the system is written in the literature with several equivalent alternative expressions, for example

$$\frac{E_b}{N_0} > \frac{2^\eta - 1}{\eta}.$$

This expression indicates that there is a minimum E_b/N_0 ratio that is necessary in order to have reliable communication. In the case of this expression, if its limit is computed when η tends to infinity we have

$$\lim_{\eta\to\infty}\frac{E_b}{N_0} = \ln 2 = 0.693 \approx -1.6 \text{ dB}.$$

Figure 4.29, splits the η vs E_b/N_0 plane in two regions. In a region, below the curve, reliable communication is possible (understood as communication in which the use of channel coding techniques allows the probability of error to be reduced to arbitrarily low levels). In the other region, above the curve, it is not possible to have reliable communication. The performance of any system can be represented by a point in the plane of this curve. The closer the point is to the curve, the higher the efficiency of the system.

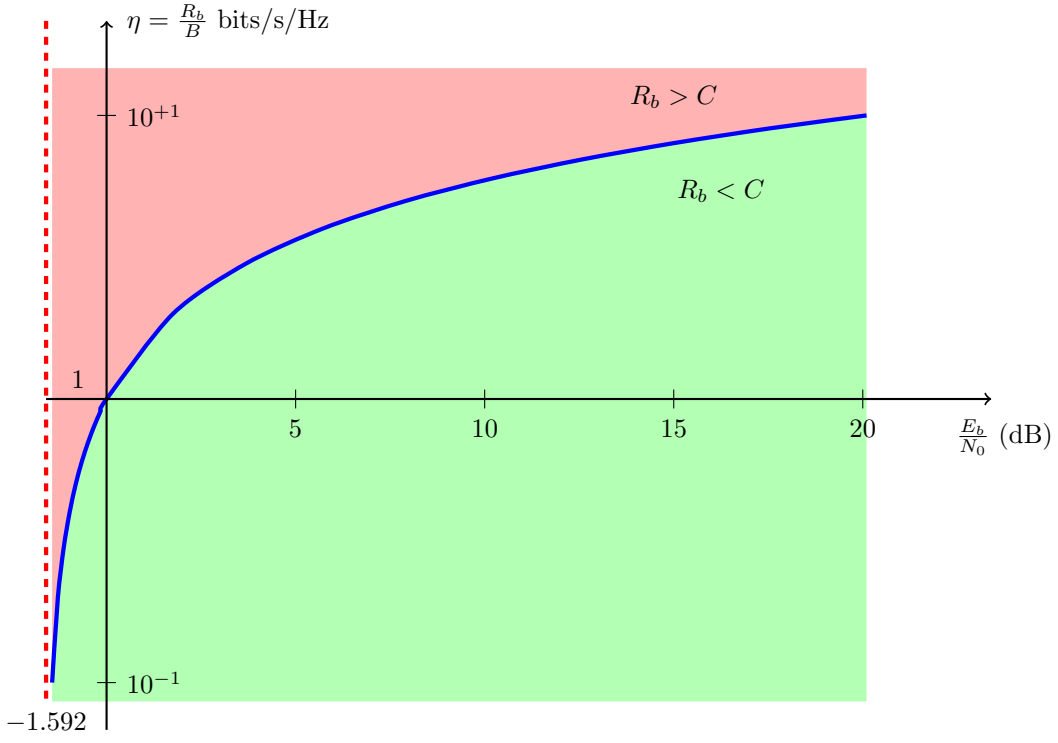


Figure 4.29: Spectral binary rate (spectral efficiency) η (bits/s/Hz) versus ratio E_b/N_0 for a Gaussian channel. The region where reliable communication is possible (in green) is distinguished from the region where it is not possible (in red).

It can be seen that this curve, when η tends to zero, the ratio E_b/N_0 tends to the value calculated above

$$\frac{E_b}{N_0} = \ln 2 = 0.693 \approx -1.6 \text{ dB}.$$

This is an absolute minimum for reliable communication, i.e., in order to have reliable communication, the ratio E_b/N_0 has to be above this limit.

In this figure, two things can also be seen regarding the value of η:

1. When $\eta << 1$, the bandwidth is large and the only limitation is power. In this case we speak of *power-limited systems*. In this case it is necessary to use simple constellations.
2. When $\eta >> 1$ the channel bandwidth is small and it is called *bandwidth-limited systems*. In this case, very dense constellations are used (eg 256-QAM).

The previous expression indicates a minimum limit that the ratio between the energy of the signal and that of the noise must reach in order to be able to transmit reliably with a certain spectral efficiency. Although in the literature this limit is usually expressed through the relationship E_b/N_0, on other occasions it is expressed from the signal-to-noise relationship, in which case the resulting expression is

$$SNR > 2^{\eta} - 1.$$

Given this minimum SNR limit, the so-called "*normalized signal-to-noise ratio*" is sometimes defined, which basically compares the working signal-to-noise ratio to that minimum required level.

$$SNR_{norm} = \frac{SNR}{2^{\eta} - 1}.$$

By the definition of this normalized signal to noise ratio, in a practical system it must take values greater than unity, i.e., greater than 0 decibels. It is therefore a relative measure that indicates how close or far the system is from the limit operating value of the reliable zone.

Appendix A

Tables of interest

Some tables of interest in the subject are shown in this appendix. Specifically:

- Pairs and properties of the Fourier transform.
- Function $Q(x)$, especially useful in calculating error probabilities.
- Relationships and integrals for some trigonometric functions.

Time Domain ($x(t)$)	Frequency Domain ($X(j\omega)$)
$\delta(t)$	1
1	$2\pi\ \delta(\omega)$
$\delta(t-t_0)$	$e^{-j\omega t_0}$
$e^{j\omega_0 t}$	$2\pi\ \delta(\omega-\omega_0)$
$\cos(\omega_0 t)$	$\pi\ \delta(\omega-\omega_0)+\pi\ \delta(\omega+\omega_0)$
$\sin(\omega_0 t)$	$\frac{\pi}{j}\ \delta(\omega-\omega_0)-\frac{\pi}{j}\ \delta(\omega+\omega_0)$
$\Pi\left(\frac{t}{T}\right)=\begin{cases}1, & \lvert t\rvert\le\frac{T}{2}\\ 0, & \lvert t\rvert>\frac{T}{2}\end{cases}$	$T\ \mathrm{sinc}\left(\frac{\omega T}{2\pi}\right)$
$\mathrm{sinc}\left(\frac{t}{T}\right)$	$T\ \Pi\left(\frac{\omega T}{2\pi}\right)$
$\Lambda\left(\frac{t}{T}\right)=\begin{cases}1-\frac{\lvert t\rvert}{T}, & \lvert t\rvert\le T\\ 0 & \lvert t\rvert>T\end{cases}$	$T\ \mathrm{sinc}^2\left(\frac{\omega T}{2\pi}\right)$
$\mathrm{sinc}^2\left(\frac{t}{T}\right)$	$T\ \Lambda\left(\frac{\omega T}{2\pi}\right)$
$u(t)$	$\frac{1}{j\omega}+\pi\ \delta(\omega)$
$\frac{1}{2}\ \delta(t)+j\ \frac{1}{2\pi t}$	$u(\omega)$
$e^{-\alpha t}\ u(t),\ \alpha>0$	$\frac{1}{\alpha+j\omega}$
$t\ e^{-\alpha t}\ u(t),\ \alpha>0$	$\frac{1}{(\alpha+j\omega)^2}$
$\frac{t^{n-1}}{(n-1)!}\ e^{-\alpha t}\ u(t),\ \alpha>0$	$\frac{1}{(\alpha+j\omega)^n}$
$e^{-\alpha\lvert t\rvert}$	$\frac{2\alpha}{\alpha^2+\omega^2}$
$e^{-\pi t^2}$	$e^{-\pi f^2}=e^{-\frac{w^2}{4\pi}}$
$\mathrm{sgn}(t)=\begin{cases}1, & t>0\\ -1, & t<0\\ 0, & t=0\end{cases}$	$\frac{2}{j\omega}$
$\delta'(t)$	$j\omega$
$\delta^{(n)}(t)$	$(j\omega)^n$
$\frac{1}{t}$	$-j\pi\ \mathrm{sgn}(\omega)$
$\sum_{n=-\infty}^{\infty}\delta(t-nT)$	$\frac{2\pi}{T}\sum_{k=-\infty}^{\infty}\delta\left(\omega-\frac{2\pi k}{T}\right)$

Table A.1: Fourier transform pairs (in $j\omega$).

Time Domain ($x(t)$)	Frequency Domain ($X(j\omega)$)
$a\ x(t) + b\ y(t)$	$a\ X(j\omega) + b\ Y(j\omega)$
$x(t - t_0)$	$e^{-j\omega t_0}\ X(j\omega)$
$e^{j\omega_0 t}\ x(t)$	$X(\omega - \omega_0)$
$x^*(t)$	$X^*(-\omega)$
$x(-t)$	$X(-\omega)$
$x(a\ t)$	$\frac{1}{\lvert a\rvert} X\left(\frac{\omega}{a}\right)$
$x(t) * y(t)$	$X(j\omega)\ Y(j\omega)$
$x(t)\ y(t)$	$\frac{1}{2\pi}\ X(j\omega) * Y(j\omega)$
$\frac{d}{dt}x(t)$	$j\omega\ X(j\omega)$
$\int_{-\infty}^{t} x(t)\ dt$	$\frac{1}{j\omega}\ X(\omega) + \pi\ X(0)\ \delta(\omega)$
$t\ x(t)$	$j\frac{d}{d\omega}X(j\omega)$
$x(t)$ real	$X(j\omega) = X^*(-j\omega)$

Parseval relationship for non-periodic signals

$$\int_{-\infty}^{\infty} |x(t)|^2\ dt = \frac{1}{2\pi}\int_{-\infty}^{\infty} |X(j\omega)|^2\ d\omega$$

Duality property

$$f(u) = \int_{-\infty}^{\infty} g(v)\ e^{-juv}\ dv$$

$$g(t) \overset{TF}{\leftrightarrow} f(\omega)$$

$$f(t) \overset{TF}{\leftrightarrow} 2\pi\ g(-\omega)$$

Table A.2: Properties of the Fourier Transform.

x	$Q(x)$	x	$Q(x)$	x	$Q(x)$	x	$Q(x)$
0.00	5.00000×10^{-1}	2.05	2.01822×10^{-2}	4.10	2.06575×10^{-5}	6.15	3.87415×10^{-10}
0.05	4.80061×10^{-1}	2.10	1.78644×10^{-2}	4.15	1.66238×10^{-5}	6.20	2.82316×10^{-10}
0.10	4.60172×10^{-1}	2.15	1.57776×10^{-2}	4.20	1.33457×10^{-5}	6.25	2.05226×10^{-10}
0.15	4.40382×10^{-1}	2.20	1.39034×10^{-2}	4.25	1.06885×10^{-5}	6.30	1.48823×10^{-10}
0.20	4.20740×10^{-1}	2.25	1.22245×10^{-2}	4.30	8.53991×10^{-6}	6.35	1.07657×10^{-10}
0.25	4.01294×10^{-1}	2.30	1.07241×10^{-2}	4.35	6.80688×10^{-6}	6.40	7.76885×10^{-11}
0.30	3.82089×10^{-1}	2.35	9.38671×10^{-3}	4.40	5.41254×10^{-6}	6.45	5.59251×10^{-11}
0.35	3.63169×10^{-1}	2.40	8.19754×10^{-3}	4.45	4.29351×10^{-6}	6.50	4.01600×10^{-11}
0.40	3.44578×10^{-1}	2.45	7.14281×10^{-3}	4.50	3.39767×10^{-6}	6.55	2.87685×10^{-11}
0.45	3.26355×10^{-1}	2.50	6.20967×10^{-3}	4.55	2.68230×10^{-6}	6.60	2.05579×10^{-11}
0.50	3.08538×10^{-1}	2.55	5.38615×10^{-3}	4.60	2.11245×10^{-6}	6.65	1.46547×10^{-11}
0.55	2.91160×10^{-1}	2.60	4.66119×10^{-3}	4.65	1.65968×10^{-6}	6.70	1.04210×10^{-11}
0.60	2.74253×10^{-1}	2.65	4.02459×10^{-3}	4.70	1.30081×10^{-6}	6.75	7.39226×10^{-12}
0.65	2.57846×10^{-1}	2.70	3.46697×10^{-3}	4.75	1.01708×10^{-6}	6.80	5.23096×10^{-12}
0.70	2.41964×10^{-1}	2.75	2.97976×10^{-3}	4.80	7.93328×10^{-7}	6.85	3.69250×10^{-12}
0.75	2.26627×10^{-1}	2.80	2.55513×10^{-3}	4.85	6.17307×10^{-7}	6.90	2.60013×10^{-12}
0.80	2.11855×10^{-1}	2.85	2.18596×10^{-3}	4.90	4.79183×10^{-7}	6.95	1.82643×10^{-12}
0.85	1.97663×10^{-1}	2.90	1.86581×10^{-3}	4.95	3.71067×10^{-7}	7.00	1.27981×10^{-12}
0.90	1.84060×10^{-1}	2.95	1.58887×10^{-3}	5.00	2.86652×10^{-7}	7.05	8.94589×10^{-13}
0.95	1.71056×10^{-1}	3.00	1.34990×10^{-3}	5.05	2.20905×10^{-7}	7.10	6.23784×10^{-13}
1.00	1.58655×10^{-1}	3.05	1.14421×10^{-3}	5.10	1.69827×10^{-7}	7.15	4.33890×10^{-13}
1.05	1.46859×10^{-1}	3.10	9.67603×10^{-4}	5.15	1.30243×10^{-7}	7.20	3.01063×10^{-13}
1.10	1.35666×10^{-1}	3.15	8.16352×10^{-4}	5.20	9.96443×10^{-8}	7.25	2.08386×10^{-13}
1.15	1.25072×10^{-1}	3.20	6.87138×10^{-4}	5.25	7.60496×10^{-8}	7.30	1.43884×10^{-13}
1.20	1.15070×10^{-1}	3.25	5.77025×10^{-4}	5.30	5.79013×10^{-8}	7.35	9.91034×10^{-14}
1.25	1.05650×10^{-1}	3.30	4.83424×10^{-4}	5.35	4.39771×10^{-8}	7.40	6.80922×10^{-14}
1.30	9.68005×10^{-2}	3.35	4.04058×10^{-4}	5.40	3.33204×10^{-8}	7.45	4.66701×10^{-14}
1.35	8.85080×10^{-2}	3.40	3.36929×10^{-4}	5.45	2.51849×10^{-8}	7.50	3.19089×10^{-14}
1.40	8.07567×10^{-2}	3.45	2.80293×10^{-4}	5.50	1.89896×10^{-8}	7.55	2.17629×10^{-14}
1.45	7.35293×10^{-2}	3.50	2.32629×10^{-4}	5.55	1.42835×10^{-8}	7.60	1.48065×10^{-14}
1.50	6.68072×10^{-2}	3.55	1.92616×10^{-4}	5.60	1.07176×10^{-8}	7.65	1.00490×10^{-14}
1.55	6.05708×10^{-2}	3.60	1.59109×10^{-4}	5.65	8.02239×10^{-9}	7.70	6.80331×10^{-15}
1.60	5.47993×10^{-2}	3.65	1.31120×10^{-4}	5.70	5.99037×10^{-9}	7.75	4.59463×10^{-15}
1.65	4.94715×10^{-2}	3.70	1.07800×10^{-4}	5.75	4.46217×10^{-9}	7.80	3.09536×10^{-15}
1.70	4.45655×10^{-2}	3.75	8.84173×10^{-5}	5.80	3.31575×10^{-9}	7.85	2.08019×10^{-15}
1.75	4.00592×10^{-2}	3.80	7.23480×10^{-5}	5.85	2.45787×10^{-9}	7.90	1.39452×10^{-15}
1.80	3.59303×10^{-2}	3.85	5.90589×10^{-5}	5.90	1.81751×10^{-9}	7.95	9.32558×10^{-16}
1.85	3.21568×10^{-2}	3.90	4.80963×10^{-5}	5.95	1.34071×10^{-9}	8.00	6.22096×10^{-16}
1.90	2.87166×10^{-2}	3.95	3.90756×10^{-5}	6.00	9.86588×10^{-10}		
1.95	2.55881×10^{-2}	4.00	3.16712×10^{-5}	6.05	7.24229×10^{-10}		
2.00	2.27501×10^{-2}	4.05	2.56088×10^{-5}	6.10	5.30342×10^{-10}		

Table A.3: Function $Q(x)$.

Trigonometric relationships

$$\sin^2(a) + \cos^2(a) = 1$$
$$\tan^2(a) + 1 = \sec^2(a)$$
$$\cot^2(a) + 1 = \csc^2(a)$$
$$\cos(a)\ \cos(b) = \frac{1}{2}\ \cos(a-b) + \frac{1}{2}\ \cos(a+b)$$
$$\sin(a)\ \sin(b) = \frac{1}{2}\ \cos(a-b) - \frac{1}{2}\ \cos(a+b)$$
$$\cos(a)\ \sin(b) = \frac{1}{2}\ \sin(a+b) - \frac{1}{2}\ \sin(a-b)$$
$$\sin(a \pm b) = \sin(a)\ \cos(b) \pm \cos(a)\ \sin(b)$$
$$\cos(a \pm b) = \cos(a)\ \cos(b) \mp \sin(a)\ \sin(b)$$
$$\sin(a) + \sin(b) = 2\ \sin\left(\frac{a+b}{2}\right)\ \cos\left(\frac{a-b}{2}\right)$$
$$\cos(a) + \cos(b) = 2\ \cos\left(\frac{a+b}{2}\right)\ \cos\left(\frac{a-b}{2}\right)$$
$$\sin(a) - \sin(b) = 2\ \cos\left(\frac{a+b}{2}\right)\ \sin\left(\frac{a-b}{2}\right)$$
$$\cos(a) - \cos(b) = -2\ \sin\left(\frac{a+b}{2}\right)\ \sin\left(\frac{a-b}{2}\right)$$
$$\sin^2(a) - \sin^2(b) = \sin(a+b)\ \sin(a-b)$$
$$\cos^2(a) - \sin^2(b) = \cos(a+b)\ \cos(a-b)$$
$$\cos(a) = \frac{e^{+ja} + e^{-ja}}{2},\ \ \sin(a) = \frac{e^{+ja} - e^{-ja}}{2j}$$
$$e^{ja} = \cos(a) + j\sin(a)$$

Integrals of some trigonometric functions

$$\int \cos(a\ t)\ dt = \frac{1}{a}\sin(a\ t)$$
$$\int \sin(a\ t)\ dt = -\frac{1}{a}\cos(a\ t)$$
$$\int \cos^2(a\ t)\ dt = \frac{t}{2} + \frac{1}{2a}\ \cos(a\ t)\ \sin(a\ t) = \frac{t}{2} + \frac{1}{4a}\ \sin(2a\ t)$$
$$\int \sin^2(a\ t)\ dt = \frac{t}{2} - \frac{1}{2a}\ \cos(a\ t)\ \sin(a\ t) = \frac{t}{2} - \frac{1}{4a}\ \sin(2a\ t)$$
$$\int \cos(a\ t)\ \cos(b\ t)\ dt = \frac{1}{2}\ \frac{\sin((a-b)\ t)}{a-b} + \frac{1}{2}\ \frac{\sin((a+b)\ t)}{a+b}$$
$$\int \sin(a\ t)\ \sin(b\ t)\ dt = \frac{1}{2}\ \frac{\sin((a-b)\ t)}{a-b} - \frac{1}{2}\ \frac{\sin((a+b)\ t)}{a+b}$$
$$\int \cos(a\ t)\ \sin(b\ t)\ dt = \frac{1}{2}\ \frac{\cos((a-b)\ t)}{a-b} - \frac{1}{2}\ \frac{\cos((a+b)\ t)}{a+b}$$

Table A.4: Relationships and integrals for some trigonometric functions.

Time domain ($x(t)$)	Frequency domain ($X(f)$)
$\delta(t)$	1
1	$\delta(f)$
$\delta(t-t_0)$	$e^{-j2\pi f t_0}$
$e^{j2\pi f_0 t}$	$\delta(f-f_0)$
$\cos(2\pi f_0 t)$	$\frac{1}{2}\delta(f-f_0)+\frac{1}{2}\delta(f+f_0)$
$\sin(2\pi f_0 t)$	$\frac{1}{2j}\delta(f-f_0)-\frac{1}{2j}\delta(f+f_0)$
$\Pi(t)=\begin{cases}1, & \lvert t\rvert<\frac{1}{2}\\ \frac{1}{2}, & t=\pm\frac{1}{2}\\ 0, & \text{in other case}\end{cases}$	$\mathrm{sinc}(f)$
$\mathrm{sinc}(t)$	$\Pi(f)$
$\Lambda(t)=\begin{cases}t+1, & -1\le t<0\\ -t+1, & 0\le t<1\\ 0, & \text{in other case}\end{cases}$	$\mathrm{sinc}^2(f)$
$\mathrm{sinc}^2(f)$	$\Lambda(f)$
$e^{-\alpha t}u(t),\ \alpha>0$	$\frac{1}{\alpha+j2\pi f}$
$te^{-\alpha t}u(t),\ \alpha>0$	$\frac{1}{(\alpha+j2\pi f)^2}$
$e^{-\alpha\lvert t\rvert}$	$\frac{2\alpha}{\alpha^2+(2\pi f)^2}$
$e^{-\pi t^2}$	$e^{-\pi f^2}$
$\mathrm{sgn}(t)=\begin{cases}1, & t>0\\ -1, & t<0\\ 0, & t=0\end{cases}$	$\frac{1}{j\pi f}$
$u(t)$	$\frac{1}{2}\delta(f)+\frac{1}{j2\pi f}$
$\delta'(t)$	$j2\pi f$
$\delta^{(n)}(t)$	$(j2\pi f)^n$
$\frac{1}{t}$	$-j\pi\mathrm{sgn}(f)$
$\sum_{n=-\infty}^{\infty}\delta(t-nT_0)$	$\frac{1}{T_0}\sum_{n=-\infty}^{\infty}\delta\left(f-\frac{n}{T_0}\right)$

Table A.5: Fourier transform pairs (in f)

Bibliography

[Artés-Rodríguez et al., 2007] A. Artés-Rodríguez, F. Pérez-Gonzalez, J. Cid-Sueiro, R. López-Valcarce, C. Mosquera-Nartallo, and F. Pérez-Cruz. *Comunicaciones Digitales*. Pearson Educación. 2007.

[Proakis and Salehi, 2002] J. G. Proakis and M. Salehi. *Communication System Engineering*. Prentice-Hall, Upper Saddle River, New Jersey, 2nd edition. 2002.