Principles Of Digital Communications

Bixio Rimoldi
School of Computer and Communication Sciences
Ecole Polytechnique Fédérale de Lausanne (EPFL)
Switzerland

Version August, 2009

Contents

Chapter 1

Introduction and Objectives

The evolution of communication technology during the past few decades has been impressive. In spite of an enormous progress, many of the challenges still lay ahead of us. While any prediction of the next big technological revolution is likely to be wrong, it is safe to say that communication devices will become smaller, lighter, more powerful, more integrated, more ubiquitous, and more reliable than they are today. Perhaps one day the input/output interface and the communication/computation hardware will be separated. The former will be the only part that we will carry on us and it will communicate wirelessly with the latter. Perhaps the communication/computation hardware will be part of the infrastructure. It will be built into cars, trains, airplanes, public places, homes, offices, etc. With the input/output device that we carry around we will have virtually unlimited access to communication and computation facilities. Search engines may be much more powerful than they are today, giving instant access to any information digitally stored. The input/output device may contain all of our preferences so that, for instance, when we sit down in front of a computer, we see the environment that we like regardless of location (home, office, someone else's desk) and regardless of the hardware and operating system. The input device may also allow us to unlock doors and make payments, making keys, credit cards, and wallets obsolete. Getting there will require joint efforts from almost all branches of electrical engineering, computer science, and system engineering.

In this course we focus on the system aspects of digital communications. Digital communications is a rather unique field in engineering in which theoretical ideas have had an extraordinary impact on actual system design. Our goal is to get acquainted with some of these ideas. Hopefully, you will appreciate the way that many of the mathematical tools you have learned so far will turn out to be exactly what we need. These tools include probability theory, stochastic processes, linear algebra, and Fourier analysis.

We will focus on systems that consist of a single transmitter, a channel, and a receiver as shown in Figure 1.1. The channel filters the incoming signal and adds noise. The noise is Gaussian since it represents the contribution of various noise sources.[1] The filter in

[1] Individual noise sources do not necessarily have Gaussian statistics. However, due to the central limit

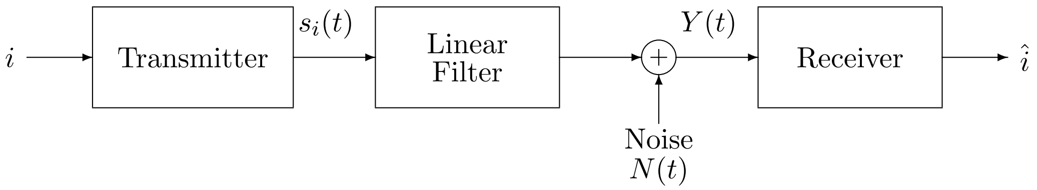


Figure 1.1: Basic point-to-point communication system over a bandlimited Gaussian channel.

the channel model has both a physical and a conceptual justification. The conceptual justification stems from the fact that most wireless communication systems are subject to a license that dictates, among other things, the frequency band that the signal is allowed to occupy. A convenient way for the system designer to deal with this constraint is to assume that the channel contains an ideal filter that blocks everything outside the intended band. The physical reason has to do with the observation that the signal emitted from the transmit antenna typically encounters obstacles that create reflections and scattering. Hence the receive antenna may capture the superposition of a number of delayed and attenuated replicas of the transmitted signal (plus noise). It is a straightforward exercise to check that this physical channel is linear and time-invariant. Thus it may be modeled by a linear filter as shown in the figure.[2] Additional filtering may occur due to the limitations of some of the components at the sender and/or at the receiver. For instance, this is the case of a linear amplifier and/or an antenna for which the amplitude response over the frequency range of interest is not flat and the phase response is not linear. The filter in Figure 1.1 accounts for all linear time-invariant transformations that act upon the communication signals as it travels from the sender to the receiver. The channel model of Figure 1.1 is meaningful for both wireline and wireless communication chanels. It is referred to as the bandlimited Gaussian channels.

Since communication means different things for different people, we need to clarify the role of the transmitter/receiver pair depicted in Figure 1.1. For the purpose of this class a transmitter implements a mapping between a message set and a signal set, both of the same cardinality, say m. The number m of elements of the message set is important but the nature of its elements is not. Typically we represent a message by an integer i between 0 and $m-1$ or, equivalently, by $\log m$ bits. During the first part of the course we will use integers to represent messages. There is a one-to-one correspondence between messages and elements of the signal set. The forms of the signals is important since signals have to be suitable to the channel. Intuitively, they should be as distinguishable as possible from the channel output. The channel model is always assumed to be given to the designer who has no control over it. By assumption, the designer can only control

theorem, their aggregate contribution is often quite well approximated by a Gaussian random process.

[2] If the scattering and reflecting objects move with respect to the transmit/receive antennae then the filter is time-varying but this case is deferred to the advanced digital communication class.

the design of the transmitter/receiver pair. A user communicates by selecting a message $i \in \{0, 1, \ldots, m-1\}$ which is converted by the transmitter into the corresponding signal s_i. The channel reacts to the signal by producing the observable y. Based on y, the receiver generates an estimate $\hat{\imath}(y)$ of i. Hence the receiver is a map from the space of channel output signals to the message set. Hopefully $i = \hat{\imath}$ most of the time. When this is not the case we say that an error event occurred. In all situations of interest to us it is not possible to reduce the probability of error to zero. This is so since, with positive probability, the channels is capable of producing an output y that could have stemmed from more than one message. One of the performance measures of a transmitter/receiver pair for a given channel is thus the probability of error. Another performance measure is the rate at which we communicate. Since we may label every message with a unique sequence of $\log m$ bits, we are sending the equivalent of $\log m$ bits every time we use the channel. By increasing the value of m we increase the rate in bits per channel use but, as we will see, under normal circumstances this increase can not be done indefinitely without increasing the probability of error.

At the end of this course you should have a good understanding of a basic communication system as depicted in Figure 1.1 and be able to make sensible design choices. In particular, you should know what a receiver does to minimize the probability of error, be able to do a quantitative analysis of some of the most important performance figures, understand the basic tradeoffs you have as a system designer, and appreciate the implications of such tradeoffs.

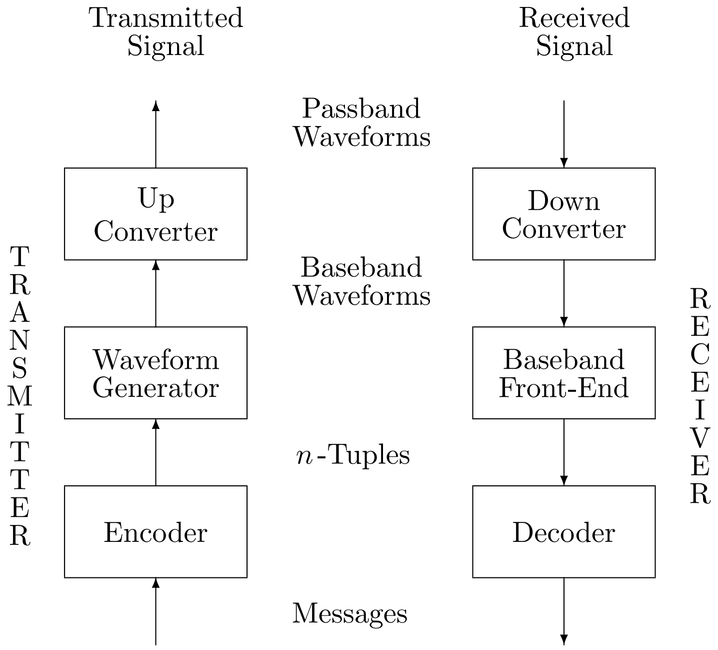


Figure 1.2: Decomposed transmitter and receiver.

A few words about the big picture and the approach that we will take are in order. We will discover that a natural way to design, analyze, and implement a transmitter/receiver pair for the Gaussian channels such as the one in Figure 1.1 (whether bandlimited or not) is in terms of the modules shown in Figure 1.2. These modules allow us to focus on selected issues while hiding others. For instance, at the very bottom level we exchange messages. At this level we may think of all modules as being inside a "black box" that hides all the implementation details and lets us see only what the user has to see from the outside. The "black box" is an abstract channel model that takes messages and delivers messages. The performance figures that are visible at this level of granularity are the cardinality m of the message set, the time T_m it takes to send a message, and the probability of error. The ratio $\frac{\log m}{T_m}$ is the rate $[\frac{\text{bits}}{\text{sec}}]$ at which we communicate. At the top level of Figure 1.2 we focus on the characteristics of the actual signals being sent over the physical medium, such as the average power of the transmitted signal and the frequency band it occupies. We will see that at the second level from the bottom we communicate n-tuples. It is at this level that we will understand the heart of the receiver. We will understand how the receiver should base its decision so as to minimize the probability of error and see how to compute the resulting error probability. Finally, one layer up we communicate using low-frequency (as opposed to radio frequency) signals. Separating the top two layers is important for implementation purposes.

There is more than one way to organize the discussion around the modules of Figure 1.2. Following the signal path, i.e., starting from the first module of the transmitter and working our way through the system until we deal with the final stage of the receiver would not be a good idea. This is so since it makes little sense to study the transmitter design without having an appreciation of the task and limitations of a receiver. More precisely, we would want to use signals that occupy a small bandwidth, have little power consumption, and that lead to a small probability of errors but we won't know how to compute the probability of error until we have studied the receiver design. We will instead make many passes over the block diagram of Figure 1.2, each time at a different level of abstraction, focussing on different issues as discussed in the previous paragraph, but each time considering the sender and the receiver together. We will start with the channel seen by the bottom modules in Figure 1.2. This approach has the advantage that you will quickly be able to appreciate what the transmitter and the receiver should do. One may argue that this approach has the disadvantage of asking the student to accept an abstract channel model that seems to be oversimplified. (It is not, but this will not be immediately clear). On the other hand one can also argue in favor of the pedagogical value of starting with highly simplified models. Shannon, the founding father of modern digital communication theory and one of the most profound engineers and mathematicians of the 20th century, was known to solve difficult problems by first reducing the problem to a much simpler version that he could almost solve "by inspection." Only after having familiarized himself with the simpler problem would he work his way back to the next level of difficulty. In this course we take a similar approach.

The choice of material covered in this course is by now more or less standard for an introductory course on digital communications. The approach depicted in Figure 1.2 has been made popular by J.M. Wozencraft and I. M. Jacobs in *Principles of Communication*

Engineering –a textbook appeared in 1965. However, the field has evolved since then and these notes reflect such evolution. Some of the exposition has benefited from the notes *Introduction to Digital Communication*, written by Profs. A. Lapidoth and R. Gallager for the MIT course Nr. 6.401/6.450, 1999. I am indebted to them for letting me use their notes during the first few editions of this course.

There is only so much that one can do in one semester. EPFL offers various possibilities for those who want to know more about digital communications and related topics. Classes for which this course is a recommended prerequisite are *Advanced Digital Communications*, *Information Theory and Coding*, *Principles of Diversity in Wireless Networks*, and *Coding Theory.* For the student interested in hands-on experience, EPFL offers *Software-Defined Radio: A Hands On Course.*

Networking is another branch of communications that has developed almost independently of the material treated in this class. It relies on quite a different set of mathematical models and tools. Networking assumes that there is a network of bit pipes which is reliable most of the time but that can fail once in a while. (How to create reliable bit pipes between network nodes is a main topic in this course). The network may fail due to network congestion, hardware failure, or queue overflow. Queues are used to temporarily store packets when the next link is congested. Networking deals with problems such as finding a route for a packet, computing the delay incurred by a packet as it goes from source to destination considering the queueing delay and the fact that packets are retransmitted if their reception is not acknowledged. We will not be dealing with networking problems in this class.

We conclude this introduction with a very brief overview of the various chapters. Not everything in this overview will make sense to you now. Nevertheless we advise you to read it now and read it again when you feel that it is time to step back and take a look at the "big picture." It will also give you an idea of which fundamental concepts will play a role in this course.

Chapter 2 deals with the *receiver design problem for discrete-time observations* with emphasis on that is seen by the bottom block of Figure 1.2. We will pay particular attention to the design of an optimal *decoder*, assuming that the encoder and the channel are given. The channel is the "black box" that contains everything above the two bottom boxes of Fig. 1.2. It takes and delivers n-tuples. Designing an optimal decoder is an application of what is know in the statistical literature as hypothesis testing (to be developed in Chapter 2). After a rather general start we will spend some time on the discrete-time additive*Gaussian* channel. In later chapters you will realize that this channel is a cornerstone of digital communications.

In Chapter 3 we will focus on the *waveform generator* and on the *baseband front-end* of Figure 1.2. The mathematical tool behind the description of the waveform generator is the notion of *orthonormal expansion* from linear algebra. We will fix an orthonormal basis and we will let the output of the encoder be the n-tuple of coefficients that determines the signal produced by the transmitter (with respect to the given orthonormal basis).

The baseband front-end of the receiver reduces the received waveform to an n-tuple that contains just as much information as needed to implement a receiver that minimizes the *error probability.* To do so, the baseband front-end *projects* the received waveform onto each element of the mentioned orthonormal basis. The resulting n-tuple is passed to the decoder. Together, the encoder and the waveform generator form the *transmitter.* Correspondingly, the baseband front-end and the decoder form the *receiver.* What we do in Chapter 3 holds irrespectively of the specific set of signals that we use to communicate.

Chapter 4 is meant to develop intuition about the high-level implications of the *signal set* used to communicate. It is in this chapter that we start shifting attention from the problem of designing the receiver for a given set of signals to the problem of designing the signal set itself.

In Chapter 5 we further explore the problem of making sensible choices concerning the signal set. We will learn to appreciate the advantages of the widespread method of communicating by modulating the amplitude of a pulse and its shifts delayed by integer multiples of the symbol time T. We will see that, when possible, one should choose the pulse to fulfill the so-called *Nyquist criterion.*

Chapter 6 is a case study on *coding.* The communication model is that of Chapter 2 with the n-tuple channel being Gaussian. The encoder will be of *convolutional* type and the decoder will be based on the *Viterbi algorithm.*

Chapter 7 is a technical one in which we learn dealing with *complex-valued Gaussian processes* and *vectors.* They will be used in Chapter 8.

Chapter 8 deals with the problem of communicating across *bandpass AWGN channels.* The idea is to learn how to *shift the spectrum* of the transmitted signal so that we can place its center frequency at any desired location in the frequency axis, without changing the baseband waveforms. This will be done using the *frequency-shift property* of the *Fourier transform.* Implementing signal processing (amplification, filtering, multiplication of signals, etc.) becomes more and more challenging as the center frequency of the signals being processed increases. This is so since simple wires meant to carry the signal inside the circuit may act as transmit antenna and irradiate the signal. This may cause all kinds of problems, including the fact that signals get mixed "in the air" and, even worse, are reabsorbed into the circuit by some short wire that acts as receive antenna causing interference, oscillations due to unwanted feedback, etc. To minimize such problems, it is common practice to design the core of the sender and of the receiver for a fixed center frequency and let the last stage of the sender and the first stage of the receiver do the frequency translation. The fixed center frequency typically ranges from zero to a few MHz. Operations done at the fixed center frequency will be referred to as being done in *baseband.* The ones at the final center frequency will be said to be in *passband.* As it turns out, the baseband representation of a general passband signal is *complex-valued.* This means that the transmitter/receiver pairs have to deal with complex-valued signals. This is not a problem per se. In fact working with complex-valued signals simplifies the notation. However, it requires a small overhead (Chapter 7) in terms of having to learn how to deal with complex-valued stochastic processes and complex-valued random vectors. In this chapter we will also “close the loop“” and understand the importance of the (discrete-time) AWGN channel considered in Chapter 2.

To emphasize the importance of the discrete-time AWGN channel, we mention that in a typical information theory course (mandatory at EPFL for master-level students) as well as in a typical coding theory course (offered at EPFL in the Ph.D. program), the channel model is always discrete-time and often AWGN. In those classes one takes it for granted that the student knows why discrete-time channel models are important.

Chapter 2

Receiver Design for Discrete-Time Observations

2.1 Introduction

As pointed out in the introduction, we will study point-to-point communications from various abstraction levels. In this chapter we will be dealing with the receiver design problem for discrete-time observations with particular emphasis on the discrete time additive white Gaussian (AWGN) channel. Later we will see that this channel is an important abstraction model. For now it suffices to say that it is the channel that we see from the input to the output of the dotted box in Figure 2.1. The goal of this chapter is to understand how to design and analyze the decoder when the channel and the encoder are given.

When the channel model is discrete time, the encoder is indeed the transmitter and the decoder is the receiver, see Figure 2.2. The figure depicts the system considered in this Chapter. Its components are:

- *A Source:* The source (not shown in the figure) is responsible for producing the message H which takes values in the message set $\mathcal{H} = \{0, 1, \ldots, (m-1)\}$. The task of the receiver would be extremely simple if the source selected the message according to some deterministic rule. In this case the receiver could reproduce the source message by following the same algorithm and there would be no need to communicate. For this reason, in communication we always assume that the source is modeled by a random variable, here denoted by the capital letter H. As usual, a random variable taking values on a finite alphabet is described by its probability mass function $P_H(i)$, $i \in \mathcal{H}$. In most cases of interest to us, H is uniformly distributed.

- *A Transmitter:* The transmitter is a mapping from the message set $\mathcal{H}$ to the signal set $\mathcal{S} = \{\boldsymbol{s_0}, \boldsymbol{s_1}, \ldots, \boldsymbol{s_{m-1}}\}$ where $\boldsymbol{s_i} \in \mathbb{C}^n$ for some n. We will start with $\boldsymbol{s}_i \in \mathbb{R}^n$ but we will see in Chapter 8 that allowing $\boldsymbol{s}_i \in \mathbb{C}^n$ is crucial.

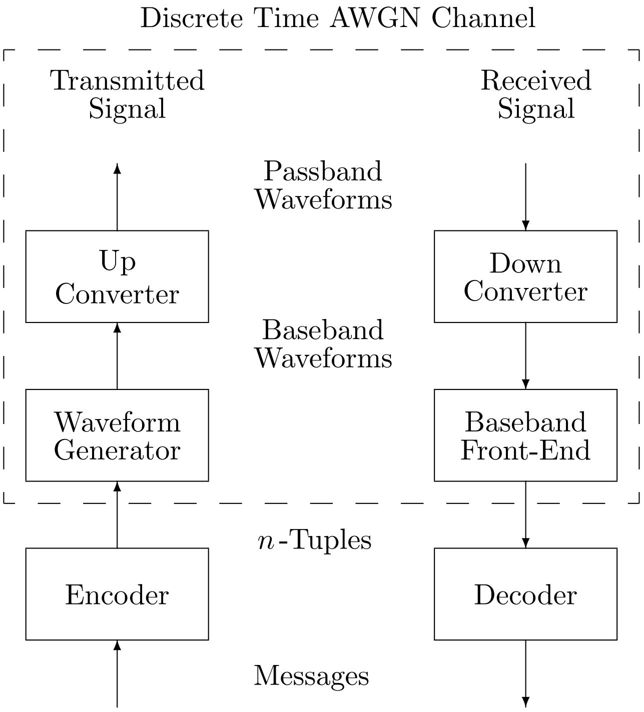


Figure 2.1: Discrete time AWGN channel abstraction.

- *A Channel:* The channel is described by the probability density of the output for each of the possible inputs. When the channel input is $\boldsymbol{s}_i$, the probability density of $\boldsymbol{Y}$ will be denoted by $f_{\boldsymbol{Y}|\boldsymbol{S}}(\cdot|\boldsymbol{s}_i)$.
- *A Receiver:* The receiver's task is to "guess" H from $\boldsymbol{Y}$. The decision made by the receiver is denoted by $\hat{H}$. Unless specified otherwise, the receiver will always be designed to minimize the probability of error defined as the probability that $\hat{H}$ differs from H. Guessing H from $\boldsymbol{Y}$ when H is a discrete random variable is the so-called *hypothesis testing* problem that comes up in various contexts (not only in communications).

First we give a few examples.

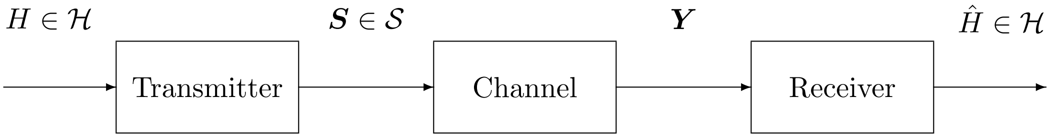


Figure 2.2: General setup for Chapter 2.

Example 1. *A common source model consist of* $\mathcal{H} = \{0, 1\}$ *and* $P_H(0) = P_H(1) = 1/2$. *This models individual bits of, say, a file. Alternatively, one could model an entire file of, say, 1 Mbit by saying that* $\mathcal{H} = \{0, 1, \ldots, (2^{10^6} - 1)\}$ *and* $P_H(i) = \frac{1}{2^{10^6}}, i \in \mathcal{H}$.

Example 2. *A transmitter for a binary source could be a map from* $\mathcal{H} = \{0, 1\}$ *to* $\mathcal{S} = \{-a, a\}$ *for some real-valued constant* a. *Alternatively, a transmitter for a 4-ary source could be a map from* $\mathcal{H} = \{0, 1, 2, 3\}$ *to* $\mathcal{S} = \{a, ia, -a, -ia\}$, *where* $i = \sqrt{-1}$.

Example 3. *The channel model that we will use mostly in this chapter is the one that maps a channel input* $\boldsymbol{s} \in \mathbb{R}^n$ *into* $\boldsymbol{Y} = \boldsymbol{s} + \boldsymbol{Z}$, *where* $\boldsymbol{Z}$ *is a Gaussian random vector of independent and uniformly distributed components. As we will see later, this is the discrete-time equivalent of the baseband continous-time channel called additive white Gaussian noise (AWGN) channel. For that reason, following common practice, we will refer to both as additive white Gaussian noise channels (AWGNs).*

The chapter is organized as follows. We first learn the basic ideas behind hypothesis testing, which is the field that deals with the problem of guessing the outcome of a random variable based on the observation of another random variable. Then we study the Q function since it is a very valuable tool in dealing with communication problems that involve Gaussian noise. At this point we are ready to consider the problem of communicating across the additive white Gaussian noise channel. We will fist consider the case that involves two messages and scalar signals, then the case of two messages and n-tuple signals, and finally the case of an arbitrary number m of messages and n-tuple signals. The last part of the chapter deals with techniques to bound the error probability when and exact expression is hard or impossible to get.

2.2 Hypothesis Testing

Detection, *decision*, and *hypothesis testing* are all synonyms. They refer to the problem of deciding the outcome of a random variable H that takes values in a finite alphabet $\mathcal{H} = \{0, 1, \ldots, m-1\}$, from the outcome of some related random variable Y. The latter is referred to as the *observable.*

The problem that a receiver has to solve is a detection problem in the above sense. Here the hypothesis H is the message selected by the source. To each message there is a signal that the transmitter plugs into the channel. The channel output is the observable Y. Its distribution depends on the input (otherwise observing Y would not help in guessing the message). The receiver guesses H from Y, assuming that the distribution of H as well as the conditional distribution of Y given H are known. The former is the source statistic and the latter depends on the sender and on the channel statistical behavior. The receiver's decision will be denoted by $\hat{H}$. We wish to make $\hat{H} = H$, but this is not always possible. The goal is to devise a decision strategy that maximizes the probability $P_c = Pr\{\hat{H} = H\}$ that the decision is correct.[1]

[1] $Pr\{\cdot\}$ is a short-hand for *probability of the enclosed event.*

We will always assume that we know the *a priori* probability P_H and that for each $i \in \mathcal{H}$ we know the conditional probability density function[2] (pdf) of Y given $H = i$, denoted by $f_{Y|H}(\cdot|i)$.

Example 4. *As a typical example of a hypothesis testing problem, consider the problem of communicating one bit of information across an optical fiber. The bit being transmitted is modeled by the random variable* $H \in \{0, 1\}$, $P_H(0) = 1/2$. *If* $H = 1$, *we switch on an LED and its light is carried across an optical fiber to a photodetector at the receiver front end. The photodetector outputs the number of photons* $Y \in \mathbb{N}$ *it detects. The problem is to decide whether* $H = 0$ *(the LED is off) or* $H = 1$ *(the LED is on). Our decision may only be based on whatever prior information we have about the model and on the actual observation* y. *What makes the problem interesting is that it is impossible to determine* H *from* Y *with certainty. Even if the LED is off, the detector is likely to detect some photons (e.g. due to "ambient light"). A good assumption is that* Y *is Poisson distributed with intensity* λ *that depends on whether the LED is on or off. Mathematically, the situation is as follows:*

$$\begin{aligned} H = 0, \quad & Y \sim p_{Y|H}(y|0) = \frac{\lambda_0^y}{y!} e^{-\lambda_0}. \\ H = 1, \quad & Y \sim p_{Y|H}(y|1) = \frac{\lambda_1^y}{y!} e^{-\lambda_1}. \end{aligned}$$

We read the first row as follows: "When the hypothesis is $H = 0$ *then the observable* Y *is Poisson distributed with intensity* λ_0*".*

Once again, the problem of deciding the value of H *from the observable* Y *when we know the distribution of* H *and that of* Y *for each value of* H *is a standard hypothesis testing problem.* □

From P_H and $f_{Y|H}$, via Bayes rule, we obtain

$$P_{H|Y}(i|y) = \frac{P_H(i) f_{Y|H}(y|i)}{f_Y(y)}$$

where $f_Y(y) = \sum_i P_H(i) f_{Y|H}(y|i)$. In the above expression $P_{H|Y}(i|y)$ is the *posterior* (also called *a posteriori probability* of H given Y). By observing $Y = y$, the probability that $H = i$ goes from $p_H(i)$ to $P_{H|Y}(i|y)$.

If we choose $\hat{H} = i$, then $P_{H|Y}(i|y)$ is the probability that we made the correct decision. Since our goal is to maximize the probability of being correct, the optimum decision rule is

$$\hat{H}(y) = \arg\max_i P_{H|Y}(i|y) \qquad \textit{(MAP decision rule).} \tag{2.1}$$

[2] In most cases of interest in communication, the random variable Y is a continuous one. That's why in the above discussion we have implicitly assumed that, given $H = i$, Y has a pdf $f_{Y|H}(\cdot|i)$. If Y is a discrete random variable, then we assume that we know the conditional probability mass function $p_{Y|H}(\cdot|i)$.

This is called *maximum a posteriori (MAP) decision rule*. In case of ties, i.e. if $P_{H|Y}(j|y)$ equals $P_{H|Y}(k|y)$ equals $\max_i P_{H|Y}(i|y)$, then it does not matter if we decide for $\hat{H} = k$ or for $\hat{H} = j$. In either case the probability that we have decided correctly is the same.

Since the MAP rule maximizes the probability of being correct for each observation y, it also maximizes the unconditional probability of being correct P_c. The former is $P_{H|Y}(\hat{H}(y)|y)$. If we plug in the random variable Y instead of y, then we obtain a random variable. (A real-valued function of a random variable is a random variable.) The expected valued of this random variable is the (unconditional) probability of being correct, i.e.,

$$P_c = E[P_{H|Y}(\hat{H}(Y)|Y)] = \int_y P_{H|Y}(\hat{H}(y)|y) f_Y(y) dy. \tag{2.2}$$

There is an important special case, namely when H is uniformly distributed. In this case $P_{H|Y}(i|y)$, as a function of i, is proportional to $f_{Y|H}(y|i)/m$. Therefore, the argument that maximizes $P_{H|Y}(i|y)$ also maximizes $f_{Y|H}(y|i)$. Then the MAP decision rule is equivalent to *the maximum likelihood (ML) decision rule*:

$$\hat{H}(y) = \arg\max_i f_{Y|H}(y|i) \qquad \textit{(ML decision rule).} \tag{2.3}$$

2.2.1 Binary Hypothesis Testing

The special case in which we have to make a binary decision, i.e., $H \in \mathcal{H} = \{0, 1\}$, is both instructive and of practical relevance. Since there are only two alternatives to be tested, the MAP test may now be written as

$$\frac{f_{Y|H}(y|1)P_H(1)}{f_Y(y)} \underset{\hat{H}=0}{\overset{\hat{H}=1}{\underset{<}{\geq}}} \frac{f_{Y|H}(y|0)P_H(0)}{f_Y(y)}.$$

Observe that the denominator is irrelevant since $f(y)$ is a positive constant — hence will not affect the decision. Thus an equivalent decision rule is

$$f_{Y|H}(y|1)P_H(1) \underset{\hat{H}=0}{\overset{\hat{H}=1}{\underset{<}{\geq}}} f_{Y|H}(y|0)P_H(0).$$

The above test is depicted in Fig. 2.3 assuming $y \in \mathbb{R}$. This is a very important figure that helps us visualize what goes on and, as we will see, will be helpful to compute the probability of error.

Yet an equivalent rule obtained by dividing both sides with the non-negative quantity

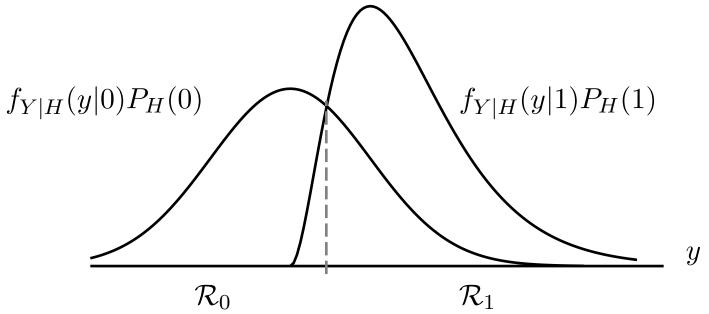


Figure 2.3: Binary MAP Decision. The decision regions $\mathcal{R}_0$ and $\mathcal{R}_1$ are the values of y (abscissa) on the left and right of the dashed line (threshold), respectively.

$f_{Y|H}(y|0)P_H(1)$. This results in the following *binary MAP test*:

$$\Lambda(y) = \frac{f_{Y|H}(y|1)}{f_{Y|H}(y|0)} \begin{array}{c} \hat{H}=1 \\ \geq \\ < \\ \hat{H}=0 \end{array} \frac{P_H(0)}{P_H(1)} = \eta. \tag{2.4}$$

The left side of the above test is called the *likelihood ratio*, denoted by $\Lambda(y)$, whereas the right side is the *threshold* η. Notice that if $P_H(0)$ increases, so does the threshold. In turn, as we would expect, the region $\{y : \hat{H}(y) = 0\}$ becomes bigger.

When $P_H(0) = P_H(1) = 1/2$ the threshold η becomes unity and the MAP test becomes a *binary ML test*:

$$f_{Y|H}(y|1) \begin{array}{c} \hat{H}=1 \\ \geq \\ < \\ \hat{H}=0 \end{array} f_{Y|H}(y|0).$$

A function $\hat{H} : \mathcal{Y} \to \mathcal{H}$ is called a *decision function* (also called *decoding function*). One way to describe a decision function is by means of the *decision regions* $\mathcal{R}_i = \{y \in \mathcal{Y} : \hat{H}(y) = i\}$, $i \in \mathcal{H}$. Hence $\mathcal{R}_i$ is the set of $y \in \mathcal{Y}$ for which $\hat{H}(y) = i$.

To compute the probability of error it is often convenient to compute the error probability for each hypothesis and then take the average. When $H = 0$, we make an incorrect decision if $Y \in \mathcal{R}_1$ or, equivalently, if $\Lambda(y) \geq \eta$. Hence, denoting by $P_e(i)$ the probability of making an error when $H = i$,

$$\begin{aligned} P_e(0) = Pr\{Y \in \mathcal{R}_1 | H = 0\} &= \int_{\mathcal{R}_1} f_{Y|H}(y|0)dy & (2.5) \\ &= Pr\{\Lambda(Y) \geq \eta | H = 0\}. & (2.6) \end{aligned}$$

Whether it is easier to work with the right side of (2.5) or that of (2.6) depends on whether it is easier to work with the conditional density of Y or of $\Lambda(Y)$. We will see examples of both cases.

Similar expressions hold for the probability of error conditioned on $H = 1$, denoted by $P_e(1)$. The unconditional error probability is then

$$P_e = P_e(1)p_H(1) + P_e(0)p_H(0).$$

From (2.4) we see that, for the purpose of performing a MAP test, having $\Lambda(Y)$ is as good as having the observable Y and this is true regardless of the prior. A function of Y that has this property is called *sufficient statistic.* The concept of sufficient statistic is developed in Section 2.5

In deriving the probability of error we have tacitly used an important technique that we use all the time in probability: conditioning as an intermediate step. Conditioning as an intermediate step may be seen as a divide-and-conquer strategy. The idea is to solve a problem that seems hard, by braking it up into subproblems that (i) we know how to solve and (ii) once we have the solution to the subproblems we also have the solution to the original problem. Here is how it works in probability. We want to compute the expected value of a random variable Z. Assume that it is not immediately clear how to compute the expected value of Z but we know that Z is related to another random variable W that tells us something useful about Z: useful in the sense that for any particular value $W = w$ we now how to compute the expected value of Z. The latter is of course $E\left[Z|W = w\right]$. If this is the case, via the theorem of total expectation we have the solution to the problem we were looking for: $E\left[Z\right] = \sum_w E\left[Z|W = w\right] P_W(w)$. The same idea applies to compute probabilities. Indeed if the random variable Z is the indicator function of an event, then the expected value of Z is the probability of that event. The indicator function of an event is 1 when the event occurs and 0 otherwise. Specifically, if Z=1 when the event $\{H \neq \hat{H}\}$ occurs and $Z = 0$ otherwise then $E\left[Z\right]$ is the probability of error.

Let us revisit what we have done in light of the above comments and see what else we could have done. The computation of the probability of error involves two random variables, H and$Y$, as well as an event $\{H \neq \hat{H}\}$. To compute the probability of error (2.5) we have first conditioned on all possible values of H. Alternatively, we could have conditioned on all possible values of Y. This is indeed a viable alternative. In fact we have already done so (without saying it) in (2.2). Between the two we use the one that seems more promising for the problem at hand. We will see examples of both.

2.2.2 m-ary Hypothesis Testing

Now we go back to the m-ary hypothesis testing problem. This means that $\mathcal{H} = \{0, 1, \cdots, m-1\}$.

Recall that the MAP decision rule, which minimizes the probability of making an error, is

$$\begin{aligned}\hat{H}_{MAP}(\boldsymbol{y}) &= \arg\max_i P_{H|\boldsymbol{Y}}(i|\boldsymbol{y}) \\ &= \arg\max_i \frac{f_{\boldsymbol{Y}|H}(\boldsymbol{y}|i)P_H(i)}{f_{\boldsymbol{Y}}(\boldsymbol{y})} \\ &= \arg\max_i f_{\boldsymbol{Y}|H}(\boldsymbol{y}|i)P_H(i),\end{aligned}$$

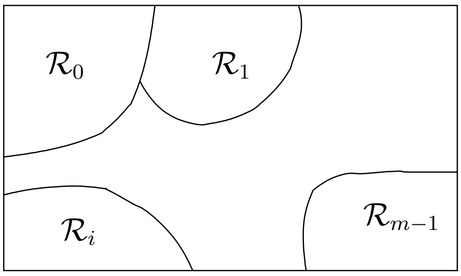
where $f_{\boldsymbol{Y}|H}(\cdot|i)$ is the probability density function of the observable $\boldsymbol{Y}$ when the hypothesis is i and $P_H(i)$ is the probability of the ith hypothesis. This rule is well defined up to ties. If there is more than one i that achieves the maximum on the right side of one (and thus all) of the above expressions, then we may decide for any such i without affecting the probability of error. If we want the decision rule to be unambiguous, we can for instance agree that in case of ties we pick the largest i that achieves the maximum.

When all hypotheses have the same probability, then the MAP rule specializes to the ML rule, i.e.,

$$\hat{H}_{ML}(\boldsymbol{y}) = \arg\max_i f_{\boldsymbol{Y}|H}(\boldsymbol{y}|i).$$

We will always assume that $f_{\boldsymbol{Y}|H}$ is either given as part of the problem formulation or that it can be figured out from the setup. In communications, one typically is given the transmitter, i.e. the map from $\mathcal{H}$ to $\mathcal{S}$, and the channel, i.e. the pdf $f_{\boldsymbol{Y}|\boldsymbol{X}}(\cdot|\boldsymbol{x})$ for all $\boldsymbol{x} \in \mathcal{X}$. From these two one immediately obtains $f_{\boldsymbol{Y}|H}(\boldsymbol{y}|i) = f_{\boldsymbol{Y}|\boldsymbol{X}}(\boldsymbol{y}|\boldsymbol{s}_i)$, where $\boldsymbol{s}_i$ is the signal assigned to i.

Note that the decoding (or decision) function $\hat{H}$ assigns an $i \in \mathcal{H}$ to each $\boldsymbol{y} \in \mathbb{R}^n$. As already mentioned, it can be described by the decoding (or decision) regions $\mathcal{R}_i$, $i \in \mathcal{H}$, where $\mathcal{R}_i$ consists of those $\boldsymbol{y}$ for which $\hat{H}(\boldsymbol{y}) = i$. It is convenient to think of $\mathbb{R}^n$ as being partitioned by decoding regions as depicted in the following figure.



We use the decoding regions to express the error probability P_e or, equivalently, the probability P_c of deciding correctly. Conditioned on $H = i$ we have

$$\begin{aligned}P_e(i) &= 1 - P_c(i) \\ &= 1 - \int_{\mathcal{R}_i} f_{\boldsymbol{Y}|H}(\boldsymbol{y}|i)d\boldsymbol{y}.\end{aligned}$$

2.3 The Q Function

The Q function plays a very important role in communications. It will come up over and over again throughout these notes. Make sure that you understand it well. It is defined as:

$$Q(x) \overset{\triangle}{=} \frac{1}{\sqrt{2\pi}} \int_x^\infty e^{-\frac{\xi^2}{2}} d\xi.$$

Hence if $Z \sim \mathcal{N}(0,1)$ (meaning that Z is a Normally distributed zero-mean random variable of unit variance) then $Pr\{Z \geq x\} = Q(x)$.

If $Z \sim \mathcal{N}(m,\sigma^2)$ the probability $Pr\{Z \geq x\}$ can also be written using the Q function. In fact the event $\{Z \geq x\}$ is equivalent to $\{\frac{Z-m}{\sigma} \geq \frac{x-m}{\sigma}\}$. But $\frac{Z-m}{\sigma} \sim \mathcal{N}(0,1)$. Hence $Pr\{Z \geq x\} = Q(\frac{x-m}{\sigma})$. Make sure you are familiar with these steps. We will use them frequently.

We now describe some of the key properties of the Q function.

(a) If $Z \sim \mathcal{N}(0,1)$, $F_Z(z) \overset{\triangle}{=} Pr\{Z \leq z\} = 1 - Q(z)$. (Draw a picture that expresses this relationship in terms of areas under the probability density function of Z.)

(b) $Q(0) = 1/2$, $Q(-\infty) = 1$, $Q(\infty) = 0$.

(c) $Q(-x) + Q(x) = 1$. (Again, draw a picture.)

(d) $\frac{1}{\sqrt{2\pi}\alpha} e^{-\frac{\alpha^2}{2}} (1 - \frac{1}{\alpha^2}) < Q(\alpha) < \frac{1}{\sqrt{2\pi}\alpha} e^{-\frac{\alpha^2}{2}}$, $\alpha > 0$.

(e) An alternative expression for the Q function with fixed integration limits is $Q(x) = \frac{1}{\pi}\int_0^{\frac{\pi}{2}} e^{-\frac{x^2}{2\sin^2\theta}} d\theta$. It holds for $x \geq 0$.

(f) $Q(\alpha) \leq \frac{1}{2} e^{-\frac{\alpha^2}{2}}$, $\alpha \geq 0$.

Proofs: The proofs or (a), (b), and (c) are immediate (a picture suffices). The proof of part (d) is left as an exercise (see Problem 34). To prove (e), let $X \sim \mathcal{N}(0,1)$ and $Y \sim \mathcal{N}(0,1)$ be independent. Hence $Pr\{X \geq 0, Y \geq \xi\} = Q(0)Q(\xi) = \frac{Q(\xi)}{2}$. Using Polar coordinates

$$\frac{Q(\xi)}{2} = \int_0^{\frac{\pi}{2}} \int_{\frac{\xi}{\sin\theta}}^\infty \frac{e^{-\frac{r^2}{2}}}{2\pi} r dr d\theta = \frac{1}{2\pi} \int_0^{\frac{\pi}{2}} \int_{\frac{\xi^2}{2\sin^2\theta}}^\infty e^{-t} dt d\theta = \frac{1}{2\pi} \int_0^{\frac{\pi}{2}} e^{-\frac{\xi^2}{2\sin^2\theta}} d\theta.$$

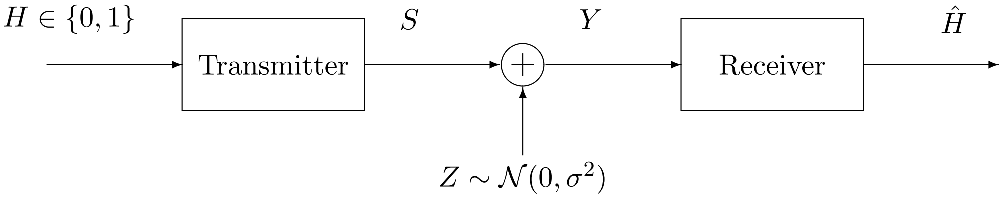
To prove (f) we use (e) and the fact that $e^{-\frac{\xi^2}{2\sin^2\theta}} \leq e^{-\frac{\xi^2}{2}}$ for $\theta \in [0, \frac{\pi}{2}]$. Hence

$$Q(\xi) \leq \frac{1}{\pi} \int_0^{\frac{\pi}{2}} e^{-\frac{\xi^2}{2}} d\theta = \frac{1}{2} e^{-\frac{\xi^2}{2}}.$$

2.4 Receiver Design for Discrete-Time AWGN Channels

2.4.1 Binary Decision for Scalar Observations

We consider the following setup



We assume that the transmitter maps $H=0$ into $a \in \mathbb{R}$ and $H=1$ into $b \in \mathbb{R}$. The output statistic for the various hypotheses is as follows:

$$\begin{aligned} H=0: &\quad Y \sim \mathcal{N}(a,\sigma^2) \\ H=1: &\quad Y \sim \mathcal{N}(b,\sigma^2). \end{aligned}$$

An equivalent way to express the output statistic for each hypothesis is

$$\begin{aligned} f_{Y|H}(y|0) &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(y-a)^2}{2\sigma^2}\right\} \\ f_{Y|H}(y|1) &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(y-b)^2}{2\sigma^2}\right\}. \end{aligned}$$

We compute the likelihood ratio

$$\Lambda(y) = \frac{f_{Y|H}(y|1)}{f_{Y|H}(y|0)} = \exp\left\{-\frac{(y-b)^2-(y-a)^2}{2\sigma^2}\right\} = \exp\left\{\frac{b-a}{\sigma^2}(y-\frac{a+b}{2})\right\}. \tag{2.7}$$

The threshold is $\eta = \frac{P_0}{P_1}$. Now we have all the ingredients for the MAP rule. Instead of comparing $\Lambda(y)$ to the threshold η we may compare $\log\Lambda(y)$ to $\log\eta$. The function $\log\Lambda(y)$ is called *log likelihood ratio*. Hence the MAP decision rule may be expressed as

$$\frac{b-a}{\sigma^2}\left(y-\frac{a+b}{2}\right) \underset{\hat{H}=0}{\overset{\hat{H}=1}{\underset{<}{\geq}}} \ln\eta.$$

Without loss of essential generality (w.l.o.g.), assume $b > a$. Then we can divide both sides by $\frac{b-a}{\sigma^2}$ without changing the outcome of the above comparison. In this case we obtain

$$\hat{H}_{\text{MAP}}(y) = \begin{cases} 1, & y > \theta \\ 0, & \text{otherwise,} \end{cases}$$

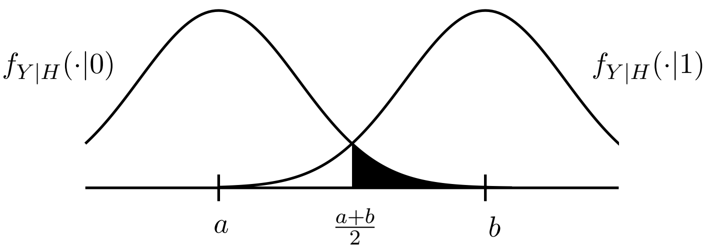


Figure 2.4: The shaded area represents the probability of error $P_e = Q(\frac{d}{2\sigma})$ when $H = 0$ and $P_H(0) = P_H(1)$.

where $\theta = \frac{\sigma^2}{b-a} \ln \eta + \frac{a+b}{2}$. Notice that if $P_H(0) = P_H(1)$, then $\ln \eta = 0$ and the threshold θ becomes the midpoint $\frac{a+b}{2}$.

We now determine the probability of error. Recall that

$$P_e(0) = Pr\{Y > \theta | H = 0\} = \int_{\mathcal{R}_1} f_{Y|H}(y|0)dy.$$

This is the probability that a Gaussian random variable with mean a and variance σ^2 exceeds the threshold θ. The situation is depicted in Figure 2.4. From our review on the Q function we know immediately that $P_e(0) = Q\left(\frac{\theta-a}{\sigma}\right)$. Similarly, $P_e(1) = Q\left(\frac{b-\theta}{\sigma}\right)$. Finally, $P_e = P_H(0)Q\left(\frac{\theta-a}{\sigma}\right) + P_H(1)Q\left(\frac{b-\theta}{\sigma}\right)$.

The most common case is when $P_H(0) = P_H(1) = 1/2$. Then $\frac{\theta-a}{\sigma} = \frac{b-\theta}{\sigma} = \frac{b-a}{2\sigma} = \frac{d}{2\sigma}$, where d is the distance between a and b. In this case

$$P_e = Q\left(\frac{d}{2\sigma}\right).$$

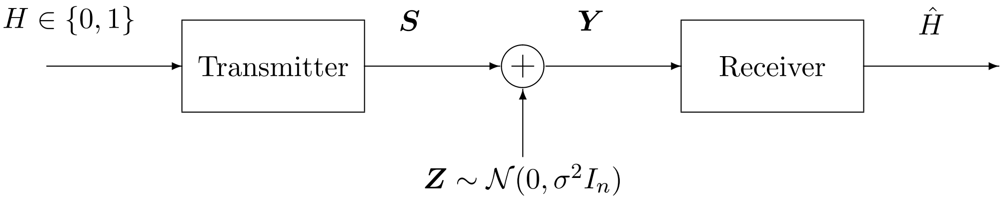
Computing P_e for the case $P_H(0) = P_H(1) = \frac{1}{2}$ is particularly straightforward. Due to symmetry, the threshold is the middle point between a and b and $P_e = P_e(0) = Q(\frac{d}{2\sigma})$, where d is the distance between a and b. (See again Figure 2.4.)

2.4.2 Binary Decision for n-Tuple Observations

The setup is the same as for the scalar case except that the transmitter output $\boldsymbol{s}$, the noise $\boldsymbol{z}$, and the observation $\boldsymbol{y}$ are now n-tuples. The new setting is represented in the figure below. Before going on we recommend reviewing the background material in Appendices 2.C and 2.E

We now assume that the hypothesis $i \in \{0, 1\}$ is mapped into the transmitter output $\boldsymbol{S}(i)$ defined by

$$\boldsymbol{S}(i) = \begin{cases} \boldsymbol{a} \in \mathbb{R}^n, & i = 0 \\ \boldsymbol{b} \in \mathbb{R}^n, & i = 1. \end{cases}$$



We also assume that $\boldsymbol{Z} \sim \mathcal{N}(0, \sigma^2 I_n)$. As we did earlier, we start by writing down the output statistic for each hypothesis

$$\begin{aligned} H = 0: &\quad \boldsymbol{Y} = \boldsymbol{a} + \boldsymbol{Z} \sim \mathcal{N}(\boldsymbol{a}, \sigma^2 I_n) \\ H = 1: &\quad \boldsymbol{Y} = \boldsymbol{b} + \boldsymbol{Z} \sim \mathcal{N}(\boldsymbol{b}, \sigma^2 I_n), \end{aligned}$$

or, equivalently,

$$\begin{aligned} f_{\boldsymbol{Y}|H}(\boldsymbol{y}|0) &= \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left\{ -\frac{\|\boldsymbol{y}-\boldsymbol{a}\|^2}{2\sigma^2} \right\} \\ f_{\boldsymbol{Y}|H}(\boldsymbol{y}|1) &= \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left\{ -\frac{\|\boldsymbol{y}-\boldsymbol{b}\|^2}{2\sigma^2} \right\}. \end{aligned}$$

Like in the scalar case we compute the likelihood ratio

$$\Lambda(\boldsymbol{y}) = \frac{f_{\boldsymbol{Y}|H}(\boldsymbol{y}|1)}{f_{\boldsymbol{Y}|H}(\boldsymbol{y}|0)} = \exp\left\{ \frac{\|\boldsymbol{y}-\boldsymbol{a}\|^2 - \|\boldsymbol{y}-\boldsymbol{b}\|^2}{2\sigma^2} \right\}.$$

Taking the logarithm on both sides and using the relationship $\langle \boldsymbol{u}+\boldsymbol{v}, \boldsymbol{u}-\boldsymbol{v} \rangle = \|\boldsymbol{u}\|^2 - \|\boldsymbol{v}\|^2$, which holds for real-valued vectors $\boldsymbol{u}$ and $\boldsymbol{v}$, we obtain

$$\begin{aligned} LLR(\boldsymbol{y}) &= \frac{\|\boldsymbol{y}-\boldsymbol{a}\|^2 - \|\boldsymbol{y}-\boldsymbol{b}\|^2}{2\sigma^2} && (2.8) \\ &= \left\langle \boldsymbol{y} - \frac{\boldsymbol{a}+\boldsymbol{b}}{2}, \frac{\boldsymbol{b}-\boldsymbol{a}}{\sigma^2} \right\rangle && (2.9) \\ &= \left\langle \boldsymbol{y}, \frac{\boldsymbol{b}-\boldsymbol{a}}{\sigma^2} \right\rangle + \frac{\|\boldsymbol{a}\|^2 - \|\boldsymbol{b}\|^2}{2\sigma^2}. && (2.10) \end{aligned}$$

From (2.10), the MAP rule is

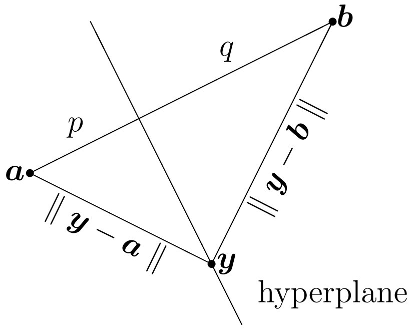
$$\langle \boldsymbol{y}, \boldsymbol{b}-\boldsymbol{a} \rangle \begin{array}{c} \hat{H}=1 \\ \geq \\ < \\ \hat{H}=0 \end{array} T,$$

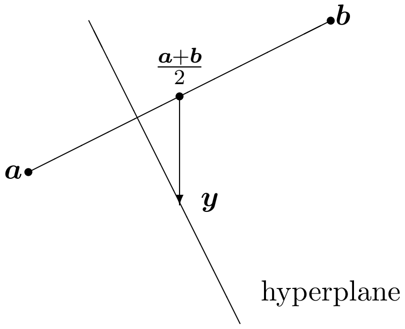
where $T = \sigma^2 \ln \eta + \frac{\|\boldsymbol{b}\|^2 - \|\boldsymbol{a}\|^2}{2}$ is a threshold and $\eta = \frac{P_H(0)}{P_H(1)}$. This says that the decision regions $\mathcal{R}_0$ and $\mathcal{R}_1$ are separated by the hyperplane[3]

$$\{\boldsymbol{y} \in \mathbb{R}^n : \langle \boldsymbol{y}, \boldsymbol{b}-\boldsymbol{a} \rangle = T\}.$$

[3] See Appendix 2.E for a review on the concept of hyperplane.

We obtain additional insight by analyzing (2.8) and (2.9). To find the boundary between $\mathcal{R}_0$ and $\mathcal{R}_1$, we look for the values of $\boldsymbol{y}$ for which (2.8) and (2.9) are constant. As shown by the left figure below, the set of points $\boldsymbol{y}$ for which (2.8) is constant is a hyperplane. Indeed, by Pythagoras, $\|\boldsymbol{y}-\boldsymbol{a}\|^2 - \|\boldsymbol{y}-\boldsymbol{b}\|^2$ equals $p^2 - q^2$. The right figure indicates that rule (2.9) performs the projection of $\boldsymbol{y} - \frac{\boldsymbol{a}+\boldsymbol{b}}{2}$ onto the linear space spanned by $\boldsymbol{b}-\boldsymbol{a}$. The set of points for which this projection is constant is again a hyperplane.


$\boldsymbol{b}$
q
p
$\boldsymbol{a}$
$\|\boldsymbol{y}-\boldsymbol{b}\|$
$\|\boldsymbol{y}-\boldsymbol{a}\|$
$\boldsymbol{y}$
hyperplane



The value of p (distance from $\boldsymbol{a}$ to the separating hyperplane) may be found by setting $\langle \boldsymbol{y}, \boldsymbol{b}-\boldsymbol{a}\rangle = T$ for $\boldsymbol{y} = \frac{\boldsymbol{b}-\boldsymbol{a}}{\|\boldsymbol{b}-\boldsymbol{a}\|}p$. This is the $\boldsymbol{y}$ where the line between $\boldsymbol{a}$ and $\boldsymbol{b}$ intersects the separating hyperplane. Inserting and solving for p we obtain

$$\begin{aligned} p &= \frac{d}{2} + \frac{\sigma^2 \ln \eta}{d} \\ q &= \frac{d}{2} - \frac{\sigma^2 \ln \eta}{d} \end{aligned}$$

with $d = \|\boldsymbol{b}-\boldsymbol{a}\|$ and $q = d - p$.

Of particular interest is the case $P_H(0) = P_H(1) = \frac{1}{2}$. In this case the hyperplane is the set of points for which (2.8) is 0. These are the points $\boldsymbol{y}$ that are at the same distance from $\boldsymbol{a}$ and from $\boldsymbol{b}$. Hence the ML decision rule for the AWGN channel decides for the transmitted vector that is closer to the observed vector.

A few additional observations are in order.

- The separating hyperplane moves towards $\boldsymbol{b}$ when the threshold T increases, which is the case when $\frac{P_H(0)}{P_H(1)}$ increases. This makes sense. It corresponds to our intuition that the decoding region $\mathcal{R}_0$ should become larger if the prior probability becomes more in favor of $H=0$.

- If $\frac{P_H(0)}{P_H(1)}$ exceeds 1, then $\ln \eta$ is positive and T increases with σ^2. This also makes sense. If the noise increases, we trust less what we observe and give more weight to the prior, which in this case favors $H=0$.

- Notice the similarity of (2.8) and (2.9) with the corresponding expressions for the scalar case, i.e., the expressions in the exponent of (2.7).

- The above comment suggest a tight relationship between the scalar and the vector case. One can gain additional insight by placing the origin of a new coordinate system at $\frac{\boldsymbol{a}+\boldsymbol{b}}{2}$ and by choosing the first coordinate in the direction of $\boldsymbol{b}-\boldsymbol{a}$. In this new coordinate system, $H=0$ is mapped into the vector $\tilde{\boldsymbol{a}}=(-\frac{d}{2},0,\dots,0)$ where $d=\|\boldsymbol{b}-\boldsymbol{a}\|$, $H=1$ is mapped into $\tilde{\boldsymbol{b}}=(\frac{d}{2},0,\dots,0)$, and the projection of the observation onto the subspace spanned by $\boldsymbol{b}-\boldsymbol{a}=(d,0,\dots,0)$ is just the first component y_1 of $\boldsymbol{y}=(y_1,y_2,\dots,y_n)$. This shows that for two hypotheses the vector case is really a scalar case embedded in an n dimensional space.

As for the scaler case, we compute the probability of error by conditioning on $H=0$ and $H=1$ and then remove the conditioning by averaging: $P_e=P_e(0)P_H(0)+P_e(1)P_H(1)$. When $H=0$, $\boldsymbol{Y}=\boldsymbol{a}+\boldsymbol{Z}$ and the MAP decoder makes the wrong decision if

$$\langle \boldsymbol{Y},\boldsymbol{b}-\boldsymbol{a}\rangle \geq T.$$

Inserting $\boldsymbol{Y}=\boldsymbol{a}+\boldsymbol{Z}$, defining the unit norm vector $\boldsymbol{\psi}_{\|}=\frac{\boldsymbol{b}-\boldsymbol{a}}{\|\boldsymbol{b}-\boldsymbol{a}\|}$ that points in the direction $\boldsymbol{b}-\boldsymbol{a}$ and rearranging terms yields the equivalent condition

$$\langle \boldsymbol{Z},\boldsymbol{\psi}_{\|}\rangle \geq \frac{d}{2}+\frac{\sigma^2\ln\eta}{d},$$

where again $d=\|\boldsymbol{b}-\boldsymbol{a}\|$. The left hand side is a zero-mean Gaussian random variable of variance σ^2 (see Appendix 2.C). Hence

$$P_e(0)=Q\Big(\frac{d}{2\sigma}+\frac{\sigma\ln\eta}{d}\Big).$$

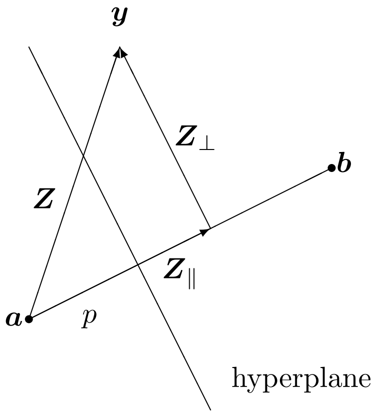
Proceeding similarly we find

$$P_e(1)=Q\Big(\frac{d}{2\sigma}-\frac{\sigma\ln\eta}{d}\Big).$$

In particular, when $P_H(0)=1/2$ we obtain

$$P_e=P_e(0)=P_e(1)=Q\Big(\frac{d}{2\sigma}\Big).$$

The figure below helps visualizing the situation. When $H=0$, a MAP decoder makes the wrong decision if the projection of $\boldsymbol{Z}$ onto the subspace spanned by $\boldsymbol{b}-\boldsymbol{a}$ lands on the other side of the separating hyperplane. The projection has the form $Z_{\|}\boldsymbol{\psi}_{\|}$ where $Z_{\|}=\langle\boldsymbol{Z},\boldsymbol{\psi}_{\|}\rangle\sim\mathcal{N}(0,\sigma^2)$. The projection lands on the other side of the separating hyperplane if $Z_{\|}\geq p$. This happens with probability $Q(\frac{p}{\sigma})$, which corresponds to the result obtained earlier.



2.4.3 m-ary Decision for n-Tuple Observations

When $H = i$, $i \in \mathcal{H}$, let $\boldsymbol{S} = \boldsymbol{s}_i \in \mathbb{R}^n$. Assume $P_H(i) = \frac{1}{m}$ (this is a common assumption in communications). The ML decision rule is

$$\begin{aligned}\hat{H}_{ML}(\boldsymbol{y}) &= \arg\max_i f_{\boldsymbol{Y}|H}(\boldsymbol{y}|i)\\ &= \arg\max_i \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\Big\{-\frac{\|\boldsymbol{y}-\boldsymbol{s}_i\|^2}{2\sigma^2}\Big\}\\ &= \arg\min_i \|\boldsymbol{y}-\boldsymbol{s}_i\|^2.\end{aligned}$$

Hence *a ML decision rule for the AWGN channel is a minimum-distance decision rule* as shown in Figure 2.5. Up to ties, $\mathcal{R}_i$ corresponds to the *Voronoi region* of $\boldsymbol{s}_i$, defined as the set of points in $\mathbb{R}^n$ that are at least as close to $\boldsymbol{s}_i$ as to any other $\boldsymbol{s}_j$.

EXAMPLE 5. *(PAM) Figure 2.6 shows the signal points and the decoding regions of a ML decoder for 6-ary Pulse Amplitude Modulation (why the name makes sense will become clear in the next chapter), assuming that the channel is AWGN. The signal points are elements of $\mathbb{R}$ and the ML decoder chooses according to the minimum-distance rule.*

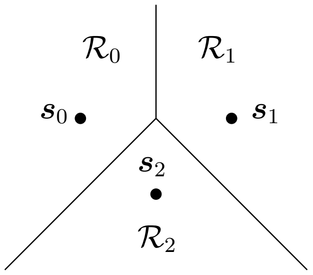


Figure 2.5: Example of Voronoi regions in $\mathbb{R}^2$.

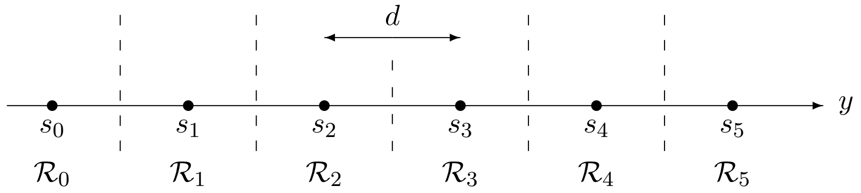


Figure 2.6: PAM signal constellation.

When the hypothesis is $H = 0$, *the receiver makes the wrong decision if the observation* $y \in \mathbb{R}$ *falls outside the decoding region* $\mathcal{R}_0$. *This is the case if the noise* $Z \in \mathbb{R}$ *is larger than* $d/2$, *where* $d = s_i - s_{i-1}$, $i = 1, \dots, 5$. *Thus*

$$P_e(0) = Pr\{Z > \frac{d}{2}\} = Q(\frac{d}{2\sigma}).$$

By symmetry, $P_e(5) = P_e(0)$. *For* $i \in \{1, 2, 3, 4\}$, *the probability of error when* $H = i$ *is the probability that the event* $\{Z \geq \frac{d}{2}\} \cup \{Z < -\frac{d}{2}\}$ *occurs. This event is the union of disjoint events. Its probability is the sum of the probability of the individual events. Hence*

$$P_e(i) = Pr\left\{\left\{Z \geq \frac{d}{2}\right\} \cup \left\{Z < -\frac{d}{2}\right\}\right\} = 2Pr\left\{Z \geq \frac{d}{2}\right\} = 2Q\left(\frac{d}{2\sigma}\right), \ i \in \{1, 2, 3, 4\}.$$

Finally,

$$P_e = \frac{2}{6}Q\left(\frac{d}{2\sigma}\right) + \frac{4}{6}2Q\left(\frac{d}{2\sigma}\right) = \frac{5}{3}Q\left(\frac{d}{2\sigma}\right).$$

□

Example 6. *(4-ary QAM) Figure 2.7 shows the signal set* $\{\boldsymbol{s}_0, \boldsymbol{s}_1, \boldsymbol{s}_2, \boldsymbol{s}_3\}$ *for 4-ary Quadrature Amplitude Modulation (QAM). We may consider signals as points in* $\mathbb{R}^2$ *or in* $\mathbb{C}$. *We choose the former since we don't know how to deal with complex valued noise yet. The noise is* $\boldsymbol{Z} \sim \mathcal{N}(0, \sigma^2 I_2)$ *and the observable, when* $H = i$, *is* $\boldsymbol{Y} = \boldsymbol{s}_i + \boldsymbol{Z}$. *We assume that the receiver implements a ML decision rule, which for the AWGN channel means minimum-distance decoding. The decoding region for* $\boldsymbol{s}_0$ *is the first quadrant, for* $\boldsymbol{s}_1$ *the second quadrant, etc. When* $H = 0$, *the decoder makes the correct decision if* $\{Z_1 > -\frac{d}{2}\} \cap \{Z_2 \geq -\frac{d}{2}\}$, *where* d *is the minimum distance among signal points. This is the intersection of independent events. Hence the probability of the intersection is the product of the probability of each event, i.e.*

$$P_c(0) = \left[Pr\left\{Z_i \geq -\frac{d}{2}\right\}\right]^2 = Q^2\left(-\frac{d}{2\sigma}\right) = \left[1 - Q\left(\frac{d}{2\sigma}\right)\right]^2.$$

By symmetry, for all i, $P_c(i) = P_c(0)$. *Hence,*

$$P_e = P_e(0) = 1 - P_c(0) = 2Q\left(\frac{d}{2\sigma}\right) - Q^2\left(\frac{d}{2\sigma}\right).$$

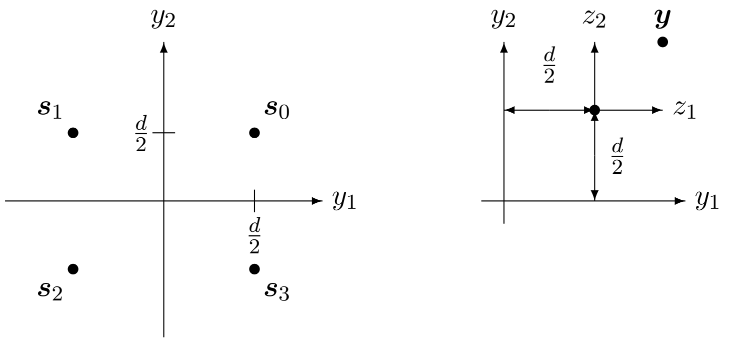


Figure 2.7: QAM signal constellation in $\mathbb{R}^2$.

When the channel is Gaussian and the decoding regions are bounded by affine planes, like in this and the previous example, one can express the error probability by means of the Q function. In this example we decided to focus on computing $P_c(0)$. It would have been possible to compute $P_e(0)$ instead of $P_c(0)$ but it would have costed slightly more work. To compute $P_e(0)$ we evaluate the probability of the union $\left\{Z_1 \leq -\frac{d}{2}\right\} \cup \left\{Z_2 \leq -\frac{d}{2}\right\}$. These are not *disjoint events. In fact they are independent events that can very well occur together. Thus the probability of the union is the sum of the individual probabilities minus the probability of the intersection. (You should verify that you obtain the same expression for P_e.)* □

Exercise 7. *Rotate and translate the signal constellation of Example 6 and evaluate the resulting error probability.*

2.5 Irrelevance and Sufficient Statistic

Have you ever tried to drink from a fire hydrant? There are situations in which the observable Y contains more data than you can handle. Some or most of that data may be irrelevant for the detection problem at hand but how to tell what is superfluous? In this section we give tests to do exactly that. We start by recalling the notion of Markov chain.

Definition 8. *Three random variables U, V, and W are said to form a Markov chain in that order, symbolized by $U \rightarrow V \rightarrow W$, if the distribution of W given both U and V is independent of U, i.e., $P_{W|V,U}(w|v,u) = P_{W|V}(w|v)$.*

The following exercise derives equivalent definitions.

Exercise 9. *Verify the following statements. (They are simple consequences of the definition of Markov chain.)*

(i) $U \to V \to W$ *if and only if* $P_{U,W|V}(u,w|v) = P_{U|V}(u|v)P_{W|V}(w|v)$*, i.e.,* U *and* W *are conditionally independent given* V*.*

(ii) $U \to V \to W$ *if and only if* $W \to V \to U$*, i.e., Markovity in one direction implies Markovity in the other direction.* □

Let Y be the observable and $T(Y)$ be a function (either stochastic or deterministic) of Y . Observe that $H \to Y \to T(Y)$ is always true but in general it is *not* true that $H \to T(Y) \to Y$.

Definition 10. *A function* $T(Y)$ *of an observable* Y *is said to be a* sufficient statistic *for* H *if* $H \to T(Y) \to Y$*.*

If $T(Y)$ is a sufficient statistic then the performance of a MAP decoder that observes $T(Y)$ is the same as that of one that observes Y. Indeed $P_{H|Y} = P_{H|Y,T} = P_{H|T}$. Hence, up to ties, $\arg\max P_{H|Y}(\cdot|y) = \arg\max P_{H|T}(\cdot|t)$. We state this important result as a theorem.

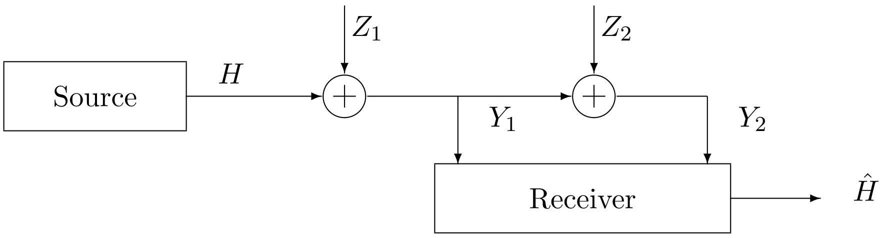
Theorem 11. *If* $T(Y)$ *is a sufficient statistic for* H *then a MAP decoder that estimates* H *from* $T(Y)$ *achieves the exact same error probability as one that estimates* H *from* Y*.*

Example 12. *Examples will be given in class.*

In some situations we make multiple measurements and want to prove that some of the measurements are relevant for the detection problem and some are not. Specifically, the observable Y may consist of two components $Y = (Y_1, Y_2)$ where Y_1 and Y_2 may be m and n tuples, respectively. If $T(Y) = Y_1$ is a sufficient statistic then we say that Y_2 is *irrelevant.* We use the two concepts interchangeably when we have two sets of observables: if one set is a sufficient statistic the other is irrelevant and vice-versa.

Exercise 13. *Assume the situation of the previous paragraph. Show that* Y_1 *is a sufficient statistic (or equivalently* Y_2 *is irrelevant) if and only if* $H \to Y_1 \to Y_2$*. (Hint: Show that* $H \to Y_1 \to Y_2$ *is equivalent to* $H \to Y_1 \to Y$*). This result is sometimes called* Theorem of Irrelevance *(See Wozencraft and Jacobs).*

Example 14. *Consider the communication system depicted in the figure where* Z_2 *is independent of* H *and* Z_1*. Then* $H \to Y_1 \to Y_2$*. Hence* Y_2 *is irrelevant for the purpose of making a MAP decision of* H *based on* (Y_1, Y_2)*.*



□

We have seen that $H \rightarrow T(Y) \rightarrow Y$ implies that Y is irrelevant to a MAP decoder that observes $T(Y)$. Is the contrary also true? Specifically, assume that a MAP decoder that observes $\big(Y, T(Y)\big)$ always makes the same decision as one that observes only $T(Y)$. Does this imply $H \rightarrow T(Y) \rightarrow Y$? The answer is "yes and no." We may expect the answer to be "no" since when $H \rightarrow U \rightarrow V$ holds then the function $P_{H|U,V}$ gives the same value as $P_{H|U}$ for all (i,u,v) whereas for v to have no effect on a MAP decision it is sufficient that for all (u,v) the *maximum* of $P_{H|U}$ and that of $P_{H|U,V}$ be achieved for the same i. In Problem 16 we give an example of this. Hence the answer to the above question is "no" in general. However, the example we give holds for a fixed distribution on H. In fact the answer to the above question becomes "yes" if Y does not affect the decision of a MAP decoder that observes $\big(Y, T(Y)\big)$ *regardless* of the distribution on H. We prove this in Problem 18 by showing that if $P_{H|U,V}(i|u,v)$ depends on v then for some distribution P_H the value of v affects the decision of a MAP decoder.

2.6 Error Probability

2.6.1 Union Bound

Here is a simple and extremely useful bound. Recall that for general events $\mathcal{A}, \mathcal{B}$

$$\begin{aligned} P(\mathcal{A} \cup \mathcal{B}) &= P(\mathcal{A}) + P(\mathcal{B}) - P(\mathcal{A} \cap \mathcal{B}) \\ &\leq P(\mathcal{A}) + P(\mathcal{B}). \end{aligned}$$

More generally, using induction, we obtain the the *Union Bound*

$$P\left(\bigcup_{i=1}^{M} \mathcal{A}_i\right) \leq \sum_{i=1}^{M} P(\mathcal{A}_i), \qquad (UB)$$

that applies to any collection of sets $\mathcal{A}_i$, $i = 1, \dots, M$. We now apply the union bound to approximate the probability of error in multi-hypothesis testing. Recall that

$$P_e(i) = Pr\{\boldsymbol{Y} \in \mathcal{R}_i^c | H = i\} = \int_{\mathcal{R}_i^c} f_{\boldsymbol{Y}|H}(\boldsymbol{y}|i) d\boldsymbol{y},$$

where $\mathcal{R}_i^c$ denotes the complement of $\mathcal{R}_i$. If we are able to evaluate the above integral for every i, then we are able to determine the probability of error exactly. The bound that we derive is useful if we are unable to evaluate the above integral.

For $i \neq j$ define

$$\mathcal{B}_{i,j} = \big\{\boldsymbol{y} : P_H(j) f_{\boldsymbol{Y}|H}(\boldsymbol{y}|j) \geq P_H(i) f_{\boldsymbol{Y}|H}(\boldsymbol{y}|i)\big\}.$$

$\mathcal{B}_{i,j}$ is the set of $\boldsymbol{y}$ for which the a posteriori probability of H given $\boldsymbol{Y} = \boldsymbol{y}$ is at least as high for $H = j$ as it is for $H = i$. Moreover,

$$\mathcal{R}_i^c \subseteq \bigcup_{j : j \neq i} \mathcal{B}_{i,j},$$

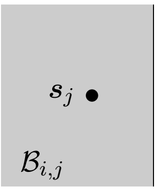


Figure 2.8: The shape of $\mathcal{B}_{i,j}$ for AWGN channels and ML decision.

with equality if ties are always resolved against i. In fact, by definition, the right side contains all the ties whereas the left side may or may not contain them. Here ties refers to those $\boldsymbol{y}$ for which equality holds in the definition of $\mathcal{B}_{i,j}$.

Now we use the union bound (with $\mathcal{A}_j = \{\boldsymbol{Y} \in \mathcal{B}_{i,j}\}$ and $P(\mathcal{A}_j) = Pr\{\boldsymbol{Y} \in \mathcal{B}_{i,j} | H = i\}$)

$$\begin{aligned} P_e(i) = Pr\left\{\boldsymbol{Y} \in \mathcal{R}_i^c | H = i\right\} &\leq Pr\Big\{\boldsymbol{Y} \in \bigcup_{j:j\neq i} \mathcal{B}_{i,j} | H = i\Big\} \\ &\leq \sum_{j:j\neq i} Pr\Big\{\boldsymbol{Y} \in \mathcal{B}_{i,j} | H = i\Big\} \qquad (2.11) \\ &= \sum_{j:j\neq i} \int_{\mathcal{B}_{i,j}} f_{\boldsymbol{Y}|H}(\boldsymbol{y}|i) d\boldsymbol{y}. \end{aligned}$$

What we have gained is that it is typically easier to integrate over $\mathcal{B}_{i,j}$ than over $\mathcal{R}_j^c$.

For instance, when the channel is the AWGN and the decision rule is ML, $\mathcal{B}_{i,j}$ is the set of points in $\mathbb{R}^n$ that are at least as close to $\boldsymbol{s}_j$ as they are to $\boldsymbol{s}_i$, as shown in the following figure. In this case,

$$\int_{\mathcal{B}_{i,j}} f_{\boldsymbol{Y}|H}(\boldsymbol{y}|i) d\boldsymbol{y} = Q\left(\frac{\|\boldsymbol{s}_j - \boldsymbol{s}_i\|}{2\sigma}\right),$$

and the union bound yields the simple expression

$$P_e(i) \leq \sum_{j:j\neq i} Q\left(\frac{\|\boldsymbol{s}_j - \boldsymbol{s}_i\|}{2\sigma}\right).$$

In the next section we derive an easy-to-compute tight upperbound on

$$\int_{\mathcal{B}_{i,j}} f_{\boldsymbol{Y}|H}(\boldsymbol{y}|i) d\boldsymbol{y}$$

for a general $f_{\boldsymbol{Y}|H}$. Notice that the above integral is the probability of error under $H = i$ when there are only two hypotheses, the other hypothesis is $H = j$, and the priors are proportional to $P_H(i)$ and $P_H(j)$.

Example 15. *(m-PSK) Figure 2.9 shows a signal set for m-ary PSK (phase-shift keying) when $m = 8$. Formally, the signal transmitted when $H = i$, $i \in \mathcal{H} = \{0, 1, \ldots, m-1\}$, is*

$$\boldsymbol{s}_i = \sqrt{\mathcal{E}_s}\Big(\cos\Big(\frac{2\pi i}{m}\Big), \sin\Big(\frac{2\pi i}{m}\Big)\Big)^T,$$

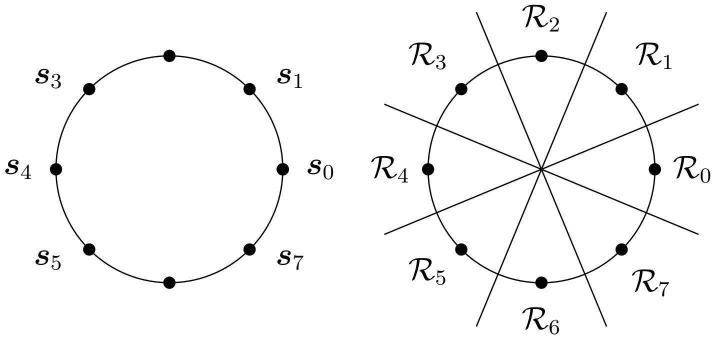


Figure 2.9: 8-ary PSK constellation in $\mathbb{R}^2$ and decoding regions.

where $\mathcal{E}_s = \|\boldsymbol{s}_i\|^2$, $i \in \mathcal{H}$. Assuming the AWGN channel, the hypothesis testing problem is specified by

$$H = i : \quad \boldsymbol{Y} \sim \mathcal{N}(\boldsymbol{s}_i, \sigma^2 I_2)$$

and the prior $P_H(i)$ is assumed to be uniform. Since we have a uniform prior, the MAP and the ML decision rule are identical. Furthermore, since the channel is the AWGN channel, the ML decoder is a minimum-distance decoder. The decoding regions (up to ties) are also shown in the figure.

One can show that

$$P_e(i) = \frac{1}{\pi} \int_0^{\pi - \frac{\pi}{m}} \exp\left\{ - \frac{\sin^2 \frac{\pi}{m}}{\sin^2(\theta + \frac{\pi}{m})} \frac{\mathcal{E}_s}{2\sigma^2} \right\} d\theta.$$

The above expression does not lead to a simple formula for the error probability.

Now we use the union bound to determine an upperbound to the error probability. With reference to Fig. 2.10 we have:

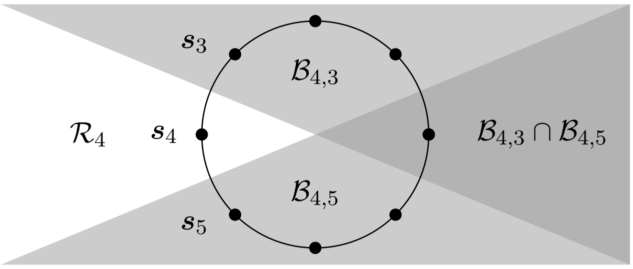


Figure 2.10: Bounding the error probability of PSK by means of the union bound.

$$\begin{aligned}P_e(i) &= Pr\{\boldsymbol{Y} \in \mathcal{B}_{i,i-1} \cup \mathcal{B}_{i,i+1} | H = i\}\\ &\leq Pr\{\boldsymbol{Y} \in \mathcal{B}_{i,i-1} | H = i\} + Pr\{\boldsymbol{Y} \in \mathcal{B}_{i,i+1} | H = i\}\\ &= 2Pr\{\boldsymbol{Y} \in \mathcal{B}_{i,i-1} | H = i\}\\ &= 2Q\left(\frac{\|\boldsymbol{s}_i - \boldsymbol{s}_{i-1}\|}{2\sigma}\right)\\ &= 2Q\left(\frac{\sqrt{\mathcal{E}_s}}{\sigma}\sin\frac{\pi}{m}\right).\end{aligned}$$

Notice that we have been using a version of the union bound adapted to the problem: we are getting a tighter bound by using the fact that $\mathcal{R}_i^c \subseteq \mathcal{B}_{i,i-1} \cup \mathcal{B}_{i,i+1}$ *(with equality with the possible exception of the boundary points) rather than* $\mathcal{R}_i^c \subseteq \cup_{j \neq i} \mathcal{B}_{i,j}$.

How good is the upper bound? Recall that

$$P_e(i) = Pr\{\boldsymbol{Y} \in \mathcal{B}_{i,i-1} | H = i\} + Pr\{\boldsymbol{Y} \in \mathcal{B}_{i,i+1} | H = i\} - Pr\{\boldsymbol{Y} \in \mathcal{B}_{i,i-1} \cap \mathcal{B}_{i,i+1} | H = i\}$$

and we obtained an upper bound by lower-bounding the last term with 0. *We now obtain a lower bound by upper-bounding the same term. To do so, observe that* $\mathcal{R}_i^c$ *is the union of* $(m-1)$ *disjoint cones, one of which is* $\mathcal{B}_{i,i-1} \cap \mathcal{B}_{i,i+1}$. *Furthermore, the integral of* $f_{\boldsymbol{Y}|H}(\cdot|i)$ *over* $\mathcal{B}_{i,i-1} \cap \mathcal{B}_{i,i+1}$ *is smaller than that over the other cones. Hence the integral over* $\mathcal{B}_{i,i-1} \cap \mathcal{B}_{i,i+1}$ *must be less than* $\frac{P_e(i)}{m-1}$. *Mathematically,*

$$Pr\{\boldsymbol{Y} \in (\mathcal{B}_{i,i-1} \cap \mathcal{B}_{i,i+1}) | H = i\} \leq \frac{P_e(i)}{m-1}.$$

Inserting in the previous expression, solving for $P_e(i)$ *and using the fact that* $P_e(i) = P_e$ *yields the desired lower bound*

$$P_e \geq 2Q\left(\sqrt{\frac{\mathcal{E}_s}{\sigma^2}}\sin\frac{\pi}{m}\right)\frac{m-1}{m}.$$

The ratio between the upper and the lower bound is the constant $\frac{m}{m-1}$. *For* m *large, the bounds become very tight. One can come up with lower bounds for which this ratio goes to* 1 *as* $\mathcal{E}_s/\sigma^2 \to \infty$. *One such bound is obtained by upper-bounding* $Pr\{\boldsymbol{Y} \in \mathcal{B}_{i,i-1} \cap \mathcal{B}_{i,i+1} | H = i\}$ *with the probability* $Q\left(\sqrt{\mathcal{E}_s}/\sigma\right)$ *that conditioned on* $H = i$, *the observable* $\boldsymbol{Y}$ *is on the other side of the hyperplane through the origine and perpendicular to* $\boldsymbol{s}_i$. □

2.6.2 Union Bhattacharyya Bound

Let us summarize. From the union bound applied to $\mathcal{R}_i^c \subseteq \bigcup_{j:j \neq i} \mathcal{B}_{i,j}$ we have obtained the upper bound

$$\begin{aligned}P_e(i) &= Pr\{\boldsymbol{Y} \in \mathcal{R}_i^c | H = i\}\\ &\leq \sum_{j:j \neq i} Pr\{\boldsymbol{Y} \in \mathcal{B}_{i,j} | H = i\}\end{aligned}$$

and we have used this bound for the AWGN channel. With the bound, instead of having to compute

$$Pr\{\boldsymbol{Y} \in \mathcal{R}_i^c | H = i\} = \int_{\mathcal{R}_i^c} f_{\boldsymbol{Y}|H}(\boldsymbol{y}|i) dy,$$

which requires integrating over a possibly complicated region $\mathcal{R}_i^c$, we only have to compute

$$Pr\{\boldsymbol{Y} \in \mathcal{B}_{i,j} | H = i\} = \int_{\mathcal{B}_{i,j}} f_{\boldsymbol{Y}|H}(\boldsymbol{y}|i) dy.$$

The latter integral is simply $Q(\frac{a}{\sigma})$, where a is the distance between $\boldsymbol{s}_i$ and the hyperplane bounding $\mathcal{B}_{i,j}$. For a ML decision rule, $a = \frac{\|\boldsymbol{s}_i - \boldsymbol{s}_j\|}{2}$.

What if the channel is *not* AWGN? Is there a relatively simple expression for $Pr\{\boldsymbol{Y} \in \mathcal{B}_{i,j} | H = i\}$ that applies for general channels? Such an expression does exist. It is the *Bhattacharyya bound* that we now derive.[4]

Given a set $\mathcal{A}$, the indicator function $1_{\mathcal{A}}$ is defined as

$$1_{\mathcal{A}}(x) = \begin{cases} 1, & x \in \mathcal{A} \\ 0, & \text{otherwise.} \end{cases}$$

From the definition of $\mathcal{B}_{i,j}$ that we repeat for convenience

$$\mathcal{B}_{i,j} = \{\boldsymbol{y} \in \mathbb{R}^n : P_H(i) f_{\boldsymbol{Y}|H}(\boldsymbol{y}|i) \le P_H(j) f_{\boldsymbol{Y}|H}(\boldsymbol{y}|j)\},$$

we immediately verify that $1_{\mathcal{B}_{i,j}}(\boldsymbol{y}) \le \sqrt{\frac{P_H(j) f_{\boldsymbol{Y}|H}(\boldsymbol{y}|j)}{P_H(i) f_{\boldsymbol{Y}|H}(\boldsymbol{y}|i)}}$. With this we obtain the Bhattacharyya bound as follows:

$$\begin{aligned} Pr\{\boldsymbol{Y} \in \mathcal{B}_{i,j} | H = i\} &= \int_{\boldsymbol{y} \in \mathcal{B}_{i,j}} f_{\boldsymbol{Y}|H}(\boldsymbol{y}|i) d\boldsymbol{y} = \int_{\boldsymbol{y} \in \mathbb{R}^n} f_{\boldsymbol{Y}|H}(\boldsymbol{y}|i) 1_{\mathcal{B}_{i,j}}(\boldsymbol{y}) d\boldsymbol{y} \\ &\le \sqrt{\frac{P_H(j)}{P_H(i)}} \int_{\boldsymbol{y} \in \mathbb{R}^n} \sqrt{f_{\boldsymbol{Y}|H}(\boldsymbol{y}|i) f_{\boldsymbol{Y}|H}(\boldsymbol{y}|j)}\, d\boldsymbol{y}. \end{aligned} \tag{2.12}$$

What makes the last integral appealing is that we integrate over the entire $\mathbb{R}^n$. As shown in Problem 29 (Bhattacharyya Bound for DMCs), for *discrete memoryless channels* the bound further simplifies.

As the name indicates, the *Union Bhattacharyya bound* combines (2.11) and (2.12), namely

$$P_e(i) \le \sum_{j:j \ne i} Pr\{\boldsymbol{Y} \in \mathcal{B}_{i,j} | H = i\} \le \sum_{j:j \ne i} \sqrt{\frac{P_H(j)}{P_H(i)}} \int_{\boldsymbol{y} \in \mathbb{R}^n} \sqrt{f_{\boldsymbol{Y}|H}(\boldsymbol{y}|i) f_{\boldsymbol{Y}|H}(\boldsymbol{y}|j)}\, d\boldsymbol{y}.$$

[4] There are two versions of the Bhattacharyya bound. Here we derive the one that has the simpler derivation. The other version, which is tighter by a factor 2, is derived in Problems 25 and 26.

We can now remove the conditioning on $H = i$ and obtain

$$P_e \le \sum_i \sum_{j:j\neq i} \sqrt{P_H(i)P_H(j)} \int_{\boldsymbol{y}\in\mathbb{R}^n} \sqrt{f_{\boldsymbol{Y}|H}(\boldsymbol{y}|i) f_{\boldsymbol{Y}|H}(\boldsymbol{y}|j)}\, d\boldsymbol{y}.$$

Example 16. (Tightness of the Bhattacharyya Bound) *Consider the following scenario*

$$\begin{aligned} H = 0: &\qquad \boldsymbol{S} = \boldsymbol{s}_0 = (0,0,\ldots,0)^T \\ H = 1: &\qquad \boldsymbol{S} = \boldsymbol{s}_1 = (1,1,\ldots,1)^T \end{aligned}$$

with $P_H(0) = 0.5$*, and where the channel is the binary erasure channel described in Figure 2.11.*

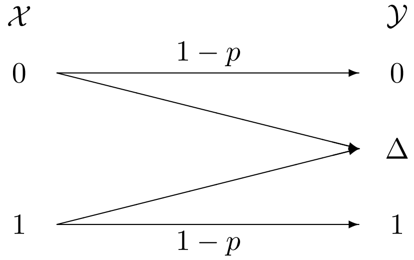


Figure 2.11: Binary erasure channel.

Evaluating the Bhattacharyya bound for this case yields:

$$\begin{aligned} Pr\{\boldsymbol{Y} \in \mathcal{B}_{0,1}|H=0\} &\le \sum_{\boldsymbol{y}\in\{0,1,\Delta\}^n} \sqrt{P_{\boldsymbol{Y}|H}(\boldsymbol{y}|1)P_{\boldsymbol{Y}|H}(\boldsymbol{y}|0)} \\ &= \sum_{\boldsymbol{y}\in\{0,1,\Delta\}^n} \sqrt{P_{\boldsymbol{Y}|\boldsymbol{X}}(\boldsymbol{y}|\boldsymbol{s}_1)P_{\boldsymbol{Y}|\boldsymbol{X}}(\boldsymbol{y}|\boldsymbol{s}_0)} \\ &\overset{(a)}{=} p^n, \end{aligned}$$

where in (a) *we used the fact that the first factor under the square root vanishes if* $\boldsymbol{y}$ *contains ones and the second vanishes if* $\boldsymbol{y}$ *contains zeros. Hence the only non-vanishing term in the sum is the one for which* $y_i = \Delta$ *for all* i*. The same bound applies for* $H = 1$*. Hence* $P_e \le \frac{1}{2}p^n + \frac{1}{2}p^n = p^n$*.*

If we use the tighter version of the union Bhattacharyya bound, which as mentioned earlier is tighter by a factor of 2*, then we obtain*

$$P_e \overset{\text{(UBB)}}{\le} \frac{1}{2}p^n.$$

For the Binary Erasure Channel and the two codewords $\boldsymbol{s}_0$ *and* $\boldsymbol{s}_1$ *we can actually compute the probability of error exactly:*

$$P_e = \frac{1}{2}Pr\{\boldsymbol{Y} = (\Delta,\Delta,\ldots,\Delta)^T\} = \frac{1}{2}p^n.$$

The Bhattacharyya bound is tight for the scenario considered in this example! □

2.7 Summary

The idea behind a MAP decision rule and the reason why it maximizes the probability that the decision is correct is quite intuitive. Let say we have two hypotheses, $H = 0$ and $H = 1$, with probability $P_H(0)$ and $P_H(1)$, respectively. If we have to guess which is the correct one without making any observation then we would choose the one that has the largest probability. This is quite intuitive yet let us repeat why. No matter how the decision is made, if it is $\hat{H} = i$ then the probability that it is correct is $P_H(i)$. Hence to maximize the probability that our decision is correct we choose $\hat{H} = \arg\max P_H(i)$. (If $P_H(0) = P_H(1) = 1/2$ then it does not matter how we decide: Either way, the probability that the decision is correct is $1/2$.)

The exact same idea applies *after* the receiver has observed the realization of the observable Y (or $\boldsymbol{Y}$). The only difference is that, after it observes $Y = y$, the receiver has an updated knowledge about the distribution of H. The new distribution is the *posterior* $P_{H|Y}(\cdot|y)$. In a typical example $P_H(i)$ may take the same value for all i whereas $P_{H|Y}(i|y)$ may be strongly biased in favor of one hypothesis. If it is strongly biased it means that the observable is very informative, which is what we hope of course.

Often $P_{H|Y}$ is not given but we can find it from P_H and $f_{Y|H}$ via Bayes' rule. While $P_{H|Y}$ is the most fundamental quantity associated to a MAP test and therefore it would make sense to write the test in terms of $P_{H|Y}$, the test is typically written in terms of P_H and $f_{Y|H}$ since those are normally the quantities that are specified as part of the model.

Notice that $f_{Y|H}$ and P_H is all we need to evaluate the union Bhattacharyya bound. Indeed the bound may be used in conjunction to any hypothesis testing problem not only for communication problems.

The following example shows how the posterior becomes more and more selective as the number of observations grows. It also shows that, as we would expect, the measurements are less informative if the channel is noisier.

Example 17. *Assume $H \in \{0, 1\}$ and $P_H(0) = P_H(1) = 1/2$. The outcome of H is communicated across a BSC of crossover probability $p < \frac{1}{2}$ via a transmitter that sends n zeros when $H = 0$ and n ones when $H = 1$. Letting k be the number of ones in the observed channel output $\boldsymbol{y}$ we have*

$$P_{\boldsymbol{Y}|H}(\boldsymbol{y}|i) = \begin{cases} p^k(1-p)^{n-k}, & H = 0 \\ p^{n-k}(1-p)^k, & H = 1. \end{cases}$$

Using Bayes rule,

$$P_{H|\boldsymbol{Y}}(i|\boldsymbol{y}) = \frac{P_{H,\boldsymbol{Y}}(i, \boldsymbol{y})}{P_{\boldsymbol{Y}}(\boldsymbol{y})} = \frac{P_H(i)P_{\boldsymbol{Y}|H}(\boldsymbol{y}|i)}{P_{\boldsymbol{Y}}(\boldsymbol{y})},$$

where $P_{\boldsymbol{Y}}(\boldsymbol{y}) = \sum_i P_{\boldsymbol{Y}|H}(\boldsymbol{y}|i)P_H(i)$ is the normalization that ensures $\sum_i P_{H|\boldsymbol{Y}}(i|\boldsymbol{y}) = 1$.

Hence

$$P_{H|\mathbf{Y}}(0|\mathbf{y}) = \frac{p^k(1-p)^{n-k}}{2P_{\mathbf{Y}}(\mathbf{y})} = \left(\frac{p}{1-p}\right)^k \frac{(1-p)^n}{2P_{\mathbf{Y}}(\mathbf{y})}$$
$$P_{H|\mathbf{Y}}(1|\mathbf{y}) = \frac{p^{n-k}(1-p)^k}{2P_{\mathbf{Y}}(\mathbf{y})} = \left(\frac{1-p}{p}\right)^k \frac{p^n}{2P_{\mathbf{Y}}(\mathbf{y})}.$$

Figure 2.12 depicts the behavior of $P_{H|\mathbf{Y}}(0|\mathbf{y})$ *as a function of the number* k *of* 1*s in* $\mathbf{y}$*. For the fist row* $n = 1$*, hence* k *may be* 0 *or* 1 *(abscissa). If* $p = .49$ *(left), the channel is very noisy and we don't learn much from the observation. Indeed we see that even if the single channel output is* 0 *(*$k = 0$ *in the figure) the posterior makes* $H = 0$ *only slightly more likely than* $H = 1$*. On the other hand if* $p = .25$ *the channel is less noisy which implies a more informative observation. Indeed we see (right top figure) that when* $k = 0$ *the posterior probability that* $H = 0$ *is significantly higher than the posterior probability that* $H = 1$*. In the bottom two figures the number of observations is* $n = 100$ *and the abscissa shows the number* k *of ones contained in the* 100 *observations. On the right (*$p = .25$*) we see that the posterior allows us to make a confident decision about* H *for almost all values of* k*. Uncertainty arises only when the number of observed ones roughly equals the number of zeros. On the other hand when* $p = .49$ *(bottom left figure) we can make a confident decision about* H *only if the observations contains a small number or a large number of* 1*s.*

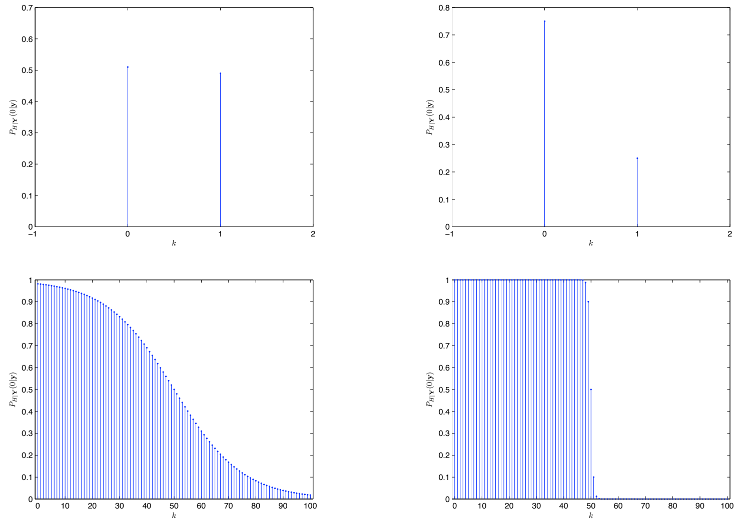


Figure 2.12: Posterior $P_{H|\mathbf{Y}}(0|\mathbf{y})$ as a function of the number k of observed 1s. The top row is for $n = 1$, $k = 0, 1$. The prior is more informative, and the decision more reliable, when $p = .25$ (right) than when $p = .49$ (left). The bottom row corresponds to $n = 100$. Now we see that we can make a reliable decision even if $p = .49$ (left), provided that k is sufficiently close to 0 or 100. When $p = .25$, as k goes from $k < \frac{n}{2}$ to $k > \frac{n}{2}$, the prior changes rapidly from being strongly in favor of $H = 0$ to strongly in favor of $H = 1$.

□

Appendix 2.A Facts About Matrices

We now review a few definitions and results that will be useful throughout. Hereafter $H^\dagger$ is the conjugate transpose of H also called the *Hermitian adjoint* of H.

Definition 18. *A matrix $U \in \mathbb{C}^{n\times n}$ is said to be* unitary *if $U^\dagger U = I$. If, in addition, $U \in \mathbb{R}^{n\times n}$, U is said to be orthogonal.* □

The following theorem lists a number of handy facts about unitary matrices. Most of them are straightforward. For a proof see [1, page 67].

Theorem 19. *if $U \in \mathbb{C}^{n\times n}$, the following are equivalent:*

(a) U is unitary;

(b) U is nonsingular and $U^\dagger = U^{-1}$;

(c) $UU^\dagger = I$;

(d) $U^\dagger$ is unitary

(e) The columns of U form an orthonormal set;

(f) The rows of U form an orthonormal set; and

(g) For all $\boldsymbol{x} \in \mathbb{C}^n$ the Euclidean length of $\boldsymbol{y} = U\boldsymbol{x}$ is the same as that of $\boldsymbol{x}$; that is, $\boldsymbol{y}^\dagger\boldsymbol{y} = \boldsymbol{x}^\dagger\boldsymbol{x}$.

Theorem 20. (Schur) *Any square matrix A can be written as $A = URU^\dagger$ where U is unitary and R is an upper-triangular matrix whose diagonal entries are the eigenvalues of A.*

Proof. Let us use induction on the size n of the matrix. The theorem is clearly true for $n = 1$. Let us now show that if it is true for $n-1$ it follows that it is true for n. Given A of size n, let $\boldsymbol{v}$ be an eigenvector of unit norm, and λ the corresponding eigenvalue. Let V be a unitary matrix whose first column is $\boldsymbol{v}$. Consider the matrix $V^\dagger AV$. The first column of this matrix is given by $V^\dagger A\boldsymbol{v} = \lambda V^\dagger\boldsymbol{v} = \lambda\boldsymbol{e}_1$ where $\boldsymbol{e}_1$ is the unit vector along the first coordinate. Thus

$$V^\dagger AV = \begin{pmatrix}\lambda & * \\ 0 & B\end{pmatrix},$$

where B is square and of dimension $n-1$. By the induction hypothesis $B = WSW^\dagger$, where W is unitary and S is upper triangular. Thus,

$$V^\dagger AV = \begin{pmatrix}\lambda & * \\ 0 & WSW^\dagger\end{pmatrix} = \begin{pmatrix}1 & 0 \\ 0 & W\end{pmatrix}\begin{pmatrix}\lambda & * \\ 0 & S\end{pmatrix}\begin{pmatrix}1 & 0 \\ 0 & W^\dagger\end{pmatrix} \tag{2.13}$$

and putting

$$U = V \begin{pmatrix} 1 & 0 \\ 0 & W \end{pmatrix} \quad \text{and} \quad R = \begin{pmatrix} \lambda & * \\ 0 & S \end{pmatrix},$$

we see that U is unitary, R is upper-triangular and $A = URU^\dagger$, completing the induction step. To see that the diagonal entries of R are indeed the eigenvalues of A it suffices to bring the characteristic polynomial of A in the following form: $\det(\lambda I - A) = \det\left[U^\dagger(\lambda I - R)U\right] = \det(\lambda I - R) = \prod_i(\lambda - r_{ii})$. □

Definition 21. *A matrix $H \in \mathbb{C}^{n\times x}$ is said to be* Hermitian *if $H = H^\dagger$. It is said to be* Skew-Hermitian *if $H = -H^\dagger$.*

Recall that an $n \times n$ matrix has exactly n eigenvalues in $\mathbb{C}$.

Lemma 22. *A Hermitian matrix $H \in \mathbb{C}^{n\times n}$ can be written as*

$$H = U\Lambda U^\dagger = \sum_i \lambda_i \boldsymbol{u}_i \boldsymbol{u}_i^\dagger$$

where U is unitary and $\Lambda = diag(\lambda_1, \ldots, \lambda_n)$ is a diagonal that consists of the eigenvalues of H. Moreover, the eigenvalues are real and the ith column of U is an eigenvector associated to λ_i.

Proof. By Theorem 20 (Schur) we can write $H = URU^\dagger$ where U is unitary and R is upper triangular with the diagonal elements consisting of the eigenvalues of A. From $R = U^\dagger HU$ we immediately see that R is Hermitian. Since it is also diagonal, the diagonal elements must be real.

If $\boldsymbol{u}_i$ is the ith column of U, then

$$H\boldsymbol{u}_i = U\Lambda U^\dagger \boldsymbol{u}_i = U\Lambda \boldsymbol{e}_i = U\lambda_i \boldsymbol{e}_i = \lambda_i \boldsymbol{u}_i$$

showing that it is indeed an eigenvector associated to the ith eigenvalue λ_i. □

The reader interested in properties of Hermitian matrices is referred to [1, Section 4.1].

Exercise 23. *Show that if $H \in \mathbb{C}^{n\times n}$ is Hermitian, then $\boldsymbol{u}^\dagger H\boldsymbol{u}$ is real for all $\boldsymbol{u} \in \mathbb{C}^n$.*

A class of Hermitian matrices with a special positivity property arises naturally in many applications, including communication theory. They provide a generalization to matrices of the notion of positive numbers.

Definition 24. *An Hermitian matrix $H \in \mathbb{C}^{n\times n}$ is said to be* positive definite *if*

$$\boldsymbol{u}^\dagger H\boldsymbol{u} > 0 \quad \textit{for all non zero} \quad \boldsymbol{u} \in \mathbb{C}^n.$$

If the above strict inequality is weakened to $\boldsymbol{u}^\dagger H\boldsymbol{u} \geq 0$, then A is said to be positive semidefinite. *Implicit in these defining inequalities is the observation that if H is Hermitian, the left hand side is always a real number.*

EXERCISE 25. *Show that a non-singular covariance matrix is always positive definite.*

THEOREM 26. (SVD) *Any matrix $A \in \mathbb{C}^{m \times n}$ can be written as a product*

$$A = UDV^{\dagger},$$

where U and V are unitary (of dimension $m \times m$ and $n \times n$, respectively) and $D \in \mathbb{R}^{m \times n}$ is non-negative and diagonal. This is called the singular value decomposition (SVD) *of A. Moreover, letting k be the rank of A, the following statements are true:*

(i) The columns of V are the eigenvectors of $A^{\dagger}A$. The last $n-k$ columns span the null space of A.

(ii) The columns of U are eigenvectors of $AA^{\dagger}$. The first k columns span the range of A.

(iii) If $m \geq n$ then

$$D = \begin{pmatrix} \operatorname{diag}(\sqrt{\lambda_1}, \ldots, \sqrt{\lambda_n}) \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ \mathbf{0}_{m-n} \end{pmatrix},$$

where $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_k > \lambda_{k+1} = \ldots = \lambda_n = 0$ are the eigenvalues of $A^{\dagger}A \in \mathbb{C}^{n \times n}$ which are non-negative since $A^{\dagger}A$ is Hermitian. If $m \leq n$ then

$$D = (\operatorname{diag}(\sqrt{\lambda_1}, \ldots, \sqrt{\lambda_m}) : \mathbf{0}_{n-m}),$$

where $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_k > \lambda_{k+1} = \ldots = \lambda_m = 0$ are the eigenvalues of $AA^{\dagger}$.

Note 1: Recall that the nonzero eigenvalues of AB equals the nonzero eigenvalues of BA, see e.g. Horn and Johnson, Theorem 1.3.29. Hence the nonzero eigenvalues in (iii) are the same for both cases.

Note 2: To remember that V is associated to $H^{\dagger}H$ (as opposed to being associated to $HH^{\dagger}$) it suffices to look at the dimensions: $V \in \mathbb{R}^{n}$ and $H^{\dagger}H \in \mathbb{R}^{n \times n}$.

Proof. It is sufficient to consider the case with $m \geq n$ since if $m < n$ we can apply the result to $A^{\dagger} = UDV^{\dagger}$ and obtain $A = VD^{\dagger}U^{\dagger}$.

Hence let $m \geq n$, and consider the matrix $A^{\dagger}A \in \mathbb{C}^{n \times n}$. This matrix is Hermitian. Hence its eigenvalues $\lambda_1 \geq \lambda_2 \geq \ldots \lambda_n \geq 0$ are real and non-negative and one can choose the eigenvectors $\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_n$ to form an orthonormal basis for $\mathbb{C}^{n}$. Let $V = (\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n)$. Let k be the number of positive eigenvectors and choose.

$$\boldsymbol{u}_i = \frac{1}{\sqrt{\lambda_i}} A\boldsymbol{v}_i, \qquad i = 1, 2, \ldots, k. \tag{2.14}$$

Observe that

$$\boldsymbol{u}_i^{\dagger}\boldsymbol{u}_j = \frac{1}{\sqrt{\lambda_i\lambda_j}} \boldsymbol{v}_i^{\dagger}A^{\dagger}A\boldsymbol{v}_j = \sqrt{\frac{\lambda_j}{\lambda_i}} \boldsymbol{v}_i^{\dagger}\boldsymbol{v}_j = \delta_{ij}, \quad 0 \leq i, j \leq k.$$

Hence $\{\boldsymbol{u}_i : i = 1, \dots, k\}$ form an orthonormal set in $\mathbb{C}^m$. Complete this set to an orthonormal basis for $\mathbb{C}^m$ by choosing $\{\boldsymbol{u}_i : i = k+1, \dots, m\}$ and let $U = (\boldsymbol{u}_1, \boldsymbol{u}_2, \dots, \boldsymbol{u}_m)$. Note that (2.14) implies

$$\boldsymbol{u}_i \sqrt{\lambda_i} = A\boldsymbol{v}_i, \quad i = 1, 2, \dots, k, k+1, \dots, n,$$

where for $i = k+1, \dots, n$ the above relationship holds since $\lambda_i = 0$ and $\boldsymbol{v}_i$ is a corresponding eigenvector. Using matrix notation we obtain

$$U \begin{pmatrix} \sqrt{\lambda_1} & & \mathbf{0} \\ & \ddots & \\ \mathbf{0} & & \sqrt{\lambda_n} \\ \cdots & \cdots & \cdots \\ & \mathbf{0}_{m-n} & \end{pmatrix} = AV, \tag{2.15}$$

i.e., $A = UDV^\dagger$. For $i = 1, 2, \dots, m$,

$$\begin{aligned} AA^\dagger \boldsymbol{u}_i &= UDV^\dagger V^\dagger D^\dagger U^\dagger \boldsymbol{u}_i \\ &= UDD^\dagger U^\dagger \boldsymbol{u}_i = \boldsymbol{u}_i \lambda_i, \end{aligned}$$

where the last equality follows from the fact that $U^\dagger \boldsymbol{u}_i$ has a 1 at position i and is zero otherwise and $DD^\dagger = \operatorname{diag}(\lambda_1, \lambda_2, \dots, \lambda_k, 0, \dots, 0)$. This shows that λ_i is also an eigenvalues of $AA^\dagger$. We have also shown that $\{\boldsymbol{v}_i : i = k+1, \dots, n\}$ spans the null space of A and from (2.15) we see that $\{\boldsymbol{u}_i : i = 1, \dots, k\}$ spans the range of A. □

The following key result is a simple application of the SVD.

Lemma 27. *The linear transformation described by a matrix $A \in \mathbb{R}^{n \times n}$ maps the unit cube into a parallelepiped of volume $|\det A|$.*

Proof. (Question to the students: do we need to review what a unit cube is, that the linear transformation maps $\boldsymbol{e}_i$ into the vector $\boldsymbol{a}_i$ that forms the i-th column of A, and that the volume of an n-dimensional object (set) $\mathcal{A}$ is $\int_{\mathcal{A}} d\boldsymbol{x}$?) From the singular value decomposition, $A = UDV^\dagger$, where D is diagonal and U and V are orthogonal matrices. The linear transformation associated to A is the same as that associated to $U^\dagger AV = D$. (We are just changing the coordinate system). But D maps the unit vectors $\boldsymbol{e}_1, \boldsymbol{e}_2, \dots, \boldsymbol{e}_n$ into $\lambda_1 \boldsymbol{e}_1, \lambda_2 \boldsymbol{e}_2, \dots, \lambda_n \boldsymbol{e}_n$. Hence, the unit cube is mapped into a rectangle of sides $\lambda_1, \lambda_2, \dots, \lambda_n$. Its volume is $|\prod \lambda_i| = |\det D| = |\det A|$. □

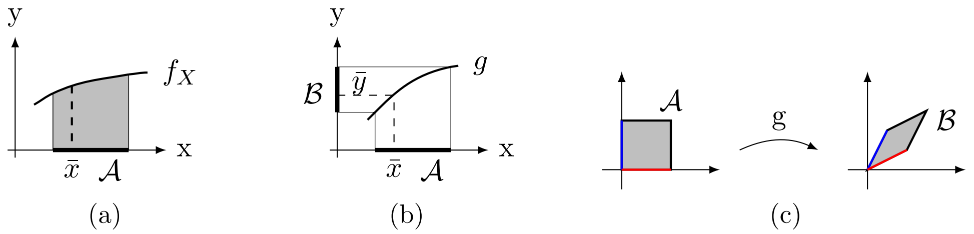


Figure 2.13: The role of a pdf (a); relationships between lengths in one-dimensional transformations (b); relationships between areas in two-dimensional transformations (c).

Appendix 2.B Densities After Linear Transformations

In this Appendix we outline how to determine the density of the random vector $\boldsymbol{Y}$ knowing the density of $\boldsymbol{X}$ and knowing that $\boldsymbol{Y} = g(\boldsymbol{X})$. This is an informal review. Or aim is to present the material in such a way that the reader sees what is gong on, hoping that in the future the student will be able to derive the density of a random variable defined in terms on another random variable without having to look up formulas.

We start with the scalar case. So X is a random variable of density f_X and $Y = g(X)$ for a given one-to-one and onto function $g : \mathcal{X} \to \mathcal{Y}$. Recall that a probability density function is to probability what pressure is to force: by integrating the probability density function over a subset $\mathcal{A}$ of $\mathcal{X}$ we obtain the probability that the event $\mathcal{A}$ occurs. If $\mathcal{A}$ is a small interval within $\mathcal{X}$ and it is small enough that we can consider f_X to be flat over $\mathcal{A}$, then $Pr\{X \in \mathcal{A}\} = f_X(\bar{x})l(\mathcal{A})$, where $l(\mathcal{A})$ denotes the length of the segment $\mathcal{A}$ and $\bar{x}$ is any point in $\mathcal{A}$. This is depicted in Fig. 2.13(a). The probability $Pr\{X \in \mathcal{A}\}$ is the shaded area, which tends to $f_X(\bar{x})l(\mathcal{A})$ as $l(\mathcal{A})$ goes to zero.

Now assume that g maps the interval $\mathcal{A}$ into the interval $\mathcal{B}$ of length $l(\mathcal{B})$ as shown in Fig. 2.13(b). The probability that $Y \in \mathcal{B}$ is the same as the probability that $X \in \mathcal{A}$. Hence f_Y must have the property

$$f_Y(\bar{y})l(\mathcal{B}) = f_X(\bar{x})l(\mathcal{A}),$$

where $\bar{y}$ is a point in $\mathcal{B}$ and $\bar{x} = g^{-1}(\bar{y})$ is the corresponding point in $\mathcal{A}$. We are making the assumption that $\mathcal{A}$ and $\mathcal{B}$ are small enough so that f_X is flat over $\mathcal{A}$ and f_Y is flat over $\mathcal{B}$. Solving we obtain

$$f_Y(\bar{y}) = f_X(\bar{x})\frac{l(\mathcal{A})}{l(\mathcal{B})}$$

From Fig.2.13(b) it is clear that in the limit of $l(\mathcal{A})$ and $l(\mathcal{B})$ becoming small we have $\frac{l(\mathcal{B})}{l(\mathcal{A})} = |g'(\bar{x})|$ where g' is the derivative of g. We have found that

$$f_Y(y) = \frac{f_X(g^{-1}(y))}{|g'(g^{-1}(y)|}$$

Example 28. *If $y = ax$ for some non-zero constant then*

$$f_Y(y) = \frac{f_X(\frac{y}{a})}{|a|}.$$

□

Next we consider the two-dimensional case. Let $\boldsymbol{X} = (X_1, X_2)^T$ have pdf $f_{\boldsymbol{X}}(\boldsymbol{x})$ and consider, as a start, the random vector $\boldsymbol{Y}$ obtained from the linear transformation

$$\boldsymbol{Y} = A\boldsymbol{X}$$

for some non-singular matrix A. The procedure to determine $f_{\boldsymbol{Y}}$ parallels the one for the scalar case. If $\mathcal{A}$ is a small rectangle, small enough that $f_{\boldsymbol{X}}(\boldsymbol{x})$ may be considered constant for all $\boldsymbol{X} \in \mathcal{A}$, then $Pr\{\boldsymbol{X} \in \mathcal{A}\}$ is approximated by $f_{\boldsymbol{X}}(\boldsymbol{x})a(\mathcal{A})$, where $a(\mathcal{A})$ is the area of $\mathcal{A}$. If $\mathcal{B}$ is the image of $\mathcal{A}$, then

$$f_{\boldsymbol{Y}}(\bar{\boldsymbol{y}})a(\mathcal{B}) = f_{\boldsymbol{X}}(\bar{\boldsymbol{x}})a(\mathcal{A})$$

where again we have made the assumption that $\mathcal{A}$ is small enough that $f_{\boldsymbol{X}}$ is constant for all $\boldsymbol{x} \in \mathcal{A}$ and $f_{\boldsymbol{Y}}$ is constant for all $\boldsymbol{y} \in \mathcal{B}$ and $\bar{\boldsymbol{x}} \in \mathcal{A}$ and $\bar{\boldsymbol{y}} \in \mathcal{B}$. Hence

$$f_{\boldsymbol{Y}}(\bar{\boldsymbol{y}}) = f_{\boldsymbol{X}}(\bar{\boldsymbol{x}})\frac{a(\mathcal{A})}{a(\mathcal{B})}.$$

For the next and final step you need to know that A maps surface $\mathcal{A}$ of area $a(\mathcal{A})$ into a surface $\mathcal{B}$ of area $a(\mathcal{B}) = a(\mathcal{A})|\det A|$. This fact, depicted in Fig. 2.13(c) for the two-dimensional case, is true in any number of dimensions n, but for $n \geq 3$ we speak of volume instead of area. The volume of $\mathcal{A}$ will be denoted by $\mathrm{Vol}(\mathcal{A})$. (The one-dimensional case is no special case: the determinant of a is a). See Lemma 27 Appendix 2.A for the outline of a proof that $\mathrm{Vol}(\mathcal{B}) = \mathrm{Vol}(\mathcal{A})|\det \mathcal{A}|$. Hence

$$f_{\boldsymbol{Y}}(\boldsymbol{y}) = \frac{f_{\boldsymbol{X}}(\mathcal{A}^{-1}\boldsymbol{y})}{|\det \mathcal{A}|}.$$

We are ready to generalize to the case

$$\bar{\boldsymbol{y}} = g(\bar{\boldsymbol{x}})$$

where g is one-to-one onto.

If we let $\mathcal{A}$ be a square of sides dx_1 and dx_2 that contains $\bar{\boldsymbol{x}}$, then the image of g will be a parallelepiped of sides dy_1 and dy_2 where

$$\begin{pmatrix} dy_1 \\ dy_2 \end{pmatrix} = J(\bar{\boldsymbol{x}}) \begin{pmatrix} dx_1 \\ dx_2 \end{pmatrix}$$

and $J = J(\bar{\boldsymbol{x}})$ is the Jacobian that at position i,j contains $\frac{\partial g_i}{\partial x_j}$ evalutated at $\bar{\boldsymbol{x}}$. The Jacobian $J(\boldsymbol{x})$ is the matrix that provides the linear approximation of g at $\boldsymbol{x}$.

Hence

$$f_{\bar{\boldsymbol{Y}}}(\bar{\boldsymbol{y}}) = \frac{f_{\boldsymbol{X}}(g^{-1}(\bar{\boldsymbol{y}}))}{|\det J(g^{-1}(\boldsymbol{y}))|}.$$

Sometimes the new random vector $\mathcal{Y}$ is described by the inverse function $\boldsymbol{x} = g^{-1}(\boldsymbol{y})$. There is no need to find g. The determinant of the Jacobian of g at $\boldsymbol{x} = g^{-1}(\boldsymbol{y})$ is one over the determinant of the Jacobian of g^{-1} at $\boldsymbol{y}$.

Example 29. (Rayleigh distribution) *Let X_1 and X_2 be two independent, zero-mean, unit-variance, Gaussian random variables. Let R and Θ be the corresponding polar coordinates, i.e., $X_1 = R\cos\Theta$ and $X_2 = R\sin\Theta$. We are interested in the probability density functions $f_{R,\Theta}$, f_R, and f_Θ. Since we are given the map g from (r,θ) to (x_1,x_2), we pretend that we know $f_{R,\Theta}$ and that we want to find f_{X_1,X_2}. Thus*

$$f_{X_1,X_2}(x_1,x_2) = \frac{1}{|\det J|} f_{R,\Theta}(r,\theta)$$

where J is the Jacobian of g, namely

$$J = \begin{pmatrix} \frac{\partial x_1(r,\theta)}{\partial r} & \frac{\partial x_1(r,\theta)}{\partial \theta} \\ \frac{\partial x_2(r,\theta)}{\partial \theta} & \frac{\partial x_2(r,\theta)}{\partial \theta} \end{pmatrix} = \begin{pmatrix} \cos\theta & -r\sin\theta \\ \sin\theta & r\cos\theta \end{pmatrix}.$$

Hence $\det J = r$ *and*

$$f_{X_1,X_2}(x_1,x_2) = \frac{1}{r} f_{R,\theta}(r,\theta).$$

Plugging in $f_{X_1,X_2}(x_1,x_2) = \frac{1}{2\pi}e^{-\frac{x_1^2+x_2^2}{2}}$, using $x_1^2 + x_2^2 = r^2$ to make it a function of the desired variables r,θ, and solving for $f_{R,\theta}$ we immediately obtain

$$f_{R,\theta}(r,\theta) = \frac{r}{2\pi}e^{-\frac{r^2}{2}}.$$

Since $f_{R,\Theta}(r,\theta)$ depends only on r we can immediately infer that R and Θ are independent random variables and that the latter is uniformly distributed in $[0,2\pi)$. Hence

$$f_\Theta(\theta) = \begin{cases} \frac{1}{2\pi} & \theta \in [0,2\pi) \\ 0 & \textit{otherwise} \end{cases}$$

and

$$f_R(r) = \begin{cases} re^{-\frac{r^2}{2}} & r \geq 0 \\ 0 & \textit{otherwise.} \end{cases}$$

We would have come to the same conclusion by integrating $f_{R,\Theta}$ over θ to obtain f_R and by integrating over r to obtain f_Θ. Notice that f_R is a Rayleigh probability density.

□

Appendix 2.C Gaussian Random Vectors

We now study Gaussian random vectors. A Gaussian random vector is nothing else than a collection of jointly Gaussian random variables. We learn to use vector notation since this will simplify matters significantly.

Recall that a random variable W is a mapping $W : \Omega \rightarrow \mathbb{R}$ from the sample space Ω to the reals $\mathbb{R}$. W is a Gaussian random variable with mean m and variance σ^2 if and only if (iff) its probability density function (pdf) is

$$f_W(w) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(w-m)^2}{2\sigma^2}\right\}.$$

Since a Gaussian random variable is completely specified by its mean m and variance σ^2, we use the short-hand notation $\mathcal{N}(m, \sigma^2)$ to denote its pdf. Hence $W \sim \mathcal{N}(m, \sigma^2)$.

An n*-dimensional random vector* (n-rv) $\boldsymbol{X}$ is a mapping $\boldsymbol{X} : \Omega \rightarrow \mathbb{R}^n$. It can be seen as a collection $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)^T$ of n random variables. The pdf of $\boldsymbol{X}$ is the joint pdf of $X_1, X_2, \ldots, X_n$. The expected value of $\boldsymbol{X}$, denoted by $E\boldsymbol{X}$ or by $\bar{\boldsymbol{X}}$, is the n-tuple $(EX_1, EX_2, \ldots, EX_n)^T$. The *covariance matrix* of $\boldsymbol{X}$ is $K_{\boldsymbol{X}} = E[(\boldsymbol{X} - \bar{\boldsymbol{X}})(\boldsymbol{X} - \bar{\boldsymbol{X}})^T]$. Notice that $\boldsymbol{X}\boldsymbol{X}^T$ is an $n \times n$ random matrix, i.e., a matrix of random variables, and the expected value of such a matrix is, by definition, the matrix whose components are the expected values of those random variables. Notice that a covariance matrix is always Hermitian.

The pdf of a vector $\boldsymbol{W} = (W_1, W_2, \ldots, W_n)^T$ that consists of independent and identically distributed (iid) $\sim \mathcal{N}(0, \sigma^2)$ components is

$$\begin{aligned} f_{\boldsymbol{W}}(\boldsymbol{w}) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{w_i^2}{2\sigma^2}\right) && (2.16) \\ &= \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{\boldsymbol{w}^T\boldsymbol{w}}{2\sigma^2}\right). && (2.17) \end{aligned}$$

The following is one of several possible ways to define a Gaussian random vector.

Definition 30. *The random vector $\boldsymbol{Y} \in \mathbb{R}^m$ is a zero-mean Gaussian random vector and $Y_1, Y_2, \ldots, Y_n$ are zero-mean jointly Gaussian random variables, iff there exists a matrix $A \in \mathbb{R}^{m \times n}$ such that $\boldsymbol{Y}$ can be expressed as*

$$\boldsymbol{Y} = A\boldsymbol{W} \qquad (2.18)$$

where $\boldsymbol{W}$ is a random vector of iid $\sim \mathcal{N}(0, 1)$ components.

Note 31. *From the above definition it follows immediately that linear combination of zero-mean jointly Gaussian random variables are zero-mean jointly Gaussian random variables. Indeed, $\boldsymbol{Z} = B\boldsymbol{Y} = BA\boldsymbol{W}$.* □

Recall from Appendix 2.B that if $\boldsymbol{Y} = A\boldsymbol{W}$ for some nonsingular matrix $A \in \mathbb{R}^{n\times n}$, then

$$f_{\boldsymbol{Y}}(\boldsymbol{y}) = \frac{f_{\boldsymbol{W}}(A^{-1}\boldsymbol{y})}{|\det A|}.$$

When $\boldsymbol{W}$ has iid $\sim \mathcal{N}(0,1)$ components,

$$f_{\boldsymbol{Y}}(\boldsymbol{y}) = \frac{\exp\left(-\frac{(A^{-1}\boldsymbol{y})^T(A^{-1}\boldsymbol{y})}{2}\right)}{(2\pi)^{n/2}|\det A|}.$$

The above expression can be simplified and brought to the standard expression

$$f_{\boldsymbol{Y}}(\boldsymbol{y}) = \frac{1}{\sqrt{(2\pi)^n \det K_{\boldsymbol{Y}}}} \exp\left(-\frac{1}{2}\boldsymbol{y}^T K_{\boldsymbol{Y}}^{-1}\boldsymbol{y}\right) \tag{2.19}$$

using $K_{\boldsymbol{Y}} = EAW(AW)^T = EAWW^TA^T = AI_nA^T = AA^T$ to obtain

$$\begin{aligned}(A^{-1}\boldsymbol{y})^T(A^{-1}\boldsymbol{y}) &= \boldsymbol{y}^T(A^{-1})^TA^{-1}\boldsymbol{y}\\ &= \boldsymbol{y}^T(AA^T)^{-1}\boldsymbol{y}\\ &= \boldsymbol{y}^TK_{\boldsymbol{Y}}^{-1}\boldsymbol{y}\end{aligned}$$

and

$$\sqrt{\det K_{\boldsymbol{Y}}} = \sqrt{\det AA^T} = \sqrt{\det A \det A} = |\det A|.$$

Fact 32. *Let $\boldsymbol{Y} \in \mathbb{R}^n$ be a zero-mean random vector with arbitrary covariance matrix $K_{\boldsymbol{Y}}$ and pdf as in (2.19). Since a covariance matrix is Hermitian, we we can write (see Appendix 2.A)*

$$K_{\boldsymbol{Y}} = U\Lambda U^\dagger \tag{2.20}$$

where U is unitary and Λ is diagonal. It is immediate to verify that $U\sqrt{\Lambda}\boldsymbol{W}$ has covariance $K_{\boldsymbol{Y}}$. This shows that an arbitrary zero-mean random vector $\boldsymbol{Y}$ with pdf as in (2.19) can always be written in the form $\boldsymbol{Y} = A\boldsymbol{W}$ where $\boldsymbol{W}$ has iid $\sim \mathcal{N}(0, I_n)$ components.

The contrary is not true in degenerated cases. We have already seen that (2.19) follows from (2.18) when A is a non-singular squared matrix. The derivation extends to any non-squared matrix A, provided that it has linearly independent rows. This result is derived as a homework exercise. In that exercise we also see that it is indeed necessary that the rows of A be linearly independent since otherwise $K_{\boldsymbol{Y}}$ is singular and $K_{\boldsymbol{Y}}^{-1}$ is not defined. Then (2.19) is not defined either. An example will show how to handle such degenerated cases.

It should be pointed out that many authors use (2.19) to define a Gaussian random vector. We favor (2.18) because it is more general, but also since it makes it straightforward to prove a number of key results associated to Gaussian random vectors. Some of these are dealt with in the examples below.

In any case, a zero-mean Gaussian random vector is completely characterized by its covariance matrix. Hence the short-hand notation $\boldsymbol{Y} \sim \mathcal{N}(0, K_{\boldsymbol{Y}})$.

Note 33. (Degenerate case) *Let $W \sim \mathcal{N}(0,1)$, $A = (1,1)^T$, and $Y = AW$. By our definition, Y is a Gaussian random vector. However, A is a matrix of linearly dependent rows implying that $\boldsymbol{Y}$ has linearly dependent components. Indeed $Y_1 = Y_2$. This also implies that $K_{\boldsymbol{Y}}$ is singular: it is a 2×2 matrix with 1 in each component. As already pointed out, we can't use (2.19) to describe the pdf of $\boldsymbol{Y}$. This immediately raises the question: how do we compute the probability of events involving $\boldsymbol{Y}$ if we don't know its pdf? The answer is easy. Any event involving $\boldsymbol{Y}$ can be rewritten as an event involving Y_1 only (or equivalently involving Y_2 only). For instance, the event $\{Y_1 \in [3,5]\} \cap \{Y_2 \in [4,6]\}$ occurs iff $\{Y_1 \in [4,5]\}$. Hence*

$$Pr\{Y_1 \in [3,5]\} \cap \{Y_2 \in [4,6]\} = Pr\{Y_1 \in [4,5]\} = Q(4) - Q(5).$$

□

Exercise 34. *Show that the ith component Y_i of a Gaussian random vector $\boldsymbol{Y}$ is a Gaussian random variable.*

Solution: $Y_i = A\boldsymbol{Y}$ when $A = \boldsymbol{e}_i^T$ is the unit row vector with 1 in the i-th component and 0 elsewhere. Hence Y_i is a Gaussian random variable. To appreciate the convenience of working with (2.18) instead of (2.19), compare this answer with the tedious derivation consisting of integrating over $f_{\boldsymbol{Y}}$ to obtain f_{Y_i} (see Problem 12). □

Exercise 35. *Let U be an orthogonal matrix. Determine the pdf of $\boldsymbol{Y} = U\boldsymbol{W}$.*

Solution: $\boldsymbol{Y}$ is zero-mean and Gaussian. Its covariance matrix is $K_{\boldsymbol{Y}} = UK_{\boldsymbol{W}}U^T = U\sigma^2 I_n U^T = \sigma^2 UU^T = \sigma^2 I_n$, where I_n denotes the $n \times n$ identiy matrix. Hence, when an n-dimensional Gaussian random vector with iid $\sim \mathcal{N}(0,\sigma^2)$ components is projected onto n orthonormal vectors, we obtain n iid $\sim \mathcal{N}(0,\sigma^2)$ random variables. This fact will be used often. □

Exercise 36. (Gaussian random variables are not necessarily jointly Gaussian) *Let $Y_1 \sim \mathcal{N}(0,1)$, let $X \in \{\pm 1\}$ be uniformly distributed, and let $Y_2 = Y_1 X$. Notice that Y_2 has the same pdf as Y_1. This follows from the fact that the pdf of Y_1 is an even function. Hence Y_1 and Y_2 are both Gaussian. However, they are not jointly Gaussian. We come to this conclusion by observing that $Z = Y_1 + Y_2 = Y_1(1+X)$ is 0 with probability 1/2. Hence Z can't be Gaussian.* □

Exercise 37. *Is it true that* uncorrelated Gaussian *random variables are always* independent*? If you think it is …think twice. The construction above labeled "Gaussian random variables are not necessarily jointly Gaussian" provides a counter example (you should be able to verify without much effort). However, the statement is true if the random variables under consideration are* jointly *Gaussian (the emphasis is on "jointly"). You should be able to prove this fact using (2.19). The contrary is always true: random variables (not necessarily Gaussian) that are independent are always uncorrelated. Again, you should be able to provide the straightforward proof. (You are strongly encouraged to brainstorm this and similar exercises with other students. Hopefully this will create healthy discussions. Let us know if you can't clear every doubt this way …we are very much interested in knowing where the difficulties are.)* □

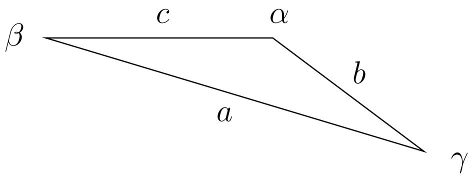
Definition 38. *The random vector* $\boldsymbol{Y}$ *is a Gaussian random vector (and* $Y_1, \ldots, Y_n$ *are jointly Gaussian random variables) iff* $\boldsymbol{Y} - m$ *is a zero mean Gaussian random vector as defined above, where* $m = E\boldsymbol{Y}$. *If the covariance* $K_{\boldsymbol{Y}}$ *is non-singular (which implies that no component of* $\boldsymbol{Y}$ *is determined by a linear combination of other components), then its pdf is*

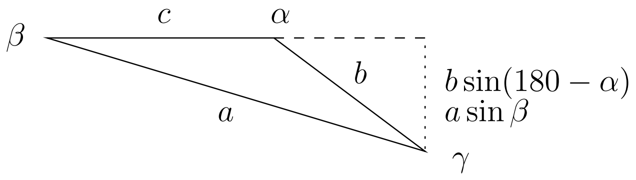
$$f_{\boldsymbol{Y}}(\boldsymbol{y}) = \frac{1}{\sqrt{(2\pi)^n \det K_{\boldsymbol{Y}}}} \exp\left(-\frac{1}{2}(\boldsymbol{y} - E\boldsymbol{y})^T K_{\boldsymbol{Y}}^{-1}(\boldsymbol{y} - E\boldsymbol{y})\right).$$

□

Appendix 2.D A Fact About Triangles

To determine an exact expression of the probability of error, in Example 15 we use the following fact about triangles.





For a triangle with edges a, b, c and angles α, β, γ (see the figure), the following relationship holds:

$$\frac{a}{\sin\alpha}=\frac{b}{\sin\beta}=\frac{c}{\sin\gamma}. \tag{2.21}$$

To prove the equality relating a and b we project the common vertex γ onto the extension of the segment connecting the other two edges (α and β). This projection gives rise to two triangles that share a common edge whose length can be written as $a\sin\beta$ and as $b\sin(180-\alpha)$ (see right figure). Using $b\sin(180-\alpha)=b\sin\alpha$ leads to $a\sin\beta=b\sin\alpha$. The second equality is proved similarly. □

Appendix 2.E Inner Product Spaces

Vector Space

We assume that you are familiar with vector spaces. In Chapter 2 we will be dealing with the vector space of n-tuples over $\mathbb{R}$ but later we will need both the vector space of n-tuples over $\mathbb{C}$ and the vector space of finite-energy complex-valued functions. To be as general as needed we assume that the vector space is over the field of complex numbers, in which case it is called a *complex vector space.* When the scalar field is $\mathbb{R}$, the vector space is called a *real vector space.*

Inner Product Space

Given a vector space and nothing more, one can introduce the notion of a basis for the vector space, but one does not have the tool needed to define an orthonormal basis. Indeed the axioms of a vector space say nothing about geometric ideas such as "length" or "angle." To remedy, one endows the vector space with the notion of inner product.

DEFINITION 39. *Let $\mathcal{V}$ be a vector space over $\mathbb{C}$. An inner product on $\mathcal{V}$ is a function that assigns to each ordered pair of vectors α,β in $\mathcal{V}$ a scalar $\langle\alpha,\beta\rangle$ in $\mathbb{C}$ in such a way* *that for all α, β, γ in $\mathcal{V}$ and all scalars c in $\mathbb{C}$*

(a) $\langle \alpha + \beta, \gamma \rangle = \langle \alpha, \gamma \rangle + \langle \beta, \gamma \rangle$
$\langle c\alpha, \beta \rangle = c\langle \alpha, \beta \rangle$;

(b) $\langle \beta, \alpha \rangle = \langle \alpha, \beta \rangle^*$; *(Hermitian Symmertry)*

(c) $\langle \alpha, \alpha \rangle \geq 0$ *with equality iff* $\alpha = 0$.

It is implicit in (c) that $\langle \alpha, \alpha \rangle$ is real for all $\alpha \in \mathcal{V}$. From (a) and (b), we obtain an additional property

(d) $\langle \alpha, \beta + \gamma \rangle = \langle \alpha, \beta \rangle + \langle \alpha, \gamma \rangle$
$\langle \alpha, c\beta \rangle = c^*\langle \alpha, \beta \rangle$.

Notice that the above definition is also valid for a vector space over the field of real numbers but in this case the complex conjugates appearing in (b) and (d) are superfluous; however, over the field of complex numbers they are necessary for the consistency of the conditions. Without these complex conjugates, for any $\alpha \neq 0$ we would have the contradiction:

$$0 < \langle i\alpha, i\alpha \rangle = -1\langle \alpha, \alpha \rangle < 0,$$

where the first inequality follows from condition (c) and the fact that $i\alpha$ is a valid vector, and the equality follows from (a) and (d) (without the complex conjugate).

On $\mathbb{C}^n$ there is an inner product that is sometimes called the *standard inner product.* It is defined on $\boldsymbol{a} = (a_1, \ldots, a_n)$ and $\boldsymbol{b} = (b_1, \ldots, b_n)$ by

$$\langle \boldsymbol{a}, \boldsymbol{b} \rangle = \sum_j a_j b_j^*.$$

On $\mathbb{R}^n$, the standard inner product is often called the dot or scalar product and denoted by $\boldsymbol{a} \cdot \boldsymbol{b}$. Unless explicitly stated otherwise, over $\mathbb{R}^n$ and over $\mathbb{C}^n$ we will always assume the standard inner product.

An *inner product space* is a real or complex vector space, together with a specified inner product on that space. We will use the letter $\mathcal{V}$ to denote a generic inner product space.

Example 40. *The vector space $\mathbb{R}^n$ equipped with the dot product is an inner product space and so is the vector space $\mathbb{C}^n$ equipped with the standard inner product.* □

By means of the inner product we introduce the notion of length, called *norm,* of a vector α, via

$$\|\alpha\| = \sqrt{\langle \alpha, \alpha \rangle}.$$

Using linearity, we immediately obtain that the *squared norm* satisfies

$$\|\alpha \pm \beta\|^2 = \langle \alpha \pm \beta, \alpha \pm \beta \rangle = \|\alpha\|^2 + \|\beta\|^2 \pm 2Re\{\langle \alpha, \beta \rangle\}. \tag{2.22}$$

The above generalizes $(a \pm b)^2 = a^2 + b^2 \pm 2ab$, $a, b \in \mathbb{R}$, and $|a \pm b|^2 = |a|^2 + |b|^2 \pm 2Re\{ab\}$, $a, b \in \mathbb{C}$.

Example 41. *Consider the vector space $\mathcal{V}$ spanned by a finite collection of complex-valued finite-energy signals, where addition of vectors and multiplication of a vector with a scalar (in $\mathbb{C}$) are defined in the obvious way. You should verify that the axioms of a vector space are fulfilled. This includes showing that the sum of two finite-energy signals is a finite-energy signal. The standard inner product for this vectors space is defined as*

$$\langle \alpha, \beta \rangle = \int \alpha(t) \beta^*(t) dt$$

which implies the norm

$$\|\alpha\| = \sqrt{\int |\alpha(t)|^2 dt}.$$

□

Example 42. *The previous example extends to the inner product space $\mathcal{L}_2$ of all complex-valued finite-energy functions. This is an infinite dimensional inner product space and to be careful one has to deal with some technicalities that we will just mention here. (If you wish you may skip the rest of this example without loosing anything important for the sequel). If α and β are two finite-energy functions that are identical except on a countable number of points, then $\langle \alpha - \beta, \alpha - \beta \rangle = 0$ (the integral is over a function that vanishes except for a countable number of points). The definition of inner product requires that $\alpha - \beta$ be the zero vector. This seems to be in contradiction with the fact that $\alpha - \beta$ is non-zero on a countable number of points. To deal with this apparent contradiction one can define vectors to be equivalence classes of finite-energy functions. In other words, if the norm of $\alpha - \beta$ vanishes then α and β are considered to be the same vector and $\alpha - \beta$ is seen as a zero vector. This equivalence may seem artificial at first but it is actually consistent with the reality that if $\alpha - \beta$ has zero energy then no instrument will be able to distinguish between α and β. The signal captured by the antenna of a receiver is finite energy, thus in $\mathcal{L}_2$. It is for this reason that we are interested in $\mathcal{L}_2$.* □

Theorem 43. *If $\mathcal{V}$ is an inner product space, then for any vectors α, β in $\mathcal{V}$ and any scalar c,*

(a) $\|c\alpha\| = |c|\|\alpha\|$

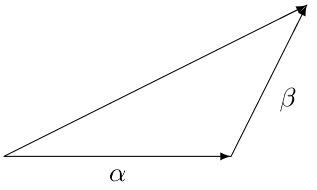
(b) $\|\alpha\| \geq 0$ *with equality iff* $\alpha = 0$

(c) $|\langle \alpha, \beta \rangle| \leq \|\alpha\|\|\beta\|$ *with equality iff* $\alpha = c\beta$ *for some* c.
(Cauchy-Schwarz inequality)

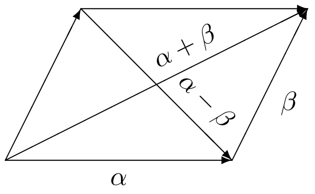
(d) $\|\alpha + \beta\| \leq \|\alpha\| + \|\beta\|$ *with equality iff* $\alpha = c\beta$ *for some non-negative* $c \in \mathbb{R}$.
(Triangle inequality)

(e) $\|\alpha + \beta\|^2 + \|\alpha - \beta\|^2 = 2(\|\alpha\|^2 + \|\beta\|^2)$
(Parallelogram equality)

Proof. *Statements (a) and (b) follow immediately from the definitions. We postpone the proof of the Cauchy-Schwarz inequality to Example 45 since it will be more insightful once we have defined the concept of a projection. To prove the triangle inequality we use (2.22) and the Cauchy-Schwarz inequality applied to* $Re\{\langle \alpha, \beta \rangle\} \leq |\langle \alpha, \beta \rangle|$ *to prove that* $\|\alpha + \beta\|^2 \leq (\|\alpha\| + \|\beta\|)^2$. *You should verify that* $Re\{\langle \alpha, \beta \rangle\} \leq |\langle \alpha, \beta \rangle|$ *holds with equality iff* $\alpha = c\beta$ *for some non-negative* $c \in \mathbb{R}$. *Hence this condition is necessary for the triangle inequality to hold with equality. It is also sufficient since then also the Cauchy-Schwarz inequality holds with equality. The parallelogram equality follows immediately from (2.22) used twice, once with each sign.* □



Triangle inequality



Parallelogram equality

At this point we could use the inner product and the norm to define the angle between two vectors but we don't have any use for that. Instead, we will make frequent use of the notion of orthogonality. Two vectors α and β are defined to be *orthogonal* if $\langle \alpha, \beta \rangle = 0$.

Theorem 44. (Pythagorean Theorem) *If* α *and* β *are orthogonal vectors in* $\mathcal{V}$, *then*

$$\|\alpha + \beta\|^2 = \|\alpha\|^2 + \|\beta\|^2.$$

Proof. The Pythagorean theorem follows immediately from the equality $\|\alpha + \beta\|^2 = \|\alpha\|^2 + \|\beta\|^2 + 2Re\{\langle \alpha, \beta \rangle\}$ and the fact that $\langle \alpha, \beta \rangle = 0$ by definition of orthogonality. □

Given two vectors $\alpha, \beta \in \mathcal{V}$, $\beta \neq 0$, we define the *projection of* α *on* β as the vector $\alpha_{|\beta}$ collinear to β (i.e. of the form $c\beta$ for some scalar c) such that $\alpha_{\perp\beta} = \alpha - \alpha_{|\beta}$ is orthogonal to β. Using the definition of orthogonality, what we want is

$$0 = \langle \alpha_{\perp\beta}, \beta \rangle = \langle \alpha - c\beta, \beta \rangle = \langle \alpha, \beta \rangle - c\|\beta\|^2.$$

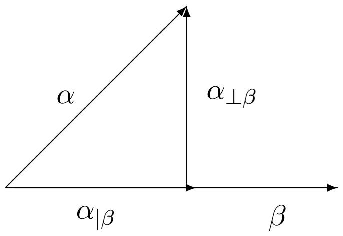
Solving for c we obtain $c = \frac{\langle \alpha, \beta \rangle}{\|\beta\|^2}$. Hence

$$\alpha_{|\beta} = \frac{\langle \alpha, \beta \rangle}{\|\beta\|^2}\beta \qquad \text{and} \qquad \alpha_{\perp\beta} = \alpha - \alpha_{|\beta}.$$

The projection of α on β does not depend on the norm of β. To see this let $\beta = b\psi$ for some $b \in \mathbb{C}$. Then

$$\alpha_{|\beta} = \langle \alpha, \psi \rangle \psi = \alpha_{|\psi},$$

regardless of b. It is immediate to verify that the norm of the projection is $|\langle \alpha, \psi \rangle| = \frac{|\langle \alpha, \beta \rangle|}{\|\beta\|}$.



Projection of α on β

Any non-zero vector β defines a *hyperplane* by the relationship

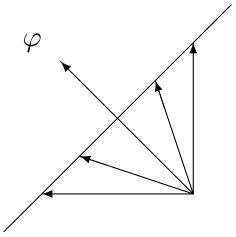
$$\{\alpha \in \mathcal{V} : \langle \alpha, \beta \rangle = 0\}.$$

It is the set of vectors that are orthogonal to β. A hyperplane always contains the zero vector.

An *affine space*, defined by a vector β and a scalar c, is an object of the form

$$\{\alpha \in \mathcal{V} : \langle \alpha, \beta \rangle = c\}.$$

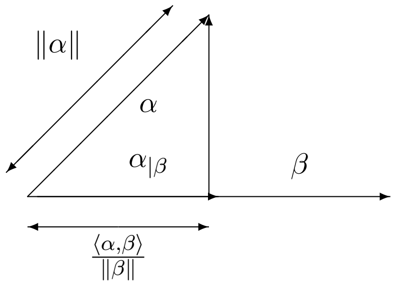
The defining vector and scalar are not unique, unless we agree that we use only normalized vectors to define hyperplanes. By letting $\varphi = \frac{\beta}{\|\beta\|}$, the above definition of affine plane may equivalently be written as $\{\alpha \in \mathcal{V} : \langle \alpha, \varphi \rangle = \frac{c}{\|\beta\|}\}$ or even as $\{\alpha \in \mathcal{V} : \langle \alpha - \frac{c}{\|\beta\|}\varphi, \varphi \rangle = 0\}$. The first shows that at an affine plane is the set of vectors that have the same projection $\frac{c}{\|\beta\|}\varphi$ on φ. The second form shows that the affine plane is a hyperplane translated by the vector $\frac{c}{\|\beta\|}\varphi$. Some authors make no distinction between affine planes and hyperplanes. In that case both are called hyperplane.



Affine plane defined by φ.

Now it is time to prove the Cauchy-Schwarz inequality stated in Theorem 43. We do it as an application of a projection.

Example 45. (Proof of the Cauchy-Schwarz Inequality). *The Cauchy-Schwarz inequality states that for any* $\alpha, \beta \in \mathcal{V}$, $|\langle \alpha, \beta \rangle| \leq \|\alpha\|\|\beta\|$ *with equality iff* $\alpha = c\beta$ *for some scalar* $c \in \mathbb{C}$. *The statement is obviously true if* $\beta = 0$. *Assume* $\beta \neq 0$ *and write* $\alpha = \alpha_{|\beta} + \alpha_{\perp\beta}$. *The Pythagorean theorem states that* $\|\alpha\|^2 = \|\alpha_{|\beta}\|^2 + \|\alpha_{\perp\beta}\|^2$. *If we drop the second term, which is always nonnegative, we obtain* $\|\alpha\|^2 \geq \|\alpha_{|\beta}\|^2$ *with equality iff* α *and* β *are collinear. From the definition of projection,* $\|\alpha_{|\beta}\|^2 = \frac{|\langle \alpha, \beta \rangle|^2}{\|\beta\|^2}$. *Hence* $\|\alpha\|^2 \geq \frac{|\langle \alpha, \beta \rangle|^2}{\|\beta\|^2}$ *with equality equality iff* α *and* β *are collinear. This is the Cauchy-Schwarz inequality.*
□



The Cauchy-Schwarz inequality

Every finite-dimensional vector space has a basis. If $\beta_1, \beta_2, \ldots, \beta_n$ is a basis for the inner product space $\mathcal{V}$ and $\alpha \in \mathcal{V}$ is an arbitrary vector, then there are scalars $a_1, \ldots, a_n$ such that $\alpha = \sum a_i \beta_i$ but finding them may be difficult. However, finding the coefficients of a vector is particularly easy when the basis is orthonormal.

A basis $\varphi_1, \varphi_2, \ldots, \varphi_n$ for an inner product space $\mathcal{V}$ is orthonormal if

$$\langle \varphi_i, \varphi_j \rangle = \begin{cases} 0, & i \neq j \\ 1, & i = j. \end{cases}$$

Finding the i-th coefficient a_i of an orthonormal expansion $\alpha = \sum a_i \psi_i$ is immediate. It suffices to observe that all but the ith term of $\sum a_i \psi_i$ are orthogonal to ψ_i and that the inner product of the ith term with ψ_i yields a_i. Hence if $\alpha = \sum a_i \psi_i$ then

$$a_i = \langle \alpha, \psi_i \rangle.$$

Observe that $|a_i|$ is the norm of the projection of α on ψ_i. This should not be surprising given that the ith term of the orthonormal expansion of α is collinear to ψ_i and the sum of all the other terms are orthogonal to ψ_i.

There is another major advantage of working with an orthonormal basis. If $\boldsymbol{a}$ and $\boldsymbol{b}$ are the n-tuples of coefficients of the expansion of α and β with respect to the same orthonormal basis then

$$\langle \alpha, \beta \rangle = \langle \boldsymbol{a}, \boldsymbol{b} \rangle$$

where the right hand side inner product is with respect to the standard inner product. Indeed

$$\begin{aligned} \langle \alpha, \beta \rangle &= \langle \sum a_i \psi_i, \sum_j b_j \psi_j \rangle = \sum a_i \langle \psi_i, \sum_j b_j \psi_j \rangle \\ &= \sum a_i \langle \psi_i, b_i \psi_i \rangle = \sum a_i b_i^* = \langle \boldsymbol{a}, \boldsymbol{b} \rangle. \end{aligned}$$

Letting $\beta = \alpha$ the above implies also

$$\|\alpha\| = \|\boldsymbol{a}\|,$$

where the right hand side is the standard norm $\|a\| = \sum |a_i|^2$.

An orthonormal set of vectors $\psi_1, \ldots, \psi_n$ of an inner product space $\mathcal{V}$ is a linearly independent set. Indeed $0 = \sum a_i \psi_i$ implies $a_i = \langle 0, \psi_i \rangle = 0$. By normalizing the vectors and recomputing the coefficients one can easily extend this reasoning to a set of orthogonal (but not necessarily orthonormal) vectors $\alpha_1, \ldots, \alpha_n$. They too must be linearly independent.

The idea of a projection on a vector generalizes to a projection on a subspace. If $\mathcal{W}$ is a subspace of an inner product space $\mathcal{V}$, and $\alpha \in \mathcal{V}$, the projection of α on $\mathcal{W}$ is defined to be a vector $\alpha_{|\mathcal{W}} \in \mathcal{W}$ such that $\alpha - \alpha_{|\mathcal{W}}$ is orthogonal to all vectors in $\mathcal{W}$. If $\psi_1, \ldots, \psi_m$ is an orthonormal basis for $\mathcal{W}$ then the condition that $\alpha - \alpha_{|\mathcal{W}}$ is orthogonal to all vectors of $\mathcal{W}$ implies $0 = \langle \alpha - \alpha_{|\mathcal{W}}, \psi_i \rangle = \langle \alpha, \psi_i \rangle - \langle \alpha_{|\mathcal{W}}, \psi_i \rangle$. This shows that $\langle \alpha, \psi_i \rangle = \langle \alpha_{|\mathcal{W}}, \psi_i \rangle$. The right side of this equality is the i-th coefficient of the orthonormal expansion of $\alpha_{|\mathcal{W}}$ with respect to the orthonormal basis. This proves that

$$\alpha_{|\mathcal{W}} = \sum_{i=1}^{m} \langle \alpha, \psi_i \rangle \psi_i$$

is the unique projection of α on $\mathcal{W}$.

Theorem 46. *Let $\mathcal{V}$ be an inner product space and let $\beta_1, \ldots, \beta_n$ be any collection of linearly independent vectors in $\mathcal{V}$. Then one may construct orthogonal vectors $\alpha_1, \ldots, \alpha_n$ in $\mathcal{V}$ such that they form a basis for the subspace spanned by $\beta_1, \ldots, \beta_n$.*

Proof. The proof is constructive via a procedure known as the Gram-Schmidt orthogonalization procedure. First let $\alpha_1 = \beta_1$. The other vectors are constructed inductively as follows. Suppose $\alpha_1, \ldots, \alpha_m$ have been chosen so that they form an orthogonal basis for the subspace $\mathcal{W}_m$ spanned by $\beta_1, \ldots, \beta_m$. We choose the next vector as

$$\alpha_{m+1} = \beta_{m+1} - \beta_{m+1|\mathcal{W}_m}, \tag{2.23}$$

where $\beta_{m+1|\mathcal{W}_m}$ is the projection of β_{m+1} on $\mathcal{W}_m$. By definition, α_{m+1} is orthogonal to every vector in $\mathcal{W}_m$, including $\alpha_1, \ldots, \alpha_m$. Also, $\alpha_{m+1} \neq 0$ for otherwise β_{m+1} contradicts the hypothesis that it is lineary independent of $\beta_1, \ldots, \beta_m$. Therefore $\alpha_1, \ldots, \alpha_{m+1}$ is an orthogonal collection of nonzero vectors in the subspace $\mathcal{W}_{m+1}$ spanned by $\beta_1, \ldots, \beta_{m+1}$. Therefore it must be a basis for $\mathcal{W}_{m+1}$. Thus the vectors $\alpha_1, \ldots, \alpha_n$ may be constructed one after the other according to (2.23). □

Corollary 47. *Every finite-dimensional vector space has an orthonormal basis.*

Proof. Let $\beta_1, \ldots, \beta_n$ be a basis for the finite-dimensionall inner product space $\mathcal{V}$. Apply the Gram-Schmidt procedure to find an orthogonal basis $\alpha_1, \ldots, \alpha_n$. Then $\psi_1, \ldots, \psi_n$, where $\psi_i = \frac{\alpha_i}{\|\alpha_i\|}$, is an orthonormal basis. □

Gram-Schmidt Orthonormalization Procedure

We summarize the Gram-Schmidt procedure, modified so as to produce orthonormal vectors. If $\beta_1, \ldots, \beta_n$ is a linearly independent collection of vectors in the inner product space $\mathcal{V}$ then we may construct a collection $\psi_1, \ldots, \psi_n$ that forms an orthonormal basis for the subspace spanned by $\beta_1, \ldots, \beta_n$ as follows: we let $\psi_1 = \frac{\beta_1}{\|\beta_1\|}$ and for $i = 2, \ldots, n$ we choose

$$\begin{aligned}\alpha_i &= \beta_i - \sum_{j=1}^{i-1} \langle \beta_i, \psi_j \rangle \psi_j \\ \psi_i &= \frac{\alpha_i}{\|\alpha_i\|}.\end{aligned}$$

We have assumed that $\beta_1, \ldots, \beta_n$ is a linearly independent collection. Now assume that this is not the case. If β_j is linearly dependent of $\beta_1, \ldots, \beta_{j-1}$, then at step $i = j$ the procedure will produce $\alpha_i = \psi_i = 0$. Such vectors are simply disregarded.

The following table gives an example of the Gram-Schmidt procedure.

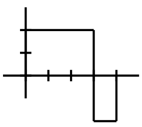
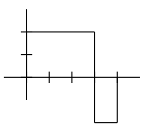
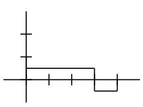
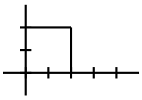
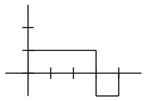
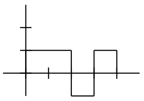
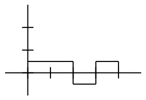
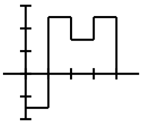
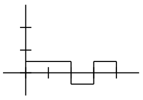
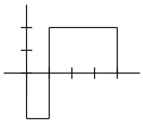
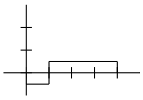
i	β_i	$\langle \beta_i, \psi_j \rangle$ $j < i$	$\beta_{i\mid\mathcal{W}_{i-1}}$	$\alpha_i = \beta_i - \beta_{i\mid\mathcal{W}_{i-1}}$	$\\|\alpha_i\\|$	ψ_i	$\boldsymbol{\beta}_i$
1		-	-		2		$\begin{pmatrix}2\\0\\0\end{pmatrix}$
2		1			1		$\begin{pmatrix}1\\1\\0\end{pmatrix}$
3		$0, 1$			4		$\begin{pmatrix}0\\1\\4\end{pmatrix}$

Table 2.1: Application of the Gram-Schmidt orthonormalization procedure. Axes are marked with unit length intervals. The starting point is the second column.

This page is kept empty

Appendix 2.F Problems

PROBLEM 1. (Probabilities of Basic Events) *Assume that* X_1 *and* X_2 *are independent random variables uniformly distributed in the interval* $[0,1]$. *Compute the probability of the following events:*

(a) $-\frac{1}{2} \leq X_1 - X_2 \leq \frac{1}{2}$.

(b) $X_2 \geq X_1^2$.

(c) $X_2 = X_1$.

(d) $X_1 + X_2 \leq 1$ *and* $X_1 \geq \frac{1}{2}$.

(e) Given that $X_1 \geq \frac{1}{2}$, *compute the probability that* $X_1 + X_2 \leq 1$.

Hint: For each event, identify the corresponding region inside the unit square.

PROBLEM 2. (Uncorrelated vs. Independent Random Variables) *Let* X *and* Y *be two continuous real-valued random variables with joint probability density function* p_{XY}.

(a) When are X *and* Y *uncorrelated? When are they independent? Write down the definitions.*

(b) Show that if X *and* Y *are independent, they are also uncorrelated.*

(c) Consider two independent and uniformly distributed random variables $U \in \{0,1\}$ *and* $V \in \{0,1\}$. *Assume that* X *and* Y *are defined as follows:* $X = U + V$ *and* $Y = |U - V|$. *Are* X *and* Y *independent? Compute the covariance of* X *and* Y. *What do you conclude?*

PROBLEM 3. (Bolt Factory) *In a bolt factory machines A, B, C manifacture, respectively* 25, 35 *and* 40 *per cent of the total. Of their product* 5, 4, *and* 2 *per cent are defective bolts. A bolt is drawn at random from the produce and is found defective. What are the probabilities that it was manufactured by machines A, B and C?*

Note: The question is taken from the book "An introduction to Probability Theory and Its Applications" *by William Feller.*

Problem 4. (One of Three) *Assume you are at a quiz show. You are shown three boxes which look identical from the outside, except they have labels 0, 1, and 2, respectively. Exactly one of them contains one million Swiss francs, the other two contain nothing. You choose one box at random with a uniform probability. Let A be the random variable which denotes your choice, $A \in \{0, 1, 2\}$.*

(a) *What is the probability that the box A contains the money?*

The quizmaster knows in which box the money is and he now opens from the remaining two boxes one that does not contain the prize. This means that if neither of the two remaining boxes contain the prize then the quizmaster opens one with uniform probability. Otherwise, he simply opens the one which does not contain the prize. Let B denote the random variable corresponding to the box that remains closed after the elimination by the quizmaster.

(b) *What is the probability that B contains the money?*

(c) *If you are now allowed to change your mind, i.e., choose B instead of sticking with A, would you do it?*

Problem 5. (The "Wetterfrosch")

Let us assume that a "weather frog" bases his forecast for tomorrow's weather entirely on today's air pressure. Determining a weather forecast is a hypothesis testing problem. For simplicity, let us assume that the weather frog only needs to tell us if the forecast for tomorrow's weather is "sunshine" or "rain". Hence we are dealing with a binary hypothesis testing problem. Let $H = 0$ mean "sunshine" and $H = 1$ mean "rain". We will assume that both values of H are equally likely, i.e. $P_H(0) = P_H(1) = 1/2$.

Measurements over several years have led the weather frog to conclude that on a day that precedes sunshine the pressure may be modeled as a random variable y with the following probability density function:

$$f_{Y|H}(y|0) = \begin{cases} A - \frac{A}{2}y, & 0 \le y \le 1 \\ 0, & \textit{otherwise.} \end{cases}$$

Similarly, the pressure on a day that precedes a rainy day is distributed according to

$$f_{Y|H}(y|1) = \begin{cases} B + \frac{B}{3}y, & 0 \le y \le 1 \\ 0, & \textit{otherwise.} \end{cases}$$

The weather frog's goal in life is to guess the value of H after measuring Y.

(a) *Determine A and B.*

(b) *Find the probability $P_{H|Y}(0|y)$ for all values of y. This probability is often called the* a posteriori probability *of hypothesis $H = 0$ given that $Y = y$. Also find the probability $P_{H|Y}(1|y)$ for all values of y.* Hint: *Use Bayes' rule.*

(c) Plot $P_{H|Y}(0|y)$ *and* $p_{H|Y}(1|y)$ *as a function of* y. *Is it true that the decision rule may be written as*

$$\hat{H}(y) \;=\; \begin{cases} 0, & \text{if } y \leq \theta \\ 1, & \text{otherwise,} \end{cases}$$

for some threshold θ*? If yes specify* θ.

(d) Determine, as a function of θ, *the probability that the decision rule in* (iii) *decides* $\hat{H} = 1$ *when, in reality,* $H = 0$. *This probability is denoted* $Pr\{\hat{H}(y) = 1|H = 0\}$.

(e) Determine, as a function of θ, *the probability of error for the decision rule that you have derived in* (iii). *Evaluate your expression at the value of* θ *that you have found in* (iii).

(f) Among decision rules that compare the pressure y *to a threshold like in Eqn. (2.24), is there a decision rule that results in a smaller probability of error than the rule derived in* (iii)*? You should be able to answer this question without further calculations. However, to double check, find the* θ *that maximizes the expression you have found in part* (iv).

PROBLEM 6. (Alternative "Wetterfrosch") *A TV "weather frog" bases his weather forecast for tomorrow entirely on today's air pressure, which is thus his observable* Y. *Here, we consider an ambitious weather frog who wants to distinguish three kinds of weather. This means, that tomorrow's weather is represented by a random variable* H *which take on value 0 if the sun shines tomorrow, 1 if it rains or 2 if the weather is unstable. We assume that the three hypotheses are a priori equally likely, i.e.* $P_H(0) = P_H(1) = P_H(2) = 1/3$.

Measurements over several years have led to the following estimate of the probability density function of today's air pressure provided that the sun shines tomorrow,

$$f_{Y|H}(y|0) = \begin{cases} A - 2Ay & , \; 0 \leq y \leq 0.5 \\ 0 & , \; \text{otherwise.} \end{cases}$$

The estimate of the probability density function of today's air pressure provided that it rains tomorrow, is

$$f_{Y|H}(y|1) = \begin{cases} B + \frac{B}{2}y & , \; 0 \leq y \leq 1 \\ 0 & , \; \text{otherwise.} \end{cases}$$

Finally, the estimate of the probability density function of today's air pressure provided that the weather is unstable tomorrow, is

$$f_{Y|H}(y|2) = \begin{cases} C & , \; 0 \leq y \leq 1 \\ 0 & , \; \text{otherwise.} \end{cases}$$

The weather frog's goal is to guess the value of H *after measuring* Y.

(a) *Determine* A, B *and* C.

(b) *Write down the optimal decision rule (i.e. the rule that minimize the probability of a wrong forecast) in general terms.*

(c) *For all values* y, *draw into one graph* $f_{y|H}(y|0)$, $f_{y|H}(y|1)$ *and* $f_{y|H}(y|2)$. *Show on the graph the decision regions corresponding to the optimal decision rule. If we let* $\hat{H}(y)$ *denote the frog's forecast for a value* y *of the measurement, can the decision rule be written in the following form:*

$$\hat{H}(y)\begin{cases} 0 & , \text{ if } y \leq \theta_1 \\ 2 & , \text{ if } \theta_1 < y < \theta_2 \\ 1 & , \text{ if } y \geq \theta_2, \end{cases}$$

where θ_1 *and* θ_2 *are some thresholds? If so, determine the values* θ_1 *and* θ_2 ?

(d) *Find the probability of a wrong forecast knowing that tomorrow's weather is unstable, i.e., determine the probability that the decision* $\hat{H}$ *is different from 2 knowing that, in reality,* $H = 2$. *This probability is denoted* $P_e(2)$.

(e) *If we assume that, instead of using the optimal rule, our weather frog always decides that tomorrow's weather is sunny, what will be his probability of error (probability of a wrong forecast)? Explain.*

Problem 7. (Hypothesis Testing in Laplacian Noise) *Consider the following hypothesis testing problem between two equally likely hypotheses. Under hypothesis* $H = 0$, *the observable* Y *is equal to* $a + Z$ *where* Z *is a random variable with Laplacian distribution*

$$f_Z(z) = \frac{1}{2}e^{-|z|}.$$

Under hypothesis $H = 1$, *the observable is given by* $-a + Z$.

(a) *Find and draw the density* $f_{Y|H}(y|0)$ *of the observable under hypothesis* $H = 0$, *and the density* $f_{Y|H}(y|1)$ *of the observable under hypothesis* $H = 1$.

(b) *Find the optimal decision rule to minimize the probability of error. Write out the expression for the likelihood ratio.*

(c) *Compute the probability of error of the optimal decision rule.*

Problem 8. (Poisson Parameter Estimation) *In this example there are two hypotheses,* $H = 0$ *and* $H = 1$ *which occur with probabilities* $P_H(0) = p_0$ *and* $P_H(1) = 1 - p_0$, *respectively. The observable is* $y \in \mathbb{N}_0$*, i.e.* y *is a nonnegative integer. Under hypothesis* $H = 0$*,* y *is distributed according to a Poisson law with parameter* λ_0*, i.e.*

$$p_{Y|H}(y|0) \quad = \quad \frac{\lambda_0^y}{y!} e^{-\lambda_0}. \tag{2.24}$$

Under hypothesis $H = 1$*,*

$$p_{Y|H}(y|1) \quad = \quad \frac{\lambda_1^y}{y!} e^{-\lambda_1}. \tag{2.25}$$

This example is in fact modeling the reception of photons in an optical fiber (for more details, see the Example in Section 2.2 of these notes).

(a) *Derive the MAP decision rule by indicating likelihood and log-likelihood ratios.* Hint: *The direction of an inequality changes if both sides are multiplied by a negative number.*

(b) *Derive the formula for the probability of error of the MAP decision rule.*

(c) *For* $p_0 = 1/3$*,* $\lambda_0 = 2$ *and* $\lambda_1 = 10$*, compute the probability of error of the MAP decision rule. You may want to use a computer program to do this.*

(d) *Repeat* (iv) *with* $\lambda_1 = 20$ *and comment.*

Problem 9. (MAP Decoding Rule: Alternative Derivation) *Consider the binary hypothesis testing problem where* H *takes values in* $\{0, 1\}$ *with probabilites* $P_H(0)$ *and* $P_H(1)$ *and the conditional probability density function of the observation* $Y \in \mathbb{R}$ *given* $H = i$*,* $i \in \{0, 1\}$ *is given by* $f_{Y|H}(\cdot|i)$*. Let* $\mathcal{R}_i$ *be the decoding region for hypothesis* i*, i.e the set of* y *for which the decision is* $\hat{H} = i$*,* $i \in \{0, 1\}$*.*

(a) *Show that the probability of error is given by*

$$P_e = P_H(1) + \int_{\mathcal{R}_1} \big(P_H(0) f_{Y|H}(y|0) - P_H(1) f_{Y|H}(y|1)\big)\, dy.$$

Hint: Note that $\mathbb{R} = \mathcal{R}_0 \bigcup \mathcal{R}_1$ *and* $\int_{\mathbb{R}} f_{Y|H}(y|i) dy = 1$ *for* $i \in \{0, 1\}$*.*

(b) *Argue that* P_e *is minimized when*

$$\mathcal{R}_1 = \{y \in \mathbb{R} : P_H(0) f_{Y|H}(y|0) < P_H(1) f_{Y|H}(y|1)\}$$

i.e the MAP rule!

PROBLEM 10. (One Bit over a Binary Channel with Memory) *Consider communicating one bit via n uses of a binary channel with memory. The channel output Y_i at time instant i is given by*

$$Y_i = X_i \oplus Z_i \qquad i = 1, \dots, n$$

where X_i is the binary channel input, Z_i is the binary noise and $\oplus$ represents modulo 2 addition. The noise sequence is generated as follows: Z_1 is generated from the distribution $Pr(Z_1 = 1) = p$ and for $i > 1$,

$$Z_i = Z_{i-1} \oplus N_i$$

where $N_2, \dots, N_n$ are i.i.d. with $Pr(N_i = 1) = p$. Let $(X_1^{(0)}, \dots, X_n^{(0)})$ and $(X_1^{(1)}, \dots, X_n^{(1)})$ denote the codewords (the sequence of symbols sent on the channel) corresponding to the message being 0 and 1 respectively.

(a) Consider the following operation by the receiver. The receiver creates the vector $(\hat{Y}_1, \hat{Y}_2, \dots, \hat{Y}_n)^T$ where $\hat{Y}_1 = Y_1$ and for $i = 2, 3, \dots, n$, $\hat{Y}_i = Y_i \oplus Y_{i-1}$. Argue that the vector created by the receiver is a sufficient statistic. Hint: Show that $(Y_1, Y_2, \dots, Y_n)^\top$ can be reconstructed from $(\hat{Y}_1, \hat{Y}_2, \dots, \hat{Y}_n)^\top$.

(b) Write down $(\hat{Y}_1, \hat{Y}_2, \dots, \hat{Y}_n)^\top$ for each of the hypotheses. Notice the similarity with the problem of communicating one bit via n uses of a binary symmetric channel.

(c) How should the receiver choose the codewords $(X_1^{(0)}, \dots, X_n^{(0)})$ and $(X_1^{(1)}, \dots, X_n^{(1)})$ so as to minimize the probability of error? Hint: When communicating one bit via n uses of a binary symmetric channel, the probability of error is minimized by choosing two codewords that differ in each component.

PROBLEM 11. (IID versus First-Order Markov) *Consider testing two equally likely hypotheses $H = 0$ and $H = 1$. The observable*

$$Y = (Y_1, \dots, Y_k) \tag{2.26}$$

is a k-dimensional binary vector. Under $H = 0$ the components of the vector Y are independent uniform random variables (also called Bernoulli$(1/2)$ random variables). Under $H = 1$, the component Y_1 is also uniform, but the components Y_i, $2 \le i \le k$, are distributed as follows:

$$Pr(Y_i = y_i | Y_{i-1} = y_{i-1}, \dots, Y_1 = y_1) = \begin{cases} 3/4, & \text{if } y_i = y_{i-1} \\ 1/4, & \text{otherwise.} \end{cases} \tag{2.27}$$

(a) Find the decision rule that minimizes the probability of error. Hint: *Write down a short sample sequence $(y_1, \dots, y_k)$ and determine its probability under each hypothesis. Then generalize.*

(b) Give a simple sufficient statistic for this decision.

(c) *Suppose that the observed sequence alternates between* 0 *and* 1 except *for one string of ones of length* s, *i.e. the observed sequence* y *looks something like*

$$y = 0101010111111\ldots111111010101\ldots. \tag{2.28}$$

What is the least s *such that we decide for hypothesis* $H = 1$*? Evaluate your formula for* $k = 20$.

Problem 12. (Real-Valued Gaussian Random Variables) *For the purpose of this problem, two zero-mean real-valued Gaussian random variables* X *and* Y *are called* jointly *Gaussian if and only if their joint density is*

$$f_{XY}(x,y) = \frac{1}{2\pi\sqrt{\det\Sigma}}\exp\left(-\frac{1}{2}\begin{pmatrix} x, & y\end{pmatrix}\Sigma^{-1}\begin{pmatrix} x \\ y\end{pmatrix}\right), \tag{2.29}$$

where (for zero-mean random vectors) the so-called covariance matrix Σ *is*

$$\Sigma = E\left[\begin{pmatrix} X \\ Y\end{pmatrix}(X,Y)\right] = \begin{pmatrix} \sigma_X^2 & \sigma_{XY} \\ \sigma_{XY} & \sigma_Y^2\end{pmatrix}. \tag{2.30}$$

(a) *Show that if* X *and* Y *are jointly Gaussian random variables, then* X *is a Gaussian random variable, and so is* Y.

(b) *How does your answer change if you use the definition of jointly Gaussian random variables given in these notes?*

(c) *Show that if* X *and* Y *are independent Gaussian random variables, then* X *and* Y *are jointly Gaussian random variables.*

(d) *However, if* X *and* Y *are Gaussian random variables but* not *independent, then* X *and* Y *are not necessarily jointly Gaussian. Give an example where* X *and* Y *are Gaussian random variables, yet they are* not *jointly Gaussian.*

(e) *Let* X *and* Y *be independent Gaussian random variables with zero mean and variance* σ_X^2 *and* σ_Y^2, *respectively. Find the probability density function of* $Z = X + Y$.

Problem 13. (Correlation versus Independence) *Let* Z *be a random variable with p.d.f.:*

$$f_Z(z) = \begin{cases} 1/2, & -1 \le z \le 1 \\ 0, & \textit{otherwise.}\end{cases}$$

Also, let $X = Z$ *and* $Y = Z^2$.

(a) *Show that* X *and* Y *are uncorrelated.*

(b) Are X and Y independent?

(c) Now let X and Y be jointly Gaussian, zero mean, uncorrelated with variances σ_X^2 and σ_Y^2 respectively. Are X and Y independent? Justify your answer.

PROBLEM 14. (Uniform Polar To Cartesian) *Let R and Φ be independent random variables. R is distributed uniformly over the unit interval, Φ is distributed uniformly over the interval $[0, 2\pi)$.*[5]

(a) Interpret R and Φ as the polar coordinates of a point in the plane. It is clear that the point lies inside (or on) the unit circle. Is the distribution of the point uniform over the unit disk? Take a guess!

(b) Define the random variables

$$\begin{aligned} X &= R\cos\Phi \\ Y &= R\sin\Phi. \end{aligned}$$

Find the joint distribution of the random variables X and Y using the Jacobian determinant.

Do you recognize a relationship between this method and the method derived in class to determine the probability density after a linear non-singular transformation?

(c) Does the result of part (ii) *support or contradict your guess from part* (i)*? Explain.*

PROBLEM 15. (Sufficient Statistic) *Consider a binary hypothesis testing problem specified by:*

$$H = 0 \quad : \quad \begin{cases} Y_1 = Z_1 \\ Y_2 = Z_1 Z_2 \end{cases}$$
$$H = 1 \quad : \quad \begin{cases} Y_1 = -Z_1 \\ Y_2 = -Z_1 Z_2 \end{cases}$$

where Z_1, Z_2 and H are independent random variables.

(a) Is Y_1 a sufficient statistic?

(Hint: *If $Y = aZ$, where a is a scalar then $f_Y(y) = \frac{1}{|a|} f_Z(\frac{y}{a})$).*

[5] This notation means: 0 is included, but 2π is excluded. It is the current standard notation in the anglo-saxon world. In the French world, the current standard for the same thing is $[0, 2\pi[$.

Problem 16. (More on Sufficient Statistic) *We have seen that if $H \rightarrow T(Y) \rightarrow Y$ then the P_e of a MAP decoder that observes both $T(Y)$ and Y is the same as that of a MAP decoder that observes only $T(Y)$. You may wonder if the contrary is also true, namely if the knowledge that Y does not help reducing the error probability that one can achieve with $T(Y)$ implies $H \rightarrow T(Y) \rightarrow Y$. Here is a counterexample. Let the hypothesis H be either 0 or 1 with equal probability (the choice of distribution on H is critical in this example). Let the observable Y take four values with the following conditional probabilities*

$$P_{Y|H}(y|0) = \begin{cases} 0.4 & \text{if } y = 0 \\ 0.3 & \text{if } y = 1 \\ 0.2 & \text{if } y = 2 \\ 0.1 & \text{if } y = 3 \end{cases} \qquad P_{Y|H}(y|1) = \begin{cases} 0.1 & \text{if } y = 0 \\ 0.2 & \text{if } y = 1 \\ 0.3 & \text{if } y = 2 \\ 0.4 & \text{if } y = 3 \end{cases}$$

and $T(Y)$ is the following function

$$T(y) = \begin{cases} 0 & \text{if } y = 0 \text{ and } y = 1 \\ 1 & \text{if } y = 2 \text{ and } y = 3. \end{cases}$$

(a) Show that the MAP decoder $\hat{H}(T(y))$ that makes its decisions based on $T(y)$ is equivalent to the MAP decoder $\hat{H}(y)$ that operates based on y.

(b) Compute the probabilities $Pr(Y = 0 \mid T(Y) = 0, H = 0)$ and $Pr(Y = 0 \mid T(Y) = 0, H = 1)$. Do we have $H \rightarrow T(Y) \rightarrow Y$?

Problem 17. (Fisher-Neyman Factorization Theorem) *Consider the hypothesis testing problem where the hypothesis is $H \in \mathcal{H} = \{0, 1, \ldots, m-1\}$, the observable is Y, and $T(Y)$ is a function of the observable. Let $f_{Y|H}(y|i)$ be given for all $i \in \mathcal{H}$. Suppose that there are functions $g_1, g_2, \ldots, g_{m-1}$ so that for each $i \in \mathcal{H}$ one can write*

$$f_{Y|H}(y|i) = g_i(T(y))h(y). \tag{2.31}$$

(a) Show that when the above conditions are satisfied, a MAP decision depends only on $T(Y)$. Hint: work directly with the definition of a MAP decision.

(b) Show that $T(Y)$ is a sufficient statistic, that is $H \rightarrow T(Y) \rightarrow Y$. Hint: Start by observing the following fact: Given a random variable Y with probability density function $f_Y(y)$ and given an arbitrary event $\mathcal{B}$, we have

$$f_{Y|Y\in\mathcal{B}} = \frac{f_Y(y)1_{\mathcal{B}}(y)}{\int_{\mathcal{B}} f_Y(y)dy}. \tag{2.32}$$

Proceed by defining $\mathcal{B}$ to be the event $\mathcal{B} = \{y : T(y) = t\}$ and make use of (2.32) applied to $f_{Y|H}(y|i)$ to prove that $f_{Y|H,T(Y)}(y|i,t)$ is independent of i.

For the following two examples, verify that condition (2.31) above is satisfied. You can then immediately conclude from part (a) and (b) that $T(Y)$ is a sufficient statistic.

(a) (Example 1) *Let $Y = (Y_1, Y_2, \dots, Y_n)$, $Y_k \in \{0,1\}$, be an independent and identically distributed (i.i.d) sequence of coin tosses of a coin such that $P_{Y_k|H}(1|i) = p_i$. Show that the function $T(y_1, y_2, \dots, y_n) = \sum_{k=1}^{n} y_k$ fulfills the condition expressed in (2.31). (Notice that $T(y_1, y_2, \dots, y_n)$ is the number of 1's in $y_1, y_2, \dots, y_n$.)*

(b) (Example 2) *Under hypothesis $H = i$, let the observable Y_k be Gaussian distributed with mean m_i and variance 1; that is*

$$f_{Y_k|H}(y|i) = \frac{1}{\sqrt{2\pi}} e^{-(y-m_i)^2},$$

and $Y_1, Y_2, \dots, Y_n$ be independently drawn according to this distribution. Show that the sample mean $T(y_1, y_2, \dots, y_n) = \frac{1}{n}\sum_{k=1}^{n} y_k$ fulfills the condition expressed in (2.31).

Problem 18. (Irrelevance and Operational Irrelevance) *Let the hypothesis H be related to the observables (U, V) via the channel $P_{U,V|H}$. We say that V is* operationally irrelevant *if a MAP decoder that observes (U, V) achieves the same probability of error as one that observes only U, and this is true regardless of P_H. We now prove that irrelevance and operational irrelevance imply one another. We have already proved that irrelevance implies operational irrelevance. Hence it suffices to show that operational irrelevance implies irrelevance or, equivalently, that if V is not irrelevant then it is not operationally irrelevant. We will prove the latter statement. We start by a few observations that are instructive and also useful to get us started. By definition, V irrelevant means $H \to U \to V$. Hence V irrelevant is equivalent to the statement that that, conditioned on U, the random variables H and V are independent. This gives us one intuitive explanation why V is operationally irrelevant: Once we have observed $U = u$, we may restate the hypothesis testing problem in terms of an hypothesis H and an observable V that are independent (conditioned on $U = u$) and because of independence, from V we don't learn anything about H. On the other hand if V is* not *irrelevant then there is at least a u, call it u^*, for which H and V are* not *independent conditioned on $U = u^*$. It is when such a u is observed that we should be able to prove that V affects the decision. This suggests that the problem we are trying to solve is intimately related to the simpler problem that involves only the hypothesis H and the observable V and the two are* not *independent. We start with this problem and then we generalize.*

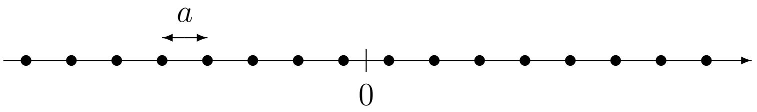
(a) *Let the hypothesis be $H \in \mathcal{H}$ (of yet unspecified distribution) and let the observable $V \in \mathcal{V}$ be related to H via an arbitrary but fixed channel $P_{V|H}$. Show that if V is* not *independent of H then there are distinct elements $i, j \in \mathcal{H}$ and distinct elements $k, l \in \mathcal{V}$ such that*

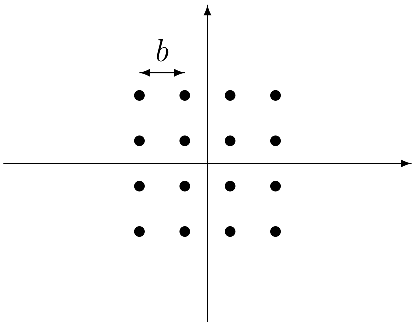
$$\begin{aligned} P_{V|H}(k|i) &< P_{V|H}(l|i) \\ P_{V|H}(k|j) &> P_{V|H}(l|j). \end{aligned}$$

(b) *Under the conditions of the previous question, show that there is a distribution P_H for which the observable V affects the decision of a MAP decoder.*

(c) *Generalize to show that if the observables are U and V and $P_{U,V|H}$ is fixed so that $H \rightarrow U \rightarrow V$ does* not *hold, then there is a distribution on H for which V is* not *operationally irrelevant.*

Problem 19. (16-PAM versus 16-QAM) *The following two signal constellations are used to communicate across an additive white Gaussian noise channel. Let the noise variance be σ^2.*





Each point represents a signal $\boldsymbol{s}_i$ for some i. Assume each signal is used with the same probabiliy.

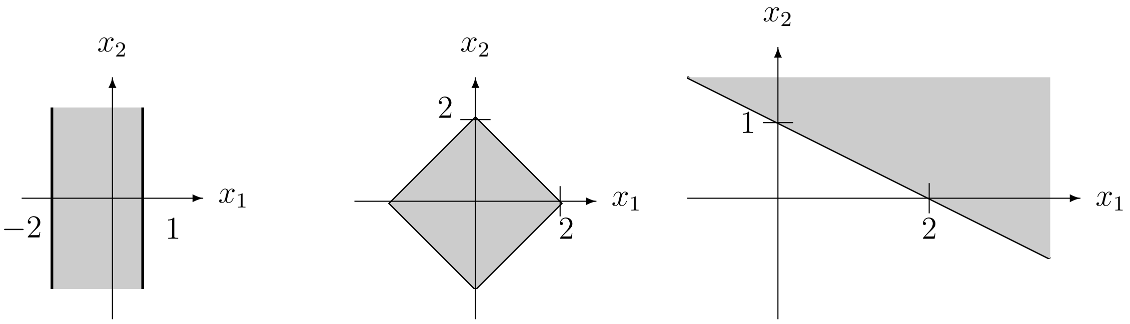
(a) *For each signal constellation, compute the average probability of error, P_e, as a function of the parameters a and b, respectively.*

(b) *For each signal constellation, compute the average energy per symbol, E_s, as a function of the parameters a and b, respectively:*

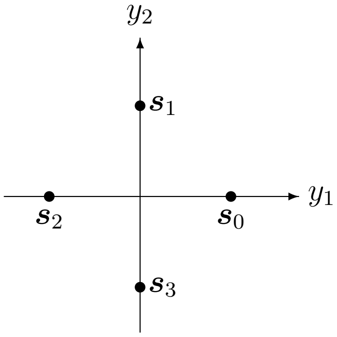
$$E_s = \sum_{i=1}^{16} P_H(i) \|\boldsymbol{s}_i\|^2 \tag{2.33}$$

(c) *Plot P_e versus E_s for both signal constellations and comment.*

Problem 20. (Q-Functions on Regions) [Wozencraft and Jacobs] *Let $\boldsymbol{X} \sim \mathcal{N}(0, \sigma^2 I_2)$. For each of the three figures below, express the probability that $\boldsymbol{X}$ lies in the shaded region. You may use the Q-function when appropriate.*



Problem 21. (QPSK Decision Regions) *Let $H \in \{0, 1, 2, 3\}$ and assume that when $H = i$ you transmit the signal $\boldsymbol{s}_i$ shown in the figure. Under $H = i$, the receiver observes $\boldsymbol{Y} = \boldsymbol{s}_i + \boldsymbol{Z}$.*



(a) Draw the decoding regions assuming that $\boldsymbol{Z} \sim \mathcal{N}(0, \sigma^2 I_2)$ and that $P_H(i) = 1/4$, $i \in \{0, 1, 2, 3\}$.

(b) Draw the decoding regions (qualitatively) assuming $\boldsymbol{Z} \sim \mathcal{N}(0, \sigma^2 I)$ and $P_H(0) = P_H(2) > P_H(1) = P_H(3)$. Justify your answer.

(c) Assume again that $P_H(i) = 1/4$, $i \in \{0, 1, 2, 3\}$ and that $\boldsymbol{Z} \sim \mathcal{N}(0, K)$, where $K = \begin{pmatrix} \sigma^2 & 0 \\ 0 & 4\sigma^2 \end{pmatrix}$. How do you decode now? Justify your answer.

Problem 22. (Antenna Array) *The following problem relates to the design of multi-antenna systems. The situation that we have in mind is one where one of two signals is transmitted over a Gaussian channel and is received through two different antennas. We shall assume that the noises at the two terminals are independent but not necessarily of equal variance. You are asked to design a receiver for this situation, and to assess its performance. This situation is made more precise as follows:*

Consider the binary equiprobable hypothesis testing problem:

$$\begin{aligned} H=0 &: Y_1 = A+Z_1, \quad Y_2 = A+Z_2 \\ H=1 &: Y_1 = -A+Z_1, \quad Y_2 = -A+Z_2, \end{aligned}$$

where Z_1, Z_2 are independent Gaussian random variables with different *variances $\sigma_1^2 \neq \sigma_2^2$, that is, $Z_1 \sim \mathcal{N}(0, \sigma_1^2)$ and $Z_2 \sim \mathcal{N}(0, \sigma_2^2)$. $A > 0$ is a constant.*

(a) *Show that the decision rule that minimizes the probability of error (based on the observable Y_1 and Y_2) can be stated as*

$$\sigma_2^2 y_1 + \sigma_1^2 y_2 \overset{0}{\underset{1}{\gtrless}} 0.$$

(b) *Draw the decision regions in the (Y_1, Y_2) plane for the special case where $\sigma_1 = 2\sigma_2$.*

(c) *Evaluate the probability of error for the optimal detector as a function of σ_1^2, σ_2^2 and A.*

Problem 23. (Multiple Choice Exam) *You are taking a multiple choice exam. Question number 5 allows for two possible answers. According to your first impression, answer 1 is correct with probability $1/4$ and answer 2 is correct with probability $3/4$.*

You would like to maximize your chance of giving the correct answer and you decide to have a look at what your left and right neighbors have to say.

The left neighbor has answered $\hat{H}_L = 1$. He is an excellent student who has a record of being correct 90% of the time.

The right neighbor has answered $\hat{H}_R = 2$. He is a weaker student who is correct 70% of the time.

(a) *You decide to use your first impression as a prior and to consider $\hat{H}_L$ and $\hat{H}_R$ as observations. Describe the corresponding hypothesis testing problem.*

(b) *What is your answer $\hat{H}$? Justify it.*

Problem 24. (QAM with Erasure) *Consider a QAM receiver that outputs a special symbol called "erasure" and denoted by δ whenever the observation falls in the shaded area shown in Figure 2.14. Assume that $\boldsymbol{s}_0$ is transmitted and that $\boldsymbol{Y} = \boldsymbol{s}_0 + \boldsymbol{N}$ is received where $\boldsymbol{N} \sim \mathcal{N}(0, \sigma^2 I_2)$. Let P_{0i}, $i = 0,1,2,3$ be the probability that the receiver outputs $\hat{H} = i$ and let$P_{0\delta}$ be the probability that it outputs δ. Determine P_{00}, P_{01}, P_{02}, P_{03} and $P_{0\delta}$.*

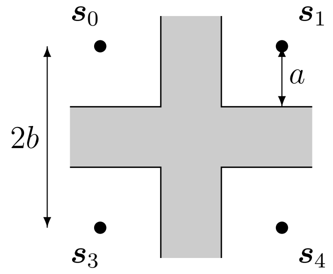


Figure 2.14: Decoding regions for QAM with erasure.

PROBLEM 25. (Repeat Codes and Bhattacharyya Bound) *Consider two equally likely hypotheses. Under hypothesis* $H = 0$, *the transmitter sends* $\boldsymbol{s}_0 = (1, \dots, 1)$ *and under* $H = 1$ *it sends* $\boldsymbol{s}_0 = (-1, \dots, -1)$. *The channel model is the AWGN with variance* σ^2 *in each component. Recall that the probability of error for a ML receiver that observes the channel output* $\boldsymbol{Y}$ *is*

$$P_{e,1} \;=\; Q\left(\frac{\sqrt{N}}{\sigma}\right).$$

Suppose now that the decoder has access only *to the sign of* Y_i, $1 \le i \le N$. *That is, the observation is*

$$W = (W_1, \dots, W_N) = (\mathrm{sign}(Y_1), \dots, \mathrm{sign}(Y_N)). \tag{2.34}$$

(a) *Determine the MAP decision rule based on the observation* $(W_1, \dots, W_N)$. *Give a simple sufficient statistic, and draw a diagram of the optimal receiver.*

(b) *Find the expression for the probability of error* $P_{e,2}$. *You may assume that* N *is odd.*

(c) *Your answer to* (ii) *contains a sum that cannot be expressed in closed form. Express the Bhattacharyya bound on* $P_{e,2}$.

(d) *For* $N = 1, 3, 5, 7$, *find the numerical values of* $P_{e,1}$, $P_{e,2}$, *and the Bhattacharyya bound on* $P_{e,2}$.

PROBLEM 26. (Tighter Union Bhattacharyya Bound: Binary Case) *In this problem we derive a tighter version of the* Union Bhattacharyya Bound *for binary hypotheses. Let*

$$\begin{aligned} H = 0 \;&:\; Y \sim f_{Y|H}(y|0) \\ H = 1 \;&:\; Y \sim f_{Y|H}(y|1). \end{aligned}$$

The MAP decision rule is

$$\hat{H}(y) = \arg\max_i P_H(i) f_{Y|H}(y|i),$$

and the resulting probability of error is

$$Pr\{e\} = P_H(0) \int_{\mathcal{R}_1} f_{Y|H}(y|0)dy + P_H(1) \int_{\mathcal{R}_0} f_{Y|H}(y|1)dy.$$

(a) *Argue that*

$$Pr\{e\} = \int_y \min\left\{P_H(0)f_{Y|H}(y|0), P_H(1)f_{Y|H}(y|1)\right\} dy.$$

(b) *Prove that for* $a, b \geq 0,\ \min(a,b) \leq \sqrt{ab} \leq \frac{a+b}{2}$. *Use this to prove the tighter version of* Bhattacharyya Bound, *i.e,*

$$Pr\{e\} \leq \frac{1}{2} \int_y \sqrt{f_{Y|H}(y|0)f_{Y|H}(y|1)}dy.$$

(c) *Compare the above bound to the one derived in class when* $P_H(0) = \frac{1}{2}$. *How do you explain the improvement by a factor* $\frac{1}{2}$*?*

Problem 27. (Tighter Union Bhattacharyya Bound: M-ary Case) *In this problem we derive a tight version of the union bound for* M*-ary hypotheses. Let us analyze the following M-ary MAP detector:*

$$\begin{aligned}\hat{H}(y) &= \text{smallest } i \text{ such that} \\ &\quad P_H(i)f_{Y/H}(y/i) = \max_j\{P_H(j)f_{Y/H}(y/j)\}\end{aligned}$$

Let

$$\mathcal{B}_{ij} = \begin{cases} y : P_H(j)f_{Y|H}(y|j) \geq P_H(i)f_{Y|H}(y|i), & j < i \\ y : P_H(j)f_{Y|H}(y|j) > P_H(i)f_{Y|H}(y|i), & j > i \end{cases}$$

(a) *Verify that* $\mathcal{B}_{ij} = \mathcal{B}_{ji}^c$.

(b) *Given* $H = i$, *the detector will make an error iff:* $y \in \bigcup_{j:j\neq i} \mathcal{B}_{ij}$ *and the probability of error is* $P_e = \sum_{i=0}^{M-1} P_e(i)P_H(i)$. *Show that:*

$$\begin{aligned} P_e &\leq \sum_{i=0}^{M-1}\sum_{j>i} [Pr\{\mathcal{B}_{ij}|H=i\}P_H(i) + Pr\{\mathcal{B}_{ji}|H=j\}P_H(j)] \\ &= \sum_{i=0}^{M-1}\sum_{j>i} \left[\int_{\mathcal{B}_{ij}} f_{Y|H}(y|i)P_H(i)dy + \int_{\mathcal{B}_{ij}^c} f_{Y|H}(y|j)P_H(j)dy\right] \\ &= \sum_{i=0}^{M-1}\sum_{j>i} \left[\int_y \min\left\{f_{Y|H}(y|i)P_H(i), f_{Y|H}(y|j)P_H(j)\right\} dy\right] \end{aligned}$$

Hint: *Use the union bound and then group the terms corresponding to* $\mathcal{B}_{ij}$ *and* $\mathcal{B}_{ji}$. *To prove the last part, go back to the definition of* $\mathcal{B}_{ij}$.

(c) Hence show that:

$$P_e \leq \sum_{i=0}^{M-1} \sum_{j>i} \left[\left(\frac{P_H(i) + P_H(j)}{2} \right) \int_y \sqrt{f_{Y|H}(y|i) f_{Y|H}(y|j) dy} \right]$$

*(*Hint: *For $a, b \geq 0, \min(a,b) \leq \sqrt{ab} \leq \frac{a+b}{2}$.)*

As an application of the above bound, consider the following binary hypothesis testing problem:

$$\begin{aligned} H = 0 &: Y \sim \mathcal{N}(-a, \sigma^2) \\ H = 1 &: Y \sim \mathcal{N}(+a, \sigma^2) \end{aligned}$$

where the two hypotheses are equiprobable. Use the above bound to show that:

$$\begin{aligned} P_e &= P_e(0) \\ &\leq \frac{1}{2} \exp\left\{ -\frac{a^2}{2\sigma^2} \right\} \end{aligned}$$

But $P_e = Q\left(\frac{a}{\sigma}\right)$. Hence we have re-derived the bound (see lecture 1):

$$Q(x) \leq \frac{1}{2} \exp\left\{ -\frac{x^2}{2} \right\}.$$

Problem 28. (Applying the Tight Bhattacharyya Bound) *As an application of the tight Bhattacharyya bound, consider the following binary hypothesis testing problem*

$$\begin{aligned} H = 0 &: Y \sim \mathcal{N}(-a, \sigma^2) \\ H = 1 &: Y \sim \mathcal{N}(+a, \sigma^2) \end{aligned}$$

where the two hypotheses are equiprobable.

(a) Use the Tight Bhattacharyya Bound *to derive a bound on P_e.*

(b) We know that the probability of error for this binary hypothesis testing problem is $Q(\frac{a}{\sigma}) \leq \frac{1}{2} \exp\left\{ -\frac{a^2}{2\sigma^2} \right\}$, where we have used the result $Q(x) \leq \frac{1}{2} \exp\left\{ -\frac{x^2}{2} \right\}$ derived in lecture 1. How do the two bounds compare? Are you surprised (and why)?

Problem 29. (Bhattacharyya Bound for DMCs) *Consider a Discrete Memoryless Channel (DMC). This is a channel model described by an input alphabet* $\mathcal{X}$, *an output alphabet* $\mathcal{Y}$ *and a transition probability*[6] $P_{Y|X}(y|x)$. *When we use this channel to transmit an n-tuple* $\boldsymbol{x} \in \mathcal{X}^n$, *the transition probability is*

$$P_{\boldsymbol{Y}|\boldsymbol{X}}(\boldsymbol{y}|\boldsymbol{x}) = \prod_{i=1}^{n} P_{Y|X}(y_i|x_i).$$

So far we have come across two DMCs, namely the BSC (Binary Symmetric Channel) and the BEC (Binary Erasure Channel). The purpose of this problem is to realize that for DMCs, the Bhattacharyya Bound *takes on a simple form, in particular when the channel input alphabet* $\mathcal{X}$ *contains only two letters.*

(a) Consider a source that sends $\boldsymbol{s}_0$ *when* $H = 0$ *and* $\boldsymbol{s}_1$ *when* $H = 1$. *Justify the following chain of inequalities.*

$$\begin{aligned}
P_e &\overset{(a)}{\leq} \sum_{\boldsymbol{y}} \sqrt{P_{\boldsymbol{Y}|\boldsymbol{X}}(\boldsymbol{y}|\boldsymbol{s}_0) P_{\boldsymbol{Y}|\boldsymbol{X}}(\boldsymbol{y}|\boldsymbol{s}_1)} \\
&\overset{(b)}{\leq} \sum_{\boldsymbol{y}} \sqrt{\prod_{i=1}^{n} P_{Y|X}(y_i|s_{0i}) P_{Y|X}(y_i|s_{1i})} \\
&\overset{(c)}{=} \sum_{y_1,\ldots,y_n} \prod_{i=1}^{n} \sqrt{P_{Y|X}(y_i|s_{0i}) P_{Y|X}(y_i|s_{1i})} \\
&\overset{(d)}{=} \sum_{y_1} \sqrt{P_{Y|X}(y_1|s_{01}) P_{Y|X}(y_1|s_{11})} \ldots \sum_{y_n} \sqrt{P_{Y|X}(y_n|s_{0n}) P_{Y|X}(y_n|s_{1n})} \\
&\overset{(e)}{=} \prod_{i=1}^{n} \sum_{y} \sqrt{P_{Y|X}(y|s_{0i}) P_{Y|X}(y|s_{1i})} \\
&\overset{(f)}{=} \prod_{a \in \mathcal{X}, b \in \mathcal{X}, a \neq b} \left(\sum_{y} \sqrt{P_{Y|X}(y|s_{0i}) P_{Y|X}(y|s_{1i})} \right)^{n(a,b)}.
\end{aligned}$$

where $n(a,b)$ *is the number of positions* i *in which* $s_{0i} = a$ *and* $s_{1i} = b$.

(b) The Hamming distance $d_H(\boldsymbol{s}_0, \boldsymbol{s}_1)$ *is defined as the number of positions in which* $\boldsymbol{s}_0$ *and* $\boldsymbol{s}_1$ *differ. Show that for a binary input channel, i.e, when* $\mathcal{X} = \{a, b\}$, *the* Bhattacharyya Bound *becomes*

$$P_e \leq z^{d_H(\boldsymbol{s}_0, \boldsymbol{s}_1)},$$

where

$$z = \sum_{y} \sqrt{P_{Y|X}(y|a) P_{Y|X}(y|b)}.$$

[6] Here we are assuming that the output alphabet is discrete. Otherwise we need to deal with densities instead of probabilities.

Notice that z depends only on the channel whereas its exponent depends only on $\boldsymbol{s}_0$ and $\boldsymbol{s}_1$.

(c) *What is z for:*

(a) *The binary input Gaussian channel described by the densities*

$$\begin{aligned} f_{Y|X}(y|0) &= \mathcal{N}(-\sqrt{E}, \sigma^2) \\ f_{Y|X}(y|1) &= \mathcal{N}(\sqrt{E}, \sigma^2). \end{aligned}$$

(b) *The Binary Symmetric Channel (BSC) with the transition probabilities described by*

$$P_{Y|X}(y|x) = \begin{cases} 1-\delta, & \textit{if } y = x, \\ \delta, & \textit{otherwise.} \end{cases}$$

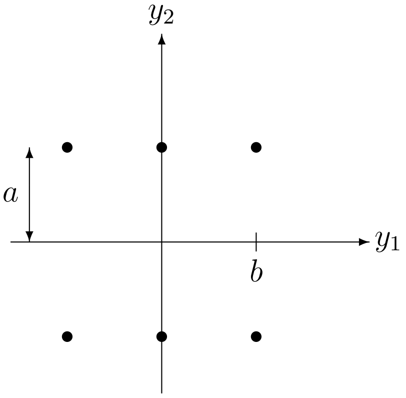
(c) *The Binary Erasure Channel (BEC) with the transition probabilities given by*

$$P_{Y|X}(y|x) = \begin{cases} 1-\delta, & \textit{if } y = x, \\ \delta, & \textit{if } y = E \\ 0, & \textit{otherwise.} \end{cases}$$

Compare your result with the the bound obtained in Example 16.

(d) *Consider a channel with input alphabet $\{\pm 1\}$, and output $Y = \text{sign}(x+Z)$, where x is the input and $Z \sim \mathcal{N}(0, \sigma^2)$. This is a BSC obtained from quantizing a Gaussian channel used with binary input alphabet. What is the crossover probability p of the BSC? Plot the z of the underlying Gaussian channel (with inputs in $\mathbb{R}$) and that of the BSC. By how much do we need to increase the input power of the quantized channel to match the z of the unquantized channel?*

Problem 30. (Signal Constellation) *The following signal constellation with six signals is used in additive white Gaussian noise of variance σ^2:*



Assume that the six signals are used with equal probability.

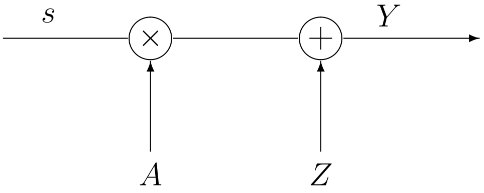
(a) *Draw the boundaries of the decision regions.*

(b) *Compute the average probability of error, P_e, for this signal constellation.*

(c) *Compute the average energy per symbol for this signal constellation.*

Problem 31. (Hypothesis Testing and Fading) *Consider the following communication problem: There are two equiprobable hypotheses. When $H = 0$, we transmit $s = -b$, where b is an arbitrary but fixed positive number. When $H = 1$, we transmit $s = b$.*

The channel is as shown in the figure below, where $Z \sim \mathcal{N}(0, \sigma^2)$ represents the noise, $A \in \{0, 1\}$ represents a random attenuation (fading) with $P_A(0) = \frac{1}{2}$, and Y is the channel output. The random variables H, A and Z are independent.



(a) *Find the decision rule that the receiver should implement to minimize the probability of error. Sketch the decision regions.*

(b) *Calculate the probability of error P_e, based on the above decision rule.*

Problem 32. (Dice Tossing) *You have two dices, one fair and one loaded. A friend told you that the loaded dice produces a 6 with probability $\frac{1}{4}$, and the other values with uniform probabilities. You do not know a priori which one is fair or which one is loaded. You pick with uniform probabilities one of the two dices, and perform N consecutive tosses. Let $\boldsymbol{Y} = (Y_1, \cdots, Y_N)$ be the sequence of numbers observed.*

(a) *Based on the sequence of observations $\boldsymbol{Y}$, find the decision rule to determine whether the dice you have chosen is loaded. Your decision rule should maximize the probability of correct decision.*

(b) *Identify a compact sufficient statistic for this hypothesis testing problem, call it S. Justify your answer. [Hint: $S \in \mathbb{N}$.]*

(c) *Find the Bhattacharyya bound on the probability of error. You can either work with the observation* $(Y_1, \dots, Y_N)$ *or with* $(Z_1, \dots, Z_N)$, *where* Z_i *indicates whether the* i*th observation is a six or not, or you can work with* S. *In some cases you may find it useful to know that* $\sum_{i=0}^{N} \binom{N}{i} x^i = (1+x)^N$ *for* $N \in \mathbb{N}$. *In other cases the following may be useful:* $\sum_{Y_1, Y_2, \dots, Y_N} \prod_{i=1}^{N} f(Y_i) = \left(\sum_{Y_1} f(Y_1) \right)^N$.

Problem 33. (Playing Darts) *Assume that you are throwing darts at a target. We assume that the target is one-dimensional,* i.e., *that the darts all end up on a line. The "bulls eye" is in the center of the line, and we give it the coordinate* 0. *The position of a dart on the target can then be measured with respect to* 0.

We assume that the position X_1 *of a dart that lands on the target is a random variable that has a Gaussian distribution with variance* σ_1^2 *and mean* 0.

Assume now that there is a second target, which is further away. If you throw dart to that target, the position X_2 *has a Gaussian distribution with variance* σ_2^2 *(where* $\sigma_2^2 > \sigma_1^2$*) and mean* 0.

You play the following game: You toss a coin which gives you "head" with probability p *and "tail" with probability* $1-p$ *for some fixed* $p \in [0,1]$. *If* $Z = 1$, *you throw a dart onto the first target. If* $Z = 0$, *you aim the second target instead. Let* X *be the relative position of the dart with respect to the center of the target that you have chosen.*

(a) *Write down* X *in terms of* X_1, X_2 *and* Z.

(b) *Compute the variance of* X. *Is* X *Gaussian?*

(c) *Let* $S = |X|$ *be the score, which is given by the distance of the dart to the center of the target (that you picked using the coin). Compute the average score* $\mathbb{E}[S]$.

Problem 34. (Properties of the Q Function) *Prove properties* (*a*) *through* (*d*) *of the Q function defined in Section 2.3. Hint: for property (d), multiple and divide inside the integral by the integration variable and integrate by parts. By upper and lowerbounding the resulting integral you will obtain the lower and upper bound.*

Problem 35. (Bhattacharyya Bound and Laplacian Noise) *When* $Y \in \mathbb{R}$ *is a continuous random variable, the Bhattacharyya bound states that*

$$Pr\{Y \in \mathcal{B}_{i,j} | H = i\} \leq \sqrt{\frac{P_H(j)}{P_H(i)}} \int_{y \in \mathbb{R}} \sqrt{f_{Y|H}(y|i) f_{Y|H}(y|j)}\, dy,$$

where i, j *are two possible hypotheses and* $\mathcal{B}_{i,j} = \{y \in \mathbb{R} : P_H(i) f_{Y|H}(y|i) \leq P_H(j) f_{Y|H}(y|j)\}$. *In this problem* $\mathcal{H} = \{0,1\}$ *and* $P_H(0) = P_H(1) = 0.5$.

(a) *Write a sentence that expresses the meaning of* $Pr\{Y \in \mathcal{B}_{0,1}|H = 0\}$. *Use words that have operational meaning.*

(b) *Do the same but for* $Pr\{Y \in \mathcal{B}_{0,1}|H = 1\}$. *(Note that we have written* $\mathcal{B}_{0,1}$ *and not* $\mathcal{B}_{1,0}$ *.)*

(c) *Evaluate the right hand side of the Bhattacharyya bound for the special case* $f_{Y|H}(y|0) = f_{Y|H}(y|1)$.

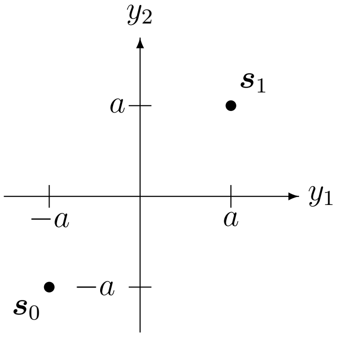
(d) *Evaluate the Bhattacharyya bound for the following (Laplacian noise) setting:*

$$\begin{aligned} H = 0 : &\quad Y = -a + Z \\ H = 1 : &\quad Y = a + Z, \end{aligned}$$

where $a \in \mathbb{R}_+$ *is a constant and* $f_Z(z) = \frac{1}{2}\exp(-|z|)$, $z \in \mathbb{R}$. *Hint: it does not matter if you evaluate the bound for* $H = 0$ *or* $H = 1$.

(e) *For which value of* a *should the bound give the result obtained in (c)? Verify that it does. Check your previous calculations if it does not.*

Problem 36. (Antipodal Signaling) *Consider the following signal constellation:*



Assume that $\boldsymbol{s}_1$ *and* $\boldsymbol{s}_0$ *are used for communication over the Gaussian vector channel. More precisely:*

$$\begin{aligned} H = 0 : &\quad \boldsymbol{Y} = \boldsymbol{s}_0 + \boldsymbol{Z}, \\ H = 1 : &\quad \boldsymbol{Y} = \boldsymbol{s}_1 + \boldsymbol{Z}, \end{aligned}$$

where $\boldsymbol{Z} \sim \mathcal{N}(\boldsymbol{0}, \sigma^2 I_2)$. *Hence,* $\boldsymbol{Y}$ *is a vector with two components* $\boldsymbol{Y} = (Y_1, Y_2)$.

(a) *Argue that* Y_1 is not *a sufficient statistic.*

(b) *Give a different signal constellation with two signals* $\tilde{\mathbf{s}}_0$ *and* $\tilde{\mathbf{s}}_1$ *such that, when used in the above communication setting,* Y_1 *is a sufficient statistic.*

Problem 37. (Hypothesis Testing: Uniform and Uniform) *Consider a binary hypothesis testing problem in which the hypotheses* $H = 0$ *and* $H = 1$ *occur with probability* $P_H(0)$ *and* $P_H(1) = 1 - P_H(0)$, *respectively. The observation* $\boldsymbol{Y}$ *is a sequence of zeros and ones of length* $2k$, *where* k *is a fixed integer. When* $H = 0$, *each component of* $\boldsymbol{Y}$ *is* 0 *or a* 1 *with probability* $\frac{1}{2}$ *and components are independent. When* $H = 1$, $\boldsymbol{Y}$ *is chosen uniformly at random from the set of all sequences of length* $2k$ *that have an equal number of ones and zeros. There are* $\binom{2k}{k}$ *such sequences.*

(a) *What is* $P_{\boldsymbol{Y}|H}(\boldsymbol{y}|0)$? *What is* $P_{\boldsymbol{Y}|H}(\boldsymbol{y}|1)$?

(b) *Find a maximum likelihood decision rule. What is the single number you need to know about* $\boldsymbol{y}$ *to implement this decision rule?*

(c) *Find a decision rule that minimizes the error probability.*

(d) *Are there values of* $P_H(0)$ *and* $P_H(1)$ *such that the decision rule that minimizes the error probability always decides for only one of the alternatives? If yes, what are these values, and what is the decision?*

Problem 38. (SIMO Channel with Laplacian Noise) *One of the two signals* $s_0 = -1, s_1 = 1$ *is transmitted over the channel shown on the left of Figure 2.15. The two noise random variables* Z_1 *and* Z_2 *are statistically independent of the transmitted signal and of each other. Their density functions are*

$$f_{Z_1}(\alpha) = f_{Z_2}(\alpha) = \frac{1}{2}\, e^{-|\alpha|}.$$

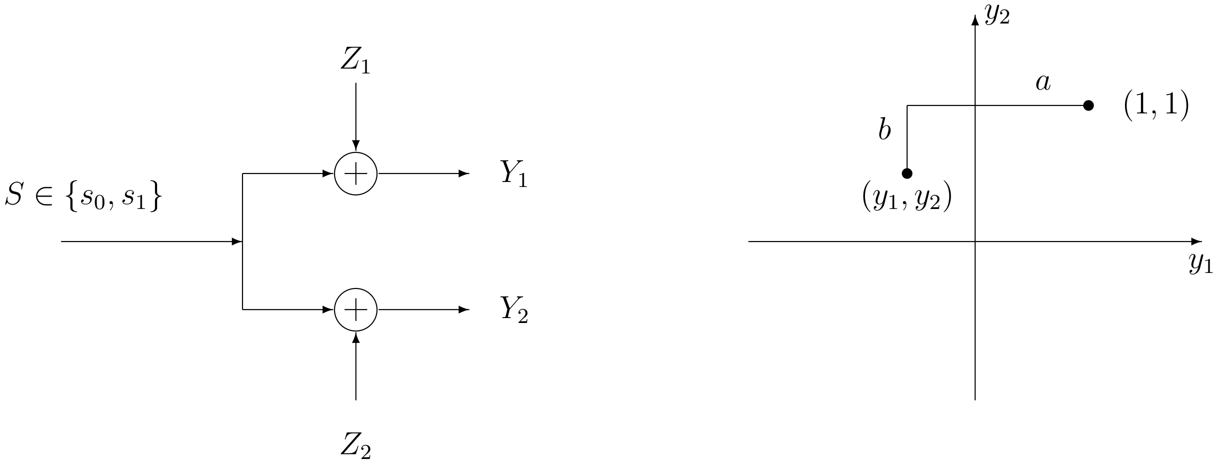


Figure 2.15: The channel (on the left) and a figure explaining the hint.

(a) *Derive a maximum likelihood decision rule.*

(b) *Describe the maximum likelihood decision regions in the (y_1, y_2) plane. Try to describe the "Either Choice" regions, i.e., the regions in which it does not matter if you decide for s_0 or for s_1. Hint: Use geometric reasoning and the fact that for a point (y_1, y_2) as shown on the right of the figure, $|y_1 - 1| + |y_2 - 1| = a + b$.*

(c) *A receiver decides that s_1 was transmitted if and only if $(y_1 + y_2) > 0$. Does this receiver minimize the error probability for equally likely messages?*

(d) *What is the error probability for the receiver in (c)? Hint: One way to do this is to use the fact that if $W = Z_1 + Z_2$ then $f_W(\omega) = \frac{e^{-\omega}}{4}(1 + \omega)$ for $w > 0$.*

(e) *Could you have derived f_W as in (d) ? If yes, say how but omit detailed calculations.*

Problem 39. (ML Receiver and UB for Orthogonal Signaling) *Let $H \in \{1, \ldots, m\}$ be uniformly distributed and consider the communication problem described by:*

$$H = i : \qquad \boldsymbol{Y} = \boldsymbol{s}_i + \boldsymbol{Z}, \quad \boldsymbol{Z} \sim \mathcal{N}(0, \sigma^2 I_m),$$

where $\boldsymbol{s}_1, \ldots, \boldsymbol{s}_m$, $\boldsymbol{s}_i \in \mathbb{R}^m$, is a set of constant-energy orthogonal signals. Without loss of generality we assume

$$\boldsymbol{s}_i = \sqrt{\mathcal{E}}\boldsymbol{e}_i,$$

where $\boldsymbol{e}_i$ is the ith unit vector in $\mathbb{R}^m$, i.e., the vector that contains 1 at position i and 0 elsewhere, and $\mathcal{E}$ is some positive constant.

(a) *Describe the maximum likelihood decision rule. (Make use of the fact that $\boldsymbol{s}_i = \sqrt{\mathcal{E}}\boldsymbol{e}_i$.)*

(b) *Find the distance $\|\boldsymbol{s}_i - \boldsymbol{s}_j\|$.*

(c) *Upper-bound the error probability $P_e(i)$ using the union bound and the Q function.*

Problem 40. (Data Storage Channel) *The process of storing and retrieving binary data on a thin-film disk may be modeled as transmitting binary symbols across an additive white Gaussian noise channel where the noise Z has a variance that depends on the transmitted (stored) binary symbol S. The noise has the following input-dependent density:*

$$f_Z(z) = \begin{cases} \frac{1}{\sqrt{2\pi\sigma_1^2}} e^{-\frac{z^2}{2\sigma_1^2}} \text{ if } S = 1 \\ \frac{1}{\sqrt{2\pi\sigma_0^2}} e^{-\frac{z^2}{2\sigma_0^2}} \text{ if } S = 0, \end{cases}$$

where $\sigma_1 > \sigma_0$. The channel inputs are equally likely.

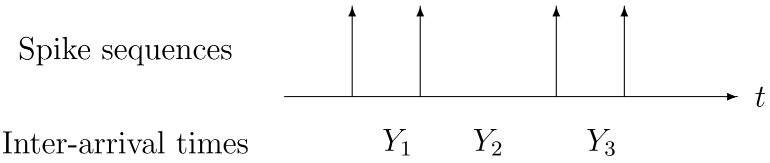
(a) *On the same graph, plot the two possible output probability density functions. Indicate, qualitatively, the decision regions.*

(b) *Determine the optimal receiver in terms of* σ_1 *and* σ_0.

(c) *Write an expression for the error probability* P_e *as a function of* σ_0 *and* σ_1.

PROBLEM 41. (Lie Detector) *You are asked to develop a "lie detector" and analyze its performance. Based on the observation of brain cell activity, your detector has to decide if a person is telling the truth or is lying.*

For the purpose of this problem, the brain cell produces a sequence of spikes as shown in the figure. For your decision you may use only a sequence of n *consecutive inter-arrival times* $Y_1, Y_2, \ldots, Y_n$. *Hence* Y_1 *is the time elapsed between the first and second spike,* Y_2 *the time between the second and third, etc.*



We assume that, a priori, a person lies with some known probability p. *When the person is telling the truth,* $Y_1, \ldots, Y_n$ *is an i.i.d. sequence of exponentially distributed random variables with intensity* α, $(\alpha > 0)$, *i.e.*

$$f_{Y_i}(y) = \alpha e^{-\alpha y}, \quad y \geq 0.$$

When the person lies, $Y_1, \ldots, Y_n$ *is i.i.d. exponentially distributed with intensity* β, $(\alpha < \beta)$.

(a) *Describe the decision rule of your lie detector for the special case* $n = 1$. *Your detector shall be designed so as to minimize the probability of error.*

(b) *What is the probability* $P_{L/T}$ *that your lie detector says that the person is lying when the person is telling the truth?*

(c) *What is the probability* $P_{T/L}$ *that your test says that the person is telling the truth when the person is lying.*

(d) *Repeat (a) and (b) for a general* n. *Hint: There is no need to repeat every step of your previous derivations.*

PROBLEM 42. (Fault Detector) *As an engineer, you are required to design the test performed by a fault-detector for a "black-box" that produces a a sequence of i.i.d. binary random variables* $\cdots, X_1, X_2, X_3, \cdots$. *Previous experience shows that this "black box" has an apriori failure probability of* $\frac{1}{1025}$. *When the "black box" works properly,* $p_{X_i}(1) = p$. *When it fails, the output symbols are equally likely to be* 0 *or* 1.

Your detector has to decide based on the observation of the past 16 *symbols, i.e., at time* k *the decision will be based on* $X_{k-16}, \ldots, X_{k-1}$.

(a) *Describe your test.*

(b) *What does your test decide if it observes the output sequence* 0101010101010101*? Assume that* $p = 1/4$.

PROBLEM 43. (A Simple Multiple-Access Scheme) *Consider the following very simple model of a multiple-access scheme. There are two users. Each user has two hypotheses. Let* $\mathcal{H}^1 = \mathcal{H}^2 = \{0, 1\}$ *denote the respective set of hypotheses and assume that both users employ a uniform prior. Further, let* X^1 *and* X^2 *be the respective signals sent by user one and two. Assume that the transmissions of both users are independent and that* $X^1 \in \{\pm 1\}$ *and* $X^2 \in \{\pm 2\}$ *where* X^1 *and* X^2 *are positive if their respective hypothesis is zero and negative otherwise. Assume that the receiver observes the signal* $Y = X^1 + X^2 + Z$, *where* Z *is a zero mean Gaussian random variable with variance* σ^2 *and is independent of the transmitted signal.*

(a) *Assume that the receiver observes* Y *and wants to estimate* both H_1 *and* H_2. *Let* $\hat{H}^1$ *and* $\hat{H}^2$ *be the estimates. Starting from first principles, what is the generic form of the optimal decision rule?*

(b) *For the specific set of signals given, what is the set of possible observations assuming that* $\sigma^2 = 0$*? Label these signals by the corresponding (joint) hypotheses.*

(c) *Assuming now that* $\sigma^2 > 0$, *draw the optimal decision regions.*

(d) *What is the resulting probability of correct decision? i.e., determine the probability* $Pr\{\hat{H}^1 = H^1, \hat{H}^2 = H^2\}$.

(e) *Finally, assume that we are only interested in the transmission of user two. What is* $Pr\{\hat{H}^2 = H^2\}$*?*

PROBLEM 44. (Uncoded Transmission) *Consider the following transmission scheme. We have two possible sequences* $\{X_j^1\}$ *and* $\{X_j^2\}$ *taking values in* $\{-1, +1\}$, *for* $j = 0, 1, 2, \cdots, k-1$. *The transmitter chooses one of the two sequences and sends it directly over an additive white Gaussian noise channel. Thus, the received value is* $Y_j = X_j^i + Z_j$, *where* $i = 1, 2$ *depending of the transmitted sequence, and* $\{Z_j\}$ *is a sequence of i.i.d. zero-mean Gaussian random variables with variance* σ^2.

(a) *Using basic principles, write down the optimal decision rule that the receiver should implement to distinguish between the two possible sequences. Simplify this rule to express it as a function of inner products of vectors.*

(b) *Let d be the number of positions in which $\{X_j^1\}$ and $\{X_j^2\}$ differ. Assuming that the transmitter sends the first sequences $\{X_j^1\}$, find the probability of error (the probability that the receiver decides on $\{X_j^2\}$), in terms of the Q function and d.*

Problem 45. (Data Dependent Noise) *Consider the following binary Gaussian hypothesis testing problem with data dependent noise.*

Under hypothesis H_0 the transmitted signal is $s_0 = -1$ and the received signal is $Y = s_0 + Z_0$, where Z_0 is zero-mean Gaussian with variance one.

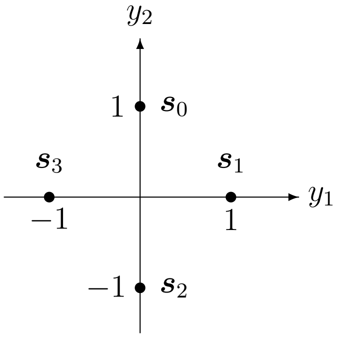
Under hypothesis H_1 the transmitted signal is $s_1 = 1$ and the received signal is $Y = s_1 + Z_1$, where Z_1 is zero-mean Gaussian with variance σ^2. Assume that the prior is uniform.

(a) *Write the optimal decision rule as a function of the parameter σ^2 and the received signal Y.*

(b) *For the value $\sigma^2 = \exp(4)$ compute the decision regions.*

(c) *Give as simple expressions as possible for the error probabilities $P_e(0)$ and $P_e(1)$.*

Problem 46. (Correlated Noise) *Consider the following decision problem. For the hypothesis $H = i$, $i \in \{0,1,2,3\}$, we send the point $\boldsymbol{s}_i$, as follows (also shown in the figure below): $\boldsymbol{s}_0 = (0,1)^T$, $\boldsymbol{s}_1 = (1,0)^T$, $\boldsymbol{s}_2 = (0,-1)^T$, $\boldsymbol{s}_3 = (-1,0)^T$.*



When $H = i$, the receiver observes the vector $\boldsymbol{Y} = \boldsymbol{s}_i + \boldsymbol{Z}$, where $\boldsymbol{Z}$ is a zero-mean Gaussian random vector whose covariance matrix is $\Sigma = \left(\begin{smallmatrix} 4 & 2 \\ 2 & 5 \end{smallmatrix}\right)$

(a) *In order to simplify the decision problem, we transform* $\boldsymbol{Y}$ *into* $\hat{\boldsymbol{Y}} = B\boldsymbol{Y}$, *where* B *is a 2-by-2 matrix, and use* $\hat{\boldsymbol{Y}}$ *to take our decision. What is the appropriated matrix* B *to choose? Hint: If* $A = \frac{1}{4}\left(\begin{smallmatrix} 2 & 0 \\ -1 & 2 \end{smallmatrix}\right)$, *then* $A\Sigma A^T = I$, *with* $I = \left(\begin{smallmatrix} 1 & 0 \\ 0 & 1 \end{smallmatrix}\right)$.

(b) *What are the new transmitted points* $\hat{\boldsymbol{s}}_i$? *Draw the resulting transmitted points and the decision regions associated to them.*

(c) *Give an upper bound to the error probability in this decision problem.*

Problem 47. (Football) *Consider four teams A,B,C,D playing in a football tournament. There are two rounds in the competition. In the first round there are two matches and the winners progress to play in the final. In the first round A plays against one of the other three teams with equal probability* $\frac{1}{3}$ *and the remaining two teams play against each other. The probability of A winning against any team depends on the number of red cards "r" A gets in the previous match. The probabilities of winning for A against B,C,D denoted by* p_b, p_c, p_d *are* $p_b = \frac{0.5}{(1+r)}, p_c = p_d = \frac{0.6}{1+r}$. *In a match against B, team A will get* 1 *red card and in a match against C or D, team B will get* 2 *red cards. Assuming that initially A has* 0 *red cards and the other teams receive no red cards in the entire tournament and among B,C,D each team has equal chances to win against each other.*

Is betting on team A as the winner a good choice ?

Problem 48. (Minimum-Energy Signals) *Consider a given signal constellation consisting of vectors* $\{\boldsymbol{s}_1, \boldsymbol{s}_2, \ldots, \boldsymbol{s}_m\}$. *Let signal* $\boldsymbol{s}_i$ *occur with probability* p_i. *In this problem, we study the influence of moving the origin of the coordinate system of the signal constellation. That is, we study the properties of the signal constellation* $\{\boldsymbol{s}_1 - \boldsymbol{a}, \boldsymbol{s}_2 - \boldsymbol{a}, \ldots, \boldsymbol{s}_m - \boldsymbol{a}\}$ *as a function of* $\boldsymbol{a}$.

(a) *Draw a sample signal constellation, and draw its shift by a sample vector* $\boldsymbol{a}$.

(b) *Does the average error probability,* P_e, *depend on the value of* $\boldsymbol{a}$? *Explain.*

(c) *The average energy per symbol depends on the value of* $\boldsymbol{a}$. *For a given signal constellation* $\{\boldsymbol{s}_1, \boldsymbol{s}_2, \ldots, \boldsymbol{s}_m\}$ *and given signal probabilities* p_i, *prove that the value of* $\boldsymbol{a}$ *that minimizes the average energy per symbol is the centroid (the center of gravity) of the signal constellation, i.e.,*

$$\boldsymbol{a} \;=\; \sum_{i=1}^{m} p_i \boldsymbol{s}_i. \tag{2.35}$$

Hint: First prove that if X *is a real-valued zero-mean random variable and* $b \in \mathbb{R}$, *then* $E[X^2] \leq E[(X-b)^2]$ *with equality iff* $b = 0$. *Then extend your proof to vectors and consider* $\boldsymbol{X} = \boldsymbol{S} - E[\boldsymbol{S}]$ *where* $\boldsymbol{S} = \boldsymbol{s}_i$ *with probability* p_i.

Chapter 3

Receiver Design for the Waveform AWGN Channel

3.1 Introduction

In the previous chapter we have learned how to communicate across the discrete-time AWGN (Additive White Gaussian Noise) channel. Given a transmitter for that channel, we now know what a receiver that minimizes the error probability should do and how to evaluate or bound the resulting error probability. In the current chapter we will deal with a channel model which is closer to reality, namely the *waveform AWGN channel*. This is the channel seen from the input to the output of the dashed box in Figure 3.1. Apart from the channel model, the main objectives of this and the previous chapters are the same: understand what the receiver has to do to minimize the error probability. We are also interested in the resulting error probability but that will come for free from what have learned in the previous chapter.

As in the previous chapter we assume that the signals used to communicate are given to us. While our primary focus will be on the receiver, we will gain valuable insight about the transmitter structure. The problem of choosing suitable signals will be studied in subsequent chapters.

The setup is the one shown in Figure 3.2. The operation of the transmitter is similar to that of the encoder of the previous chapter except that the output is now an element of a set of m finite-energy waveforms $\mathcal{S} = \{s_0(t), \dots, s_{m-1}(t)\}$. The channel adds white Gaussian noise $N(t)$ (defined in the next section). Unless otherwise specified, we assume that the (double-sided) power spectral density of the noise is $\frac{N_0}{2}$.

To emphasize the fact that we are now dealing with waveforms, in the above paragraph as well as in Figure 3.2 we have made an exception to the convention we will use henceforth, namely to use single letters (possibly with an index) to denote waveforms and stochastic processes such as s_i and R. When we want to emphasize the time dependency we may

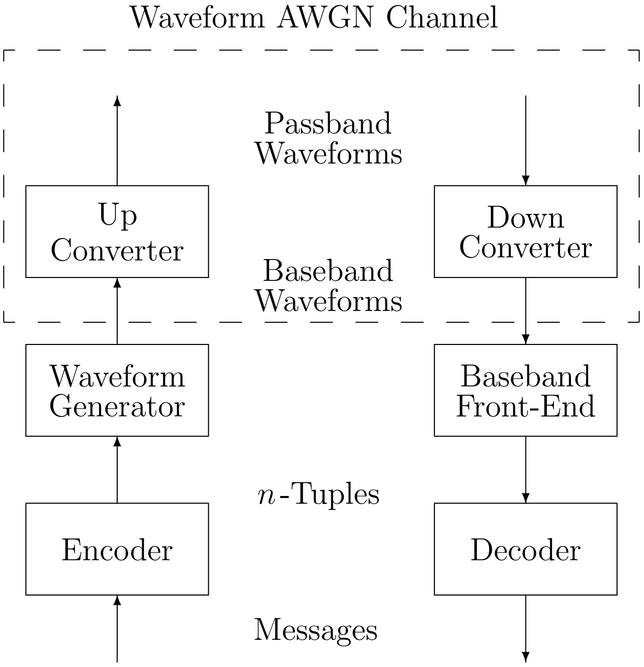


Figure 3.1: Waveform channel abstraction.

also use the equivalent notation $\{s_i(t) : t \in \mathbb{R}\}$ and $\{R(t) : t \in \mathbb{R}\}$.

The highlight of the chapter is the power of abstraction. In the previous chapter we have seen that the receiver design problem for the discrete-time AWGN channel relies on geometrical ideas that may be formulated whenever we are in an inner-produce space (i.e. a vector space endowed with an inner product). Since finite-energy waveforms also form an inner-product space, the methods developed in the previous chapter are appropriate tools also to deal with the waveform AWGN channel.

The main result of this chapter is a decomposition of the sender and the receiver for the waveform AWGN channel into the building blocks that form the bottom two layers in Figure 3.1. We will see that, without loss of generality, we may (and should) think of the transmitter as consisting of a part that maps the message $i \in \mathcal{H}$ into an n-tuple $\boldsymbol{s}_i$, as in the previous chapter, followed by a *waveform generator* that maps $\boldsymbol{s}_i$ into a waveform s_i. Similarly, we will see that the receiver may consist of a *front-end* that takes the channel output and produces an n-tuple $\boldsymbol{Y}$ which is a sufficient statistic. From the waveform generator input to the receiver front-end output we see the discrete-time AWGN channel considered in the previous chapter. Hence we know already what the decoder of Fig. 3.1 should do with the sufficient statistic produced by the receiver front-end.

In this chapter we assume familiarity with the linear space $\mathcal{L}_2$ of finite energy functions. See Appendix 2.E for a review.

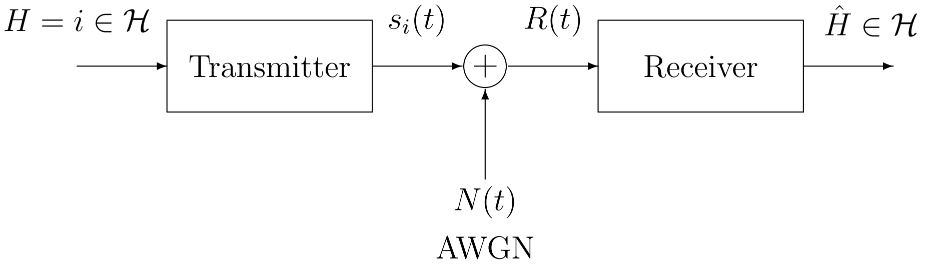


Figure 3.2: Communication across the AWGN channel.

3.2 Gaussian Processes and White Gaussian Noise

We assume that the reader is familiar with: (i) the definition of a wide-sense-stationary (wss) stochastic process; (ii) the notion of autocorrelation and power spectral density; (iii) the definition of a Gaussian random vector.

Definition 48. $\{N(t) : t \in \mathbb{R}\}$ *is a* Gaussian random process *if for any finite collection of times* $t_1, t_2, \ldots, t_k$ *, the vector* $\boldsymbol{Z} = (N(t_1), N(t_2), \ldots, N(t_k))^T$ *of samples is a Gaussian random vector. A second process* $\{\tilde{N}(t) : t \in \mathbb{R}\}$ *is* jointly Gaussian *with* N *if* $\boldsymbol{Z}$ *and* $\tilde{\boldsymbol{Z}}$ *are jointly Gaussian random vectors for any vector* $\tilde{\boldsymbol{Z}}$ *consisting of samples from* $\tilde{N}$ *.*

The definition of white Gaussian noise requires an introduction. Many communication textbooks define white Gaussian noise to be a zero-mean wide-sense-stationary Gaussian random process $\{N(t) : t \in \mathbb{R}\}$ of autocorrelation $K_N(\tau) = \frac{N_0}{2}\delta(\tau)$. This definition is simple and useful but mathematically problematic. To see why, recall that a Gaussian random variable has finite variance.[1] The sample $N(t)$ at an arbitrary epoch t is a Gaussian random variable of variance $K_N(0) = \frac{N_0}{2}\delta(0)$. But $\delta(0)$ is not defined. One may be tempted to say that $\delta(0) = \infty$ but this would mean that the sample is a Gaussian random variable of infinite variance.[2]

Our goal is a consistent model that leads to the correct observations. The noise we are trying to model shows up when we make real-world measurements. If $N(t)$ models electromagnetic noise, then its effect will show up as a voltage at the output of an antenna. If $N(t)$ models the noise in an electrical cable, then it shows up when we measure the voltage on the cable. In any case the measurement is done via some piece of wire (the antenna or the tip of a probe) which is modeled as a *linear time invariant system* of some

[1] The Gaussian probability density is not defined when the variance is infinite.

[2] One way to deal with this problem is to define $\{N(t) : t \in \mathbb{R}\}$ as generalized Gaussian random process. We choose a different approach that allows us to rely on familiar tools.

finite energy impulse response g.[3] Hence we are limited to observations of the kind

$$Z(t) = \int N(\alpha)g(t-\alpha)d\alpha.$$

We define white Gaussian noise $N(t)$ by defining what we obtain from an arbitrary but finite collection of such measurements.

Definition 49. $\{N(t) : t \in \mathbb{R}\}$ *is zero-mean white Gaussian noise of power spectral density* $\frac{N_0}{2}$ *if for any finite collection of* $\mathcal{L}_2$ *functions* $g_1(t), g_2(t), \dots, g_k(t)$,

$$Z_i(t) = \int N(\alpha)g_i(t-\alpha)d\alpha, \quad i = 1, 2, \dots, k$$

is a collection of zero-mean jointly Gaussian random processes with covariances

$$cov\big(Z_i(\beta), Z_j(\gamma)\big) = E\left[Z_i(\beta)Z_j^*(\gamma)\right] = \frac{N_0}{2}\int g_i(t)g_j^*(t+\gamma-\beta)dt. \tag{3.1}$$

Exercise 50. *Show that (3.1) is precisely what we obtain if we define white Gaussian noise to be a Gaussian noise process of autocorrelation* $K_N(\tau) = \frac{N_0}{2}\delta(\tau)$. □

A few comments are in order. First, the fact that we are defining $N(t)$ indirectly is consistent with the fact that for no time t we can observe $N(t)$. Second, defining an object—zero-mean white Gaussian noise in this case—via what we see when we integrate that object against a finite-energy function g is not new: we do the same when we define a delta Dirac $\delta(t)$ by saying that $\int g(t)\delta(t) = g(0)$. Third, our definition does not require proving that a Gaussian process $N(t)$ that has the desired properties exits. In fact $N(t)$ may not be Gaussian but such that when filtered and then sampled at a finite number of times forms a collection of zero-mean jointly Gaussian random variables. If the reader is uncomfortable with the idea that we are integrating against an object that we have not defined—and in fact may not even exist— then he/she can choose to think of $N(t)$ as being the *name* of some undefined physical phenomenon that we call zero-mean Gaussian noise and think of $\int N(\alpha)g_i(t-\alpha)d\alpha$ not as a convolution between two functions of time but rather as a place holder for *what we see when we observe zero-mean Gaussian noise through a filter of impulse response* g_i. In doing so we model the result of the measurement and not the signal we are measuring. Finally, no matter whether we use the more common definition of white Gaussian noise mentioned earlier or the one we are using, a model is only an approximation of reality: if we could make measurements with arbitrary impulse responses g_i, at some point we would discover that our model is not accurate. To be specific, if g_i is the impulse response of an ideal bandpass filter of 1 Hz of bandwidth, then the idealized model of white Gaussian noise says that for any fixed t the random variable $Z_i(t)$ has variance $N_0/2$. If we could increase the center frequency indefinitely, at some point we would observe that the variance of the real measurements starts decreasing. This must be the case since the underlying physical signal can not

[3] We neglect the noise introduced by the measurement since it can be accounted for by $N(t)$.

have infinite power. We are not concerned about this potential discrepancy between the model and real measurements since we are unable to make measurements involving filers of arbitrarily large center frequency.

By far the most common measurements we will be concerned with in relationship to white Gaussian noise N of power spectral density $\frac{N_0}{2}$ are of the kind

$$Z_i = \int N(\alpha) g_i(\alpha) dt, \quad i = 1, 2, \ldots, k.$$

Then $\boldsymbol{Z} = (Z_1, \ldots, Z_k)^T$ is a zero-mean Gaussian random vector and the i, j element of its covariance matrix is

$$E[Z_i, Z_j] = \frac{N_0}{2} \int g_i(t) g_j^*(t) dt. \tag{3.2}$$

Of particular interest is the special case when the waveforms $g_1(t), \ldots, g_k(t)$ form an orthonormal set. Then $\boldsymbol{Z} \sim \mathcal{N}(0, \frac{N_0}{2} I_k)$.

3.2.1 Observables and Sufficient Statistic

By assumption the channel output is $R = s_i + N$ for some $i \in \mathcal{H}$ and N is white Gaussian noise. As discussed in the previous section, due to the nature of the white noise the channel output R is not observable. What we can observe via measurements is the integral of R against any number of finite-energy waveforms. Hence we may consider as the observable any k-tuple $\boldsymbol{V} = (V_1, \ldots, V_k)^T$ such that

$$V_i = \int_{\infty}^{\infty} R(\alpha) g_i^*(\alpha) d\alpha, \qquad i = 1, 2, \ldots, k \tag{3.3}$$

We are choosing k to be finite as part of our model since no one can make infinite measurements.[4]

Notice that the kind of measurements we are considering is quite general. For instance, we can pass R through an ideal lowpass filter of cutoff frequency B for some huge B (say 10^{10} Hz) and collect an arbitrary large number of samples taken every $\frac{1}{2B}$ seconds so as to fulfill the sampling theorem. In fact, by choosing $g_i(t) = h(\frac{i}{2B} - t)$, where $h(t)$ is the impulse response of the lowpass filter, V_i becomes the filter output sampled at time $t = \frac{i}{2B}$.

Let $\mathcal{W}$ be the inner-product space spanned by $\mathcal{S}$ and let $\{\psi_1, \ldots, \psi_n\}$ be an arbitrary orthonormal basis for $\mathcal{W}$. We claim that the n-tuple $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^T$ with i-th component

$$Y_i = \int R(\alpha) \psi_i^*(\alpha) d\alpha$$

[4] By letting k be infinite we would have to deal with subtle issues of infinity without gaining anything in practice.

is a sufficient statistic among any collection of measurements that contains $\boldsymbol{Y}$. To prove this claim, let $\boldsymbol{U} = (U_1, U_2, \dots, U_k)^T$ be the vector of all the other measurements we may want to consider. The only requirement is that they be consistent with (3.3). Let $\mathcal{V}$ be the inner product space spanned by $\mathcal{S} \cup \{g_1, g_2, \dots, g_k\}$ and let $\{\psi_1, \dots, \psi_n, \phi_1, \phi_2, \dots, \phi_{\tilde{n}}\}$ be an orthonormal basis for $\mathcal{V}$. Define

$$V_i = \int R(\alpha)\phi_i^*(\alpha)d\alpha, \qquad i = 1, \dots, \tilde{n}.$$

There is a one-to-one correspondence between $(\boldsymbol{Y}, \boldsymbol{U})$ and $(\boldsymbol{Y}, \boldsymbol{V})$. Hence the latter may be considered as *the observable.* Note that when $H = i$,

$$\begin{aligned}
Y_j &= \int R(\alpha)\psi_j^*(\alpha) = \int \big(s_i(\alpha) + N(\alpha)\big)\psi_j^*(\alpha)d\alpha = s_{i,j} + \int N(\alpha)\psi_j^*(\alpha)d\alpha,\\
V_j &= \int R(\alpha)\phi_j^*(\alpha) = \int \big(s_i(\alpha) + N(\alpha)\big)\phi_j^*(\alpha)d\alpha = \int N(\alpha)\phi_j^*(\alpha)d\alpha,
\end{aligned}$$

where we used the fact that s_i is in the subspace spanned by $\{\psi_1, \dots, \psi_n\}$ and therefore it is orthogonal to ϕ_j for each $j = 1, 2, \dots, \tilde{n}$. Hence when $H = i$,

$$\begin{aligned}
\boldsymbol{Y} &= \boldsymbol{s}_i + \boldsymbol{N}_{|\mathcal{W}},\\
\boldsymbol{V} &= \boldsymbol{N}_{\perp},
\end{aligned}$$

where $\boldsymbol{N}_{|\mathcal{W}} \sim \mathcal{N}(0, \frac{N_0}{2}I_n)$ and $\boldsymbol{N}_{\perp} \sim \mathcal{N}(0, \frac{N_0}{2}I_{\tilde{n}})$. Furthermore, $\boldsymbol{N}_{|\mathcal{W}}$ and $\boldsymbol{N}_{\perp}$ are independent of each other and of H. In particular, $H \rightarrow \boldsymbol{Y} \rightarrow \big(\boldsymbol{Y}, \boldsymbol{V}\big)$, showing that $\boldsymbol{Y}$ is indeed a sufficient statistic. Hence $\boldsymbol{V}$ is irrelevant as claimed.[5]

To gain additional insight, let Y be the waveform associated to $\boldsymbol{Y}$, i.e., $Y(t) = \sum Y_i\psi_i(t)$, and similarly let $N_{|\mathcal{W}}$ and $N_{\perp}$ be the waveforms associated to $\boldsymbol{N}_{|\mathcal{W}}$ and $\boldsymbol{N}_{\perp}$, respectively. Then we may define $\tilde{N}$ via the equality $R = Y + \tilde{N}$. These quantities have the following interpretation

$$\begin{aligned}
Y = s_i + N_{|\mathcal{W}} &= \text{“Projection” of the received signal } R \text{ onto } \mathcal{W}\\
N_{|\mathcal{W}} &= \text{“Projection” of the noise } N \text{ onto } \mathcal{W}\\
N_{\perp} &= \text{Noise component captured by the measurement but orthogonal to } \mathcal{W}\\
\tilde{N} &= \text{Noise component “orthogonal to } \mathcal{W}\text{”}
\end{aligned}$$

where quotations are due since projection and orthogonality are defined for elements of an inner product space and have made no claim about the belonging of R and N to such a space. Nevertheless one can compute the integral of R against ϕ_i^* or ψ_i^* so that the meaning of “projection” is well defined. Hereafter we will drop the quotation when we speak about the ”projection” of R onto $\mathcal{W}$. Similarly, by “orthogonality” of $\tilde{N}$ and $\mathcal{W}$ we mean that the integral $\tilde{N}$ against any element of $\mathcal{W}$ vanishes. Figure 3.3 gives a geometric interpretation of the various quantities.

[5] We have not proved that R is irrelevant. There is no reason to prove that since R is not observable.

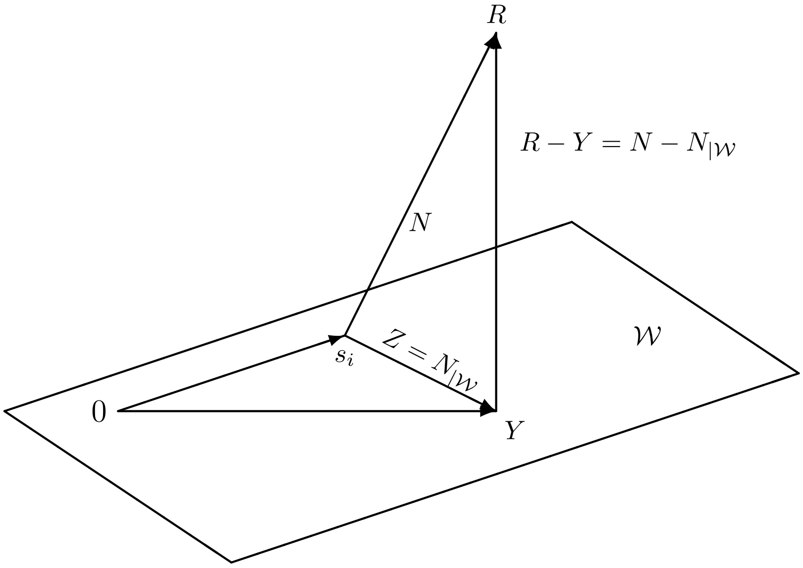


Figure 3.3: Projection of the received signal R onto $\mathcal{W}$ when $H = i$.

The receiver *front-end* that computes $\boldsymbol{Y}$ from R is shown in Figure 3.4. The figure also shows that one can single out a corresponding block at the sender, namely the *waveform generator* that produces s_i from the n-tuple $\boldsymbol{s}_i$. Of course one can generate the signal s_i without the intermediate step of generating the n-tuple of coefficients $\boldsymbol{s}_i$ but thinking in terms of the two-step procedure underlines the symmetry between the sender and the receiver and emphasizes the fact that dealing with the waveform AWGN channel just adds a layer of processing with respect to dealing with the discrete-time AWGN counterpart.

From the waveform generator input to the baseband front-end output we "see" the discrete-time AWGN channel studied in Chapter 2. In fact the decoder faces precisely the same decision problem which is to do a ML decision for the hypothesis testing problem specified by

$$H = i : \qquad \boldsymbol{Y} = \boldsymbol{s}_i + \boldsymbol{Z}.$$

where $\boldsymbol{Z} \sim \mathcal{N}(0, \frac{N_0}{2} I_n)$ is independent of H.

It would seem that we are done with the receiver design problem for the waveform AWGN channel. In fact we are done with the conceptual part. In the rest of the chapter we will gain additional insight by looking at some of the details and by working out a few examples. What we can already say at this point is that the sender and the receiver may be decomposed as shown in Figure 3.5 and that the channel seen between the encoder/decoder pair is the *discrete-time* AWGN channel considered in the previous chapter. Later we will see that the decomposition is not just useful at a conceptual level. In fact *coding* is a subarea of digital communication devoted to the study of encoders/decoders. In a broad sense, *modulation* is likewise an area devoted to the study of waveform generators and

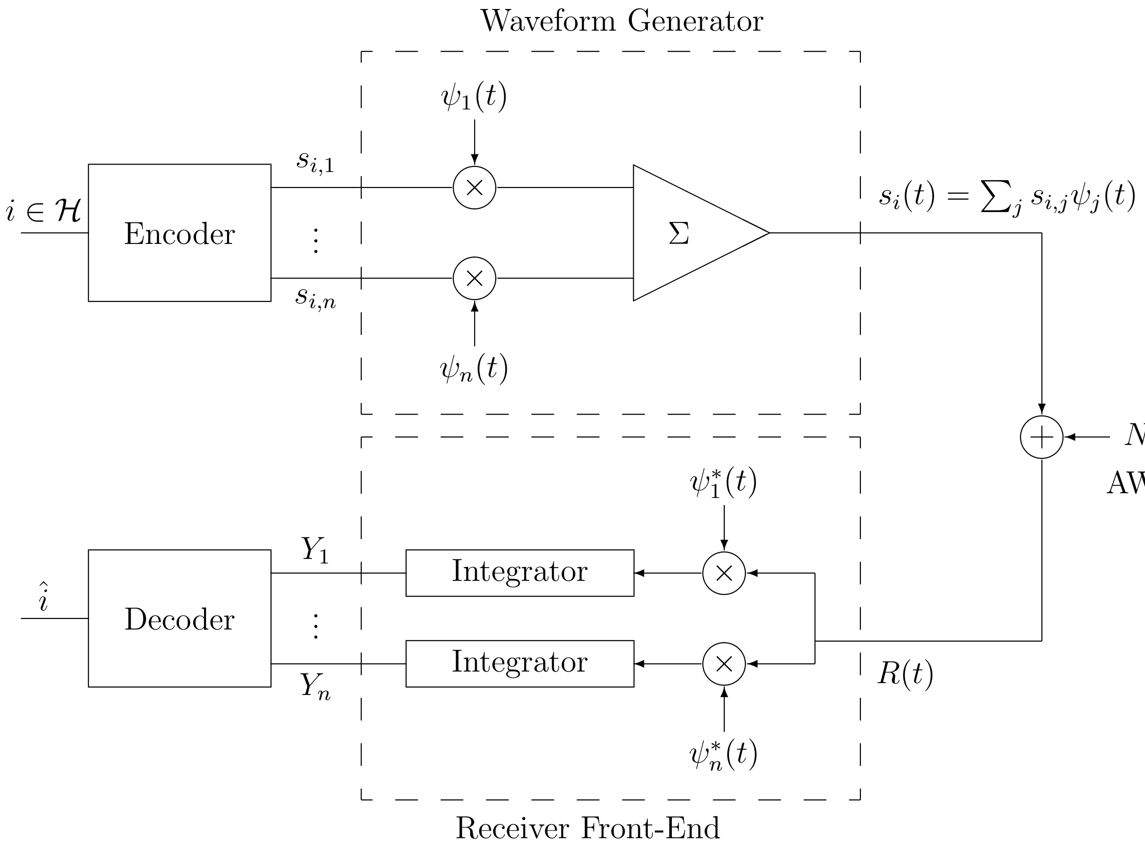


Figure 3.4: Waveform sender/receiver pair

baseband front-ends.

3.3 The Binary Equiprobable Case

We start with the binary hypothesis case since it allows us to focus on the essential. Generalizing to m hypotheses will be straightforward. We also assume $P_H(0) = P_H(1) = 1/2$.

3.3.1 Optimal Test

The test that minimizes the error probability is the ML decision rule:

$$\|\boldsymbol{y} - \boldsymbol{s}_0\|^2 \underset{\hat{H}=0}{\overset{\hat{H}=1}{\gtrless}} \|\boldsymbol{y} - \boldsymbol{s}_1\|^2.$$

As usual, ties may be resolved either way.

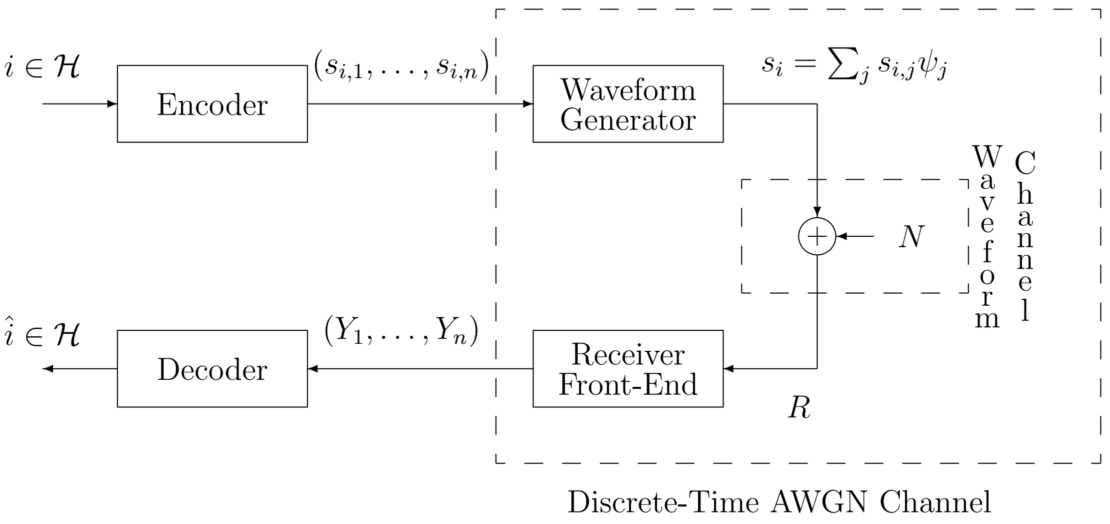


Figure 3.5: Canonical decomposition of the transmitter for the waveform AWGN channel into and encoder and a waveform generator. The receiver decomposes into a front-end and a decoder. From the waveform generator input to the receiver front end output we see the n-tuple AWGN channel

3.3.2 Receiver Structures

There are various ways to implement the receiver since :

(a) the ML test can be rewritten in various ways

(b) there are two basic ways to implement an inner product.

Hereafter are three equivalent ML tests. The fist is conceptual whereas the the second and third suggest receiver implementations. They are:

$$\|\boldsymbol{y} - \boldsymbol{s}_0\| \underset{\hat{H}=0}{\overset{\hat{H}=1}{\gtrless}} \|\boldsymbol{y} - \boldsymbol{s}_1\| \qquad \text{(T1)}$$

$$\Re\big[\langle \boldsymbol{y}, \boldsymbol{s}_1\rangle\big] - \frac{\|\boldsymbol{s}_1\|^2}{2} \underset{\hat{H}=0}{\overset{\hat{H}=1}{\gtrless}} \Re\big[\langle \boldsymbol{y}, \boldsymbol{s}_0\rangle\big] - \frac{\|\boldsymbol{s}_0\|^2}{2} \qquad \text{(T2)}$$

$$\Re\Big[\int R(t)s_1^*(t)dt\Big] - \frac{\|s_1\|^2}{2} \underset{\hat{H}=0}{\overset{\hat{H}=1}{\gtrless}} \Re\Big[\int R(t)s_0^*(t)dt\Big] - \frac{\|s_0\|^2}{2} \qquad \text{(T3)}$$

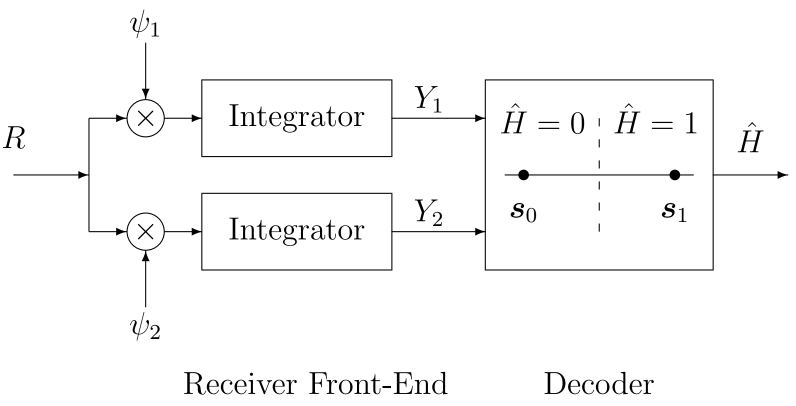


Figure 3.6: Implementation of test (T1). The front-end is based on correlators. This is the part that converts the received waveform into an $n$tuple $\boldsymbol{Y}$ which is a sufficient statistic. From this point on the decision problem is the one considered in the previous chapter.

Test (T1) is the test described in the previous section after taking the square root on both sides. Since the square root of a nonnegative number is a monotonic operation, the test outcome remains unchanged. Test (T1) is useful to visualize decoding regions and to compute the probability of error. It says that the decoding region of $\boldsymbol{s}_0$ is the set of $\boldsymbol{y}$ that are closer to $\boldsymbol{s}_0$ than to $\boldsymbol{s}_1$. (We knew this already from the previous chapter.)

Figure 3.6 shows the block diagram of a receiver inspired by (T1). The receiver front-end maps R into $\boldsymbol{Y}=(Y_1,Y_2)^T$. This part of the receiver deals with waveforms and in the past it has been implemented via analog circuitry. A modern implementation would typically sample the received signal after passing it through a filter to ensure that the condition of the sampling theorem is fulfilled. The filter is designed so as to be transparent to the signal waveforms. The filter removes part of the noise that would anyhow be removed by the receiver front end. The decoder chooses the index i of the $\boldsymbol{s}_i$ that minimizes $\|\boldsymbol{y}-\boldsymbol{s}_i\|$. Test (T1) does not explicitly say how to find that index. We imagine the decoder as a conceptual device that knows the decoding regions and checks which decoding region contains $\boldsymbol{y}$. The decoder shown in Figure 3.6 assumes *antipodal signals*, i.e., $s_0=-s_1$, and $\psi_1=s_1/\|s_1\|$. In this case the signal space is one-dimensional. A decoder such as this one that decides upon comparing the components of $\boldsymbol{Y}$ (in this case one component) to thresholds is sometimes called a *slicer*.

A decoder for a 2-dimensional signal space spanned by orthogonal signals s_0 and s_1 would decide based on the decoding regions shown in Figure 3.7, where we defined $\psi_1=s_0/\|s_0\|$ and $\psi_2=s_1/\|s_1\|$.

Perhaps the biggest advantage of test (T1) is the geometrical insight it provides which is often useful to determine the error probability.

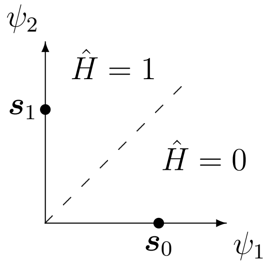


Figure 3.7: Decoding regions for two orthogonal signals

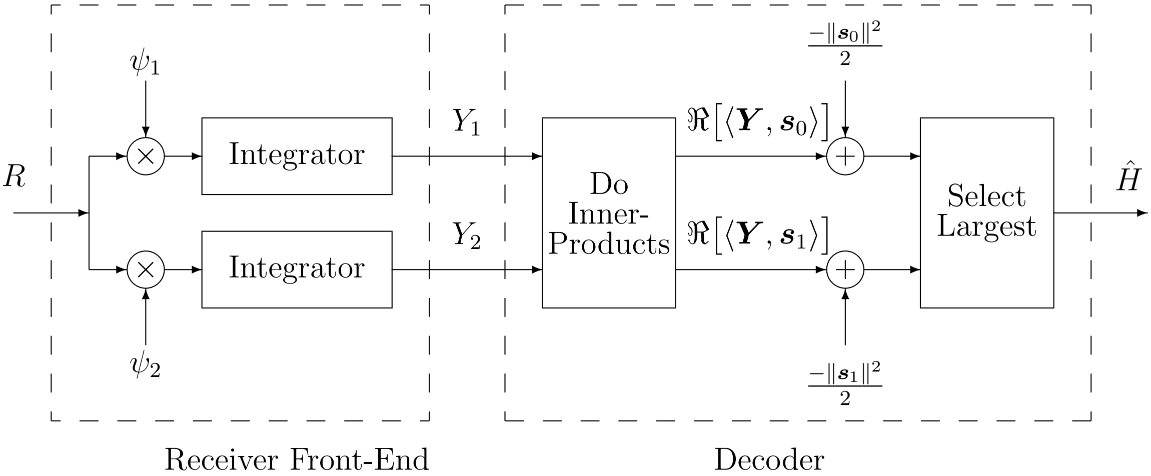

ψ_1
R
Integrator
Y_1
Integrator
Y_2
ψ_2
Do Inner-Products
$\Re[\langle \boldsymbol{Y}, \boldsymbol{s}_0\rangle]$
$\Re[\langle \boldsymbol{Y}, \boldsymbol{s}_1\rangle]$
$\frac{-\|\boldsymbol{s}_0\|^2}{2}$
$\frac{-\|\boldsymbol{s}_1\|^2}{2}$
Select Largest
$\hat{H}$
Receiver Front-End
Decoder

Figure 3.8: Receiver implementation following (T2). This implementation requires an orthonormal basis. Finding and implementing waveforms that constitute an orthonormal basis may or may not be easy.

Test (T2) is obtained from (T1) using the relationship

$$\begin{aligned}\|\boldsymbol{y} - \boldsymbol{s}_i\|^2 &= \langle \boldsymbol{y} - \boldsymbol{s}_i, \boldsymbol{y} - \boldsymbol{s}_i \rangle \\ &= \|\boldsymbol{y}\|^2 - 2\Re\{\langle \boldsymbol{y}, \boldsymbol{s}_i \rangle\} + \|\boldsymbol{s}_i\|^2,\end{aligned}$$

after canceling out common terms, multiplying each side by $-1/2$, and using the fact that $a > b$ iff $-a < -b$. Test (T2) is implemented by the block diagram of Figure 3.8. The added value of the decoder in Figure 3.8 is that it is completely specified in terms of easy-to-implement operations. However, it looses some of the geometrical insight present in a decoder that depicts the decoding regions as in Figure 3.6.

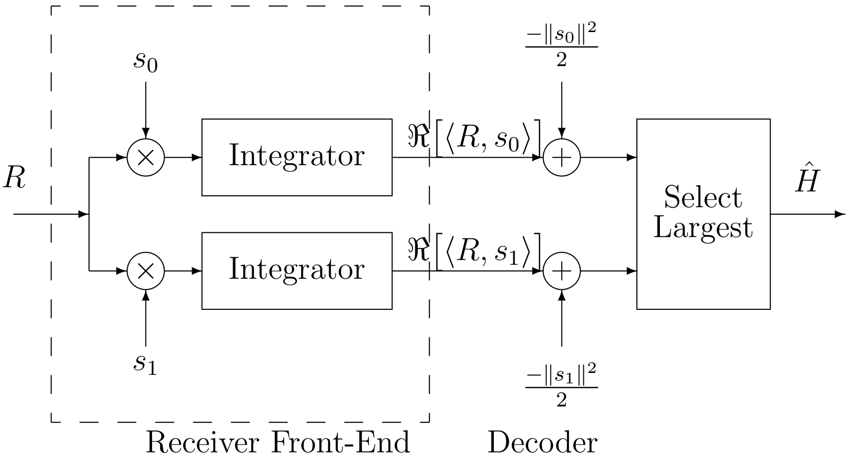


Figure 3.9: Receiver implementation following (T3). Notice that this implementation does not rely on an orthonormal basis.

Test (T3) is obtained from (T2) via Parseval's relationship and a bit more to account for the fact that projecting R onto s_i is the same as projecting Y. Specifically, for $i = 1, 2$,

$$\begin{aligned}\langle \boldsymbol{y}, \boldsymbol{s}_i\rangle &= \langle Y, s_i\rangle \\ &= \langle Y + N_\perp, s_i\rangle \\ &= \langle R, s_i\rangle.\end{aligned}$$

Test (T3) is implemented by the block diagram in Figure 3.9. The subtraction of half the signal energy in (T2) and (T3) is of course superfluous when all signals have the same energy.

Even tough the mathematical expression for the test (T2) and (T3) look similar, the tests differ fundamentally and practically. First of all, (T3) does not require finding a basis for the signal space spanned by $\mathcal{W}$. As a side benefit, this proves that the receiver performance does not depend on the basis used to perform (T2) (or (T1) for that matter). Second, Test (T2) requires an extra layer of computation, namely that needed to perform the inner products $\langle \boldsymbol{y}, \boldsymbol{s}_i\rangle$. This step comes for free in (T3) (compare Figures 3.8 and 3.9). However, the number of integrators needed in Figure 3.9 equals the number m of hypotheses (2 in our case), whereas that in Figure 3.8 equals to dimensionality n of the signal space $\mathcal{W}$. We know that $n \leq m$ and one can easily construct examples where equality holds or where $n \ll m$. In the latter case it is preferable to implement test (T2). This point will become clearer and more relevant when the number m of hypotheses is large. It should also be pointed out that the block diagram of Figure 3.9 does *not* quite fit into the decomposition of Figure 3.5 (the n-tuple $\boldsymbol{Y}$ is not produced).

Each of the tests (T1), (T2), and (T3) can be implemented two ways. One way is shown in Figs. 3.6, 3.8 and 3.9, respectively. The other way makes use of the fact that the

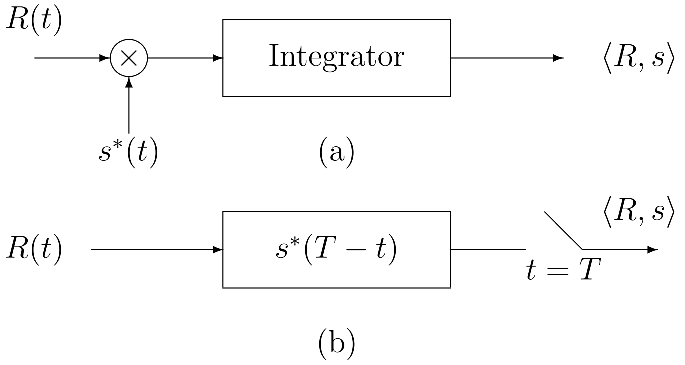


Figure 3.10: Two ways to implement the projection $\langle R, s\rangle$, namely via a "correlator" (a) and via a "matched filter" (b).

operation

$$\langle R, s\rangle = \int R(t)s^*(t)dt$$

can always be implemented by means of a filter of impulse response $h(t) = s^*(T-t)$ as shown in Figure 3.10 (b), where T is an arbitrary delay selected in such a way as to make h a causal impulse response. To verify that the implementation of Figure (3.10)(b) also leads to $\langle R, s\rangle$, we proceed as follows. Let y be the filter output when the input is R. If $h(t) = s^*(T-t)$, $t \in \mathbb{R}$, is the filter impulse response, then

$$y(t) = \int R(\alpha)\ h(t-\alpha)\ d\alpha = \int R(\alpha)\ s^*(T+\alpha-t)\ d\alpha.$$

At $t = T$ the output is

$$y(T) = \int R(\alpha)\ s^*(\alpha)\ d\alpha,$$

which is indeed $\langle R, s\rangle$ (by definition). The implementation of Figure 3.10(b) is referred to as *matched-filter implementation* of the receiver front-end. In each of the receiver front ends shown in Figs. 3.6, 3.8 and 3.9, we can substitute matched filters for correlators.

3.3.3 Probability of Error

We compute the probability of error the exact same way as we did in Section 2.4.2. As we have seen, the computation is straightforward when we have only two hypotheses. From test (T1) we see that when $H = 0$ we make an error if $\boldsymbol{Y}$ is closer to $\boldsymbol{s}_1$ than to $\boldsymbol{s}_0$. This happens if the projection of the noise N in direction $\boldsymbol{s}_1 - \boldsymbol{s}_0$ has length exceeding $\frac{\|\boldsymbol{s}_1-\boldsymbol{s}_0\|}{2}$. This event has probability $P_e(0) = Q\big(\frac{\|\boldsymbol{s}_1-\boldsymbol{s}_0\|}{2\sigma}\big)$ where $\sigma^2 = \frac{N_0}{2}$ is the variance of the projection of the noise in any direction. By symmetry, $P_e(1) = P_e(0)$. Hence

$$P_e = \frac{1}{2}P_e(1) + \frac{1}{2}P_e(0) = Q\left(\frac{\|\boldsymbol{s}_1-\boldsymbol{s}_0\|}{\sqrt{2N_0}}\right) = Q\left(\frac{\|s_1-s_0\|}{\sqrt{2N_0}}\right),$$

where we use the fact that

$$\|\boldsymbol{s}_1 - \boldsymbol{s}_0\| = \|s_1 - s_0\| = \sqrt{\int [s_1(t) - s_0(t)]^2 dt}.$$

It is interesting to observe that the probability of error depends only on the distance $\|s_1 - s_0\|$ and not on the particular shape of the waveforms s_0 and s_1. This fact is illustrated in the following example.

Example 51. *Consider the following signal choices and verify that, in all cases, the corresponding* n*-tuples are* $\boldsymbol{s}_0 = (\sqrt{\mathcal{E}}, 0)^T$ *and* $\boldsymbol{s}_1 = (0, \sqrt{\mathcal{E}})^T$. *To reach this conclusion, it is enough to verify that* $\langle s_i, s_j \rangle = \mathcal{E}\delta_{ij}$, *where* δ_{ij} *equals* 1 *if* $i = j$ *and* 0 *otherwise. This means that, in each case,* s_0 *and* s_1 *are orthogonal and have squared norm* $\mathcal{E}$.

Choice 1 (Rectangular Pulse Position Modulation) :

$$\begin{aligned} s_0(t) &= \sqrt{\frac{\mathcal{E}}{T}}\, 1_{[0,T]}(t) \\ s_1(t) &= \sqrt{\frac{\mathcal{E}}{T}}\, 1_{[T,2T]}(t), \end{aligned}$$

where we have used the indicator function $1_{\mathcal{I}}(t)$ *to denote a rectangular pulse which is* 1 *in the interval* $\mathcal{I}$ *and* 0 *elsewhere. Rectangular pulses can easily be generated, e.g. by a switch. They are used to communicate binary symbols within a circuit. A drawback of rectangular pulses is that they have infinite support in the frequency domain.*

Choice 2 (Frequency Shift Keying):

$$\begin{aligned} s_0(t) &= \sqrt{\frac{2\mathcal{E}}{T}} \sin\left(\pi k \frac{t}{T}\right) 1_{[0,T]}(t) \\ s_1(t) &= \sqrt{\frac{2\mathcal{E}}{T}} \sin\left(\pi l \frac{t}{T}\right) 1_{[0,T]}(t), \end{aligned}$$

where k *and* l *are positive integers,* $k \neq l$. *With a large value of* k *and* l, *these signals could be used for wireless communication. Also these signals have infinite support in the frequency domain. Using the trigonometric identity* $\sin(\alpha)\sin(\beta) = \cos(\alpha - \beta) - \cos(\alpha + \beta)$, *it is straightforward to verify that the signals are orthogonal.*

Choice 3 (Sinc Pulse Position Modulation):

$$\begin{aligned} s_0(t) &= \sqrt{\frac{\mathcal{E}}{T}}\, \operatorname{sinc}\left(\frac{t}{T}\right) \\ s_1(t) &= \sqrt{\frac{\mathcal{E}}{T}}\, \operatorname{sinc}\left(\frac{t-T}{T}\right) \end{aligned}$$

The biggest advantage of sinc *pulses is that they have finite support in the frequency domain. This means that they have infinite support in the time domain. In practice one uses a truncated version of the time domain signal.*

Choice 4 (Spread Spectrum):

$$s_0(t) = \sqrt{\frac{\mathcal{E}}{T}} \sum_{j=1}^{n} s_{0j} 1_{[0,\frac{T}{n}]}\left(t - j\frac{T}{n}\right)$$
$$s_1(t) = \sqrt{\frac{\mathcal{E}}{T}} \sum_{j=1}^{n} s_{1j} 1_{[0,\frac{T}{n}]}\left(t - j\frac{T}{n}\right)$$

where $\underline{s}_0 = (s_{01}, \dots, s_{0n})^T$ *and* $\underline{s}_1 = (s_{11}, \dots, s_{1n})^T$ *are orthogonal and have square norm* $\mathcal{E}$*. This signaling method is called spread spectrum. It uses much bandwidth but it has an inherent robustness with respect to interfering (non-white) signals.*

As a function of time, the above signal constellations are all quite different. Nevertheless, when used to signal across the waveform AWGN channel they all lead to the same probability of error. □

3.4 The m-ary Case

Generalizing to the m-ary case is straightforward. In this section we let the prior P_H be general (not necessarily uniformly distributed as thus far in this chapter). So $H = i$ with probability $P_H(i)$, $i \in \mathcal{H}$. When $H = i$, $R = s_i + N$ where $s_i \in \mathcal{S}$, $\mathcal{S} = \{s_0, s_1, \dots, s_{m-1}\}$ is the signal constellation assumed to be known to the receiver, and N is white Gaussian noise.

We assume that we have selected an orthonormal basis $\{\psi_1, \psi_2, \dots, \psi_n\}$ for the vector space $\mathcal{W}$ spanned by $\mathcal{S}$. Like for the binary case, it will turn out that an optimal receiver can be implemented without going through the step of finding an orthonormal basis. At the receiver we obtain a sufficient statistic by projecting the received signal R onto each of the basis vector. The result is:

$$\boldsymbol{Y} = (Y_1, Y_2, \dots, Y_n)^T \text{ where}$$
$$Y_i = \langle R, \psi_i \rangle, \quad i = 1, \dots, n.$$

The decoder "sees" the vector hypothesis testing problem

$$H = i: \qquad \boldsymbol{Y} = \boldsymbol{s}_i + \boldsymbol{Z} \sim \mathcal{N}(\boldsymbol{s}_i, \frac{N_0}{2} I_n)$$

studied in Chapter 2. The receiver observes $\boldsymbol{y}$ and decides for $\hat{H} = i$ only if

$$P_H(i) f_{\boldsymbol{Y}|H}(\boldsymbol{y}|i) = \max_k \{P_H(k) f_{\boldsymbol{Y}|H}(\boldsymbol{y}|k)\}.$$

Any receiver that satisfies this decision rule minimizes the probability of error. If the maximum is not unique, the receiver may declare any of the hypotheses that achieves the maximum.

For the additive white Gaussian channel under consideration

$$f_{\boldsymbol{Y}|H}(\boldsymbol{y}|i) = \frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}} \exp\left(-\frac{\|\boldsymbol{y}-\boldsymbol{s}_i\|^2}{2\sigma^2}\right)$$

where $\sigma^2 = \frac{N_0}{2}$. Plugging into the above decoding rule, taking the log which is a monotonic function, multiplying by minus N_0, and canceling terms that do not depend on i, we obtain that a MAP decoder decides for one of the $i \in \mathcal{H}$ that minimizes

$$-N_0 \ln P_H(i) + \|\boldsymbol{y}-\boldsymbol{s}_i\|^2.$$

The expression should be compared to test (T1) of the previous section. The manipulations of $\|y-s_i\|^2$ that have led to test (T2) and (T3) are valid also here. In particular, the equivalent of (T2) consists of maximizing.

$$\langle \boldsymbol{y}, \boldsymbol{s}_i \rangle + c_i$$

where $c_i = \frac{1}{2}(N_0 \ \ln \ P_H(i) - \|s_i\|^2)$. Finally, we can use Parseval's relationship to substitute $\langle R, s_i \rangle$ for $\langle \boldsymbol{Y}, \boldsymbol{s}_i \rangle$ and get rid of the need to find an orthonormal basis. This leads to the generalization of (T3), namely

$$\langle R, s_i \rangle + c_i.$$

Figure 3.11 shows three MAP receivers where the receiver front end is implemented via a bank of matched filters. Three alternative forms are obtained by using correlators instead of matched filters. In the first figure, the decoder partitions $\mathbb{C}^n$ into decoding regions. The decoding region for $H = i$ is the set of points $\boldsymbol{y} \in \mathbb{C}^n$ for which

$$-N_0 \ \ln \ P_H(k) + \|\boldsymbol{y} - \boldsymbol{s}_k\|^2$$

is minimized when $k = i$. Notice that in the first two implementations there are n matched filters, where n is the dimension of the signal space $\mathcal{W}$ spanned by the signals in $\mathcal{S}$, whereas in the third implementation the number of matched filters equals the number m of signals in $\mathcal{S}$. In general, $n \leq m$. If $n = m$, the third implementation is preferable to the second since it does not require the weighing matrix and does not require finding a basis for $\mathcal{W}$. If n is small and m is large, the second implementation is preferable since it requires fewer filters.

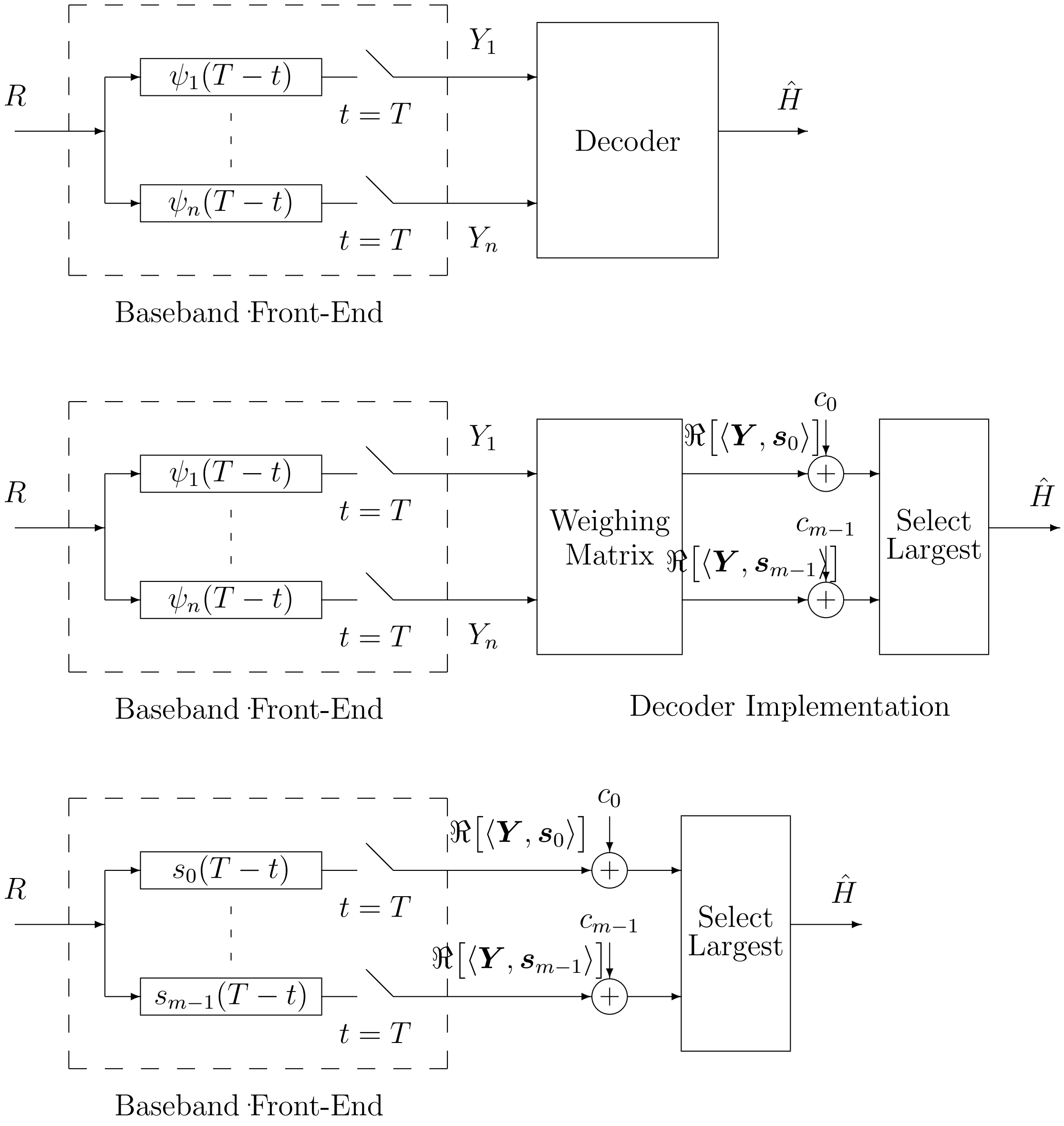


Figure 3.11: Three block diagrams of an optimal receiver for the waveform AWGN channel . Each baseband front end may alternatively be implemented via correlators.

3.5 Summary

In this chapter we have made the important transition from dealing with the discrete-time AWGN channels to the waveform AWGN channel. From a mathematical point of view we may summarize the essence as follows. Whatever we do, we send signals that are finite energy—hence in $\mathcal{L}_2$. We may see the collection of all possible signals as elements of an inner product space $\mathcal{W} \subset \mathcal{L}_2$ of some dimensionality n. The received signal consists of a component in $\mathcal{W}$ and one orthogonal to $\mathcal{W}$. The latter contains no signal component and can be removed by the receiver front end without loss of optimality. The elimination of the orthogonal component may be done by projecting the received signal onto $\mathcal{W}$. After we pick an orthonormal basis for $\mathcal{W}$, we can represent the transmitted signal and the projected received signal by means of n-tuples. Since the projected noise can also be represented as an n-tuple of i.i.d. zero-mean Gaussian random variables of variance $\sigma^2 = \frac{N_0}{2}$, the received n-tuple has the statistic of the output of a discrete-time AWGN channel that has the transmitter n-tuple as its input. An immediate consequence of this point of view is that there is no loss of generality in viewing a waveform sender and the corresponding maximum a posteriori (or maximum likelihood) receiver as being decomposed into the blocks of Figure 3.5. This implies that to design the decoder and to compute the error probability we can directly use what we have learned in Chapter 2 for the discrete-time AWGN channel.

Appendix 3.A Rectangle and Sinc as Fourier Transform Pairs

The Fourier transform of a rectangular pulse is a sinc pulse. Often one has to go back and forth between such Fourier pairs. The purpose of this appendix is to make it easier to figure out the details.

First of all let us recall that a function g and its Fourier transform $g_{\mathcal{F}}$ are related by

$$g(u) = \int g_{\mathcal{F}}(\alpha) \exp(j2\pi u\alpha) d\alpha$$
$$g_{\mathcal{F}}(v) = \int g(\alpha) \exp(-j2\pi v\alpha) d\alpha.$$

Notice that $g_{\mathcal{F}}(0)$ is the area under g and $g(0)$ is the area under $g_{\mathcal{F}}$.

Next let us recall that $\operatorname{sinc}(x) = \frac{\sin(\pi x)}{\pi x}$ is the function that equals 1 at $x = 0$ and equals 0 at all other integer values of x. Hence if $a, b \in \mathbb{R}$ are arbitrary constants, $a \operatorname{sinc}(bx)$ equals a at $x = 0$ and and equals 0 at nonzero multiples of $1/b$.

If you could remember that the area under $a \operatorname{sinc}(bx)$ is a/b then, from the two facts above, you could conclude that its Fourier transform, which you know is a rectangle, has height a/b and area a. Hence the width of this rectangle must be b.

It is actually easy to remember that the area under $a \operatorname{sinc}(bx)$ is a/b: it is the area of the triangle described by the main lobe of $a \operatorname{sinc}(bx)$, namely the area of the triangle with coordinates $(-1/b, 0)$, $(0, a)$, $(1/b, 0)$.

Appendix 3.B Problems

Problem 1. (Gram-Schmidt Procedure On Tuples) *Use the Gram-Schmidt orthonormalization procedure to find an orthonormal basis for the subspace spanned by the vectors* $\beta_1, \ldots, \beta_4$ *where* $\beta_1 = (1,0,1,1)^T$, $\beta_2 = (2,1,0,1)^T$, $\beta_3 = (1,0,1,-2)^T$, *and* $\beta_4 = (2,0,2,-1)^T$.

Problem 2. (Matched Filter Implementation)

In this problem, we consider the implementation of matched filter receivers. In particular, we consider Frequency Shift Keying (FSK) with the following signals:

$$s_j(t) = \begin{cases} \sqrt{\frac{2}{T}} \cos 2\pi \frac{n_j}{T} t, & \text{for } 0 \le t \le T, \\ 0, & \text{otherwise,} \end{cases} \tag{3.4}$$

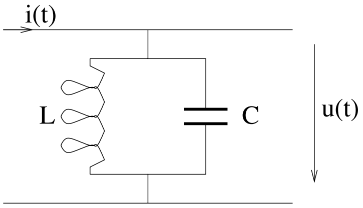
where $n_j \in \mathbb{Z}$ *and* $0 \le j \le m-1$. *Thus, the communications scheme consists of* m *signals* $s_j(t)$ *of different frequencies* $\frac{n_j}{T}$

(i) *Determine the impulse response* $h_j(t)$ *of the matched filter for the signal* $s_j(t)$. *Plot* $h_j(t)$.

(ii) *Sketch the matched filter receiver. How many matched filters are needed?*

(iii) *For* $-T \le t \le 3T$, *sketch the output of the matched filter with impulse response* $h_j(t)$ *when the input is* $s_j(t)$. *(*Hint: We recommend you to use Matlab.*)*

(iv) *Consider the following ideal resonance circuit:*



For this circuit, the voltage response to a unit impulse of current is

$$h(t) = \frac{1}{C} \cos \frac{t}{\sqrt{LC}}. \tag{3.5}$$

Show how this can be used to implement the matched filter for signal $s_j(t)$. *Determine how* L *and* C *should be chosen.* Hint: *Suppose that* $i(t) = s_j(t)$. *In that case, what is* $u(t)$*?*

Problem 3. (On-Off Signaling) *Consider the following binary hypothesis testing problem specified by:*

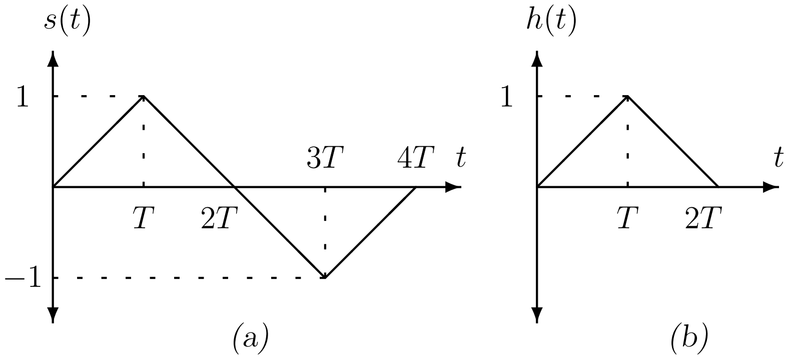
$$\begin{aligned} H = 0 &: \; Y(t) = s(t) + N(t) \\ H = 1 &: \; Y(t) = N(t) \end{aligned}$$

where $N(t)$ is AWGN (Additive White Gaussian Noise) of power spectral density $N_0/2$ and $s(t)$ is the signal shown in the Figure (a) below.

(a) Describe the maximum-likelihood receiver for the observable $Y(t)$, $t \in \mathbb{R}$.

(b) Determine the error probability for the receiver you described in (a)*.*

(c) Can you realize your receiver of part (a) *using a filter with impulse response $h(t)$ shown in Figure (a)?*



Problem 4. (Matched Filter Basics) *Let the transmitted signal be*

$$S(t) \;=\; \sum_{k=1}^{K} S_k \; h(t - kT)$$

where $S_i \in \{-1, 1\}$ and $h(t)$ is a given function. Assume that the function $h(t)$ and its shifts by multiples of T form an othonormal set, i.e.,

$$\int_{-\infty}^{\infty} h(t)h(t-kT)dt \;=\; \begin{cases} 0, & k \neq 0 \\ 1, & k = 0. \end{cases}$$

(a) Suppose $S(t)$ is filtered at the receiver by the matched filter with impulse response $h(-t)$. That is, the filtered waveform is $R(t) = \int_{-\infty}^{\infty} S(\tau)h(\tau - t)d\tau$. Show that the samples of this waveform at multiples of T are $R(mT) = S_m$, for $1 \leq m \leq K$.

(b) *Now suppose that the channel has an echo in it and behaves like a filter of impulse response* $f(t) = \delta(t) + \rho\delta(t-T)$, *where* ρ *is some constant between* -1 *and* 1. *Assume that the transmitted waveform* $S(t)$ *is filtered by* $f(t)$, *then filtered at the receiver by* $h(-t)$. *The resulting waveform* $\tilde{R}(t)$ *is again sampled at multiples of* T. *Determine the samples* $\tilde{R}(mT)$, *for* $1 \le m \le K$.

(c) *Suppose that the* k *th received sample is* $Y_k = S_k + \alpha S_{k-1} + Z_k$, *where* $Z_k \sim \mathcal{N}(0, \sigma^2)$ *and* $0 \le \alpha < 1$ *is a constant.* S_k *and* S_{k-1} *are independent random variables that take on the values* 1 *and* -1 *with equal probability. Suppose that the detector decides* $\hat{S}_k = 1$ *if* $Y_k > 0$, *and decides* $\hat{S}_k = -1$ *otherwise. Find the probability of error for this receiver.*

Problem 5. (Matched Filter Intuition) *In this problem, we develop some further intuition about matched filters. We have seen that an optimal receiver front end for the signal set* $\{s_j(t)\}_{j=0}^{m-1}$ *reduces the received (noisy) signal* $R(t)$ *to the* m *real numbers* $\langle R, s_j \rangle$, $j = 0, \ldots, m-1$. *We gain additional intuition about the operation* $\langle R, s_j \rangle$ *by considering*

$$R(t) \quad = \quad s(t) + N(t), \tag{3.6}$$

where $N(t)$ *is additive white Gaussian noise of power spectral density* $N_0/2$ *and* $s(t)$ *is an arbitrary but fixed signal. Let* $h(t)$ *be an arbitrary waveform, and consider the receiver operation*

$$Y \quad = \quad \langle R, h \rangle = \langle s, h \rangle + \langle N, h \rangle. \tag{3.7}$$

The signal-to-noise ratio (SNR) is thus

$$SNR \quad = \quad \frac{|\langle s, h \rangle|^2}{E\left[|\langle N, h \rangle|^2\right]}. \tag{3.8}$$

Notice that the SNR is not changed when $h(t)$ *is multiplied by a constant. Therefore, we assume that* $h(t)$ *is a unit energy signal and denote it by* $\phi(t)$. *Then,*

$$E\left[|\langle N, \phi \rangle|^2\right] \quad = \quad \frac{N_0}{2}. \tag{3.9}$$

(a) *Use Cauchy-Schwarz inequality to give an upper bound on the SNR. What is the condition for equality in the Cauchy-Schwarz inequality? Find the* $\phi(t)$ *that maximizes the SNR. What is the relationship between the maximizing* $\phi(t)$ *and the signal* $s(t)$ *?*

(b) *Let* $s = (s_1, s_2)^T$ *and use calculus (instead of the Cauchy-Schwarz inequality) to find the* $\phi = (\phi_1, \phi_2)^T$ *that maximizes* $\langle s, \phi \rangle$ *subject to the constraint that* ϕ *has unit energy.*

(c) *Hence to maximize the SNR, for each value of* t *we have to weigh (multiply)* $R(t)$ *with* $s(t)$ *and then integrate. Verify with a picture (convolution) that the output at time* T *of a filter with input* $s(t)$ *and impulse response* $h(t) = s(T-t)$ *is indeed* $\int_0^T s^2(t)dt$.

(d) *We may also look at the situation in terms of Fourier transforms. Write out the filter operation in the frequency domain.*

PROBLEM 6. (Receiver for Non-White Gaussian Noise) *We consider the receiver design problem for signals used in non-white additive Gaussian noise. That is, we are given a set of signals* $\{s_j(t)\}_{j=0}^{m-1}$ *as usual, but the noise added to those signals is no longer white; rather, it is a Gaussian stochastic process with a given power spectral density*

$$S_N(f) \;=\; G^2(f), \tag{3.10}$$

where we assume that $G(f) \neq 0$ *inside the bandwidth of the signal set* $\{s_j(t)\}_{j=0}^{m-1}$. *The problem is to design the receiver that minimizes the probability of error.*

(a) *Find a way to transform the above problem into one that you can solve, and derive the optimum receiver.*

(b) *Suppose there is an interval* $[f_0, f_0 + \Delta]$ *inside the bandwidth of the signal set* $\{s_j(t)\}_{j=0}^{m-1}$ *for which* $G(f) = 0$. *What do you do? Describe in words.*

PROBLEM 7. (Antipodal Signaling in Non-White Gaussian Noise) *In this problem, antipodal signaling (i.e.* $s_0(t) = -s_1(t)$ *) is to be used in non-white additive Gaussian noise of power spectral density*

$$S_N(f) \;=\; G^2(f), \tag{3.11}$$

where we assume that $G(f) \neq 0$ *inside the bandwidth of the signal* $s(t)$. *How should the signal* $s(t)$ *be chosen (as a function of* $G(f)$ *) such as to minimize the probability of error?* Hint: *For ML decoding of antipodal signaling in AWGN (of fixed variance), the* $Pr\{e\}$ *depends only on the signal energy.*

PROBLEM 8. (Mismatched Receiver) *Let the received waveform* $Y(t)$ *be given by*

$$Y(t) \;=\; c\,X\,s(t) + N(t), \tag{3.12}$$

where $c > 0$ *is some deterministic constant,* X *is a uniformly distributed random variable that takes values in* $\{3, 1, -1, -3\}$, $s(t)$ *is the deterministic waveform*

$$s(t) \;=\; \begin{cases} 1, & \text{if } 0 \leq t < 1 \\ 0, & \text{otherwise,} \end{cases} \tag{3.13}$$

and $N(t)$ *is white Gaussian noise of spectral density* $\frac{N_0}{2}$.

(a) *Describe the receiver that, based on the received waveform* $Y(t)$, *decides on the value of* X *with least probability of error. Be sure to indicate precisely when your decision rule would declare "+3", "+1", "−1", and "−3".*

(b) *Find the probability of error of the detector you have found in Part* (a).

(c) *Suppose now that you still use the detector you have found in Part* (a), *but that the received waveform is actually*

$$Y(t) \quad = \quad \frac{3}{4} c\, X\, s(t) + N(t), \tag{3.14}$$

i.e., you were mis-informed about the signal amplitude. What is the probability of error now?

(d) *Suppose now that you still use the detector you have found in Part* (a) *and that* $Y(t)$ *is according to Equation (3.12), but that the noise is* colored. *In fact,* $N(t)$ *is a zero-mean stationary Gaussian noise process of auto-covariance function*

$$K_N(\tau) \quad = \quad E[N(t)N(t+\tau)] = \frac{1}{4\alpha} e^{-|\tau|/\alpha}, \tag{3.15}$$

where $0 < \alpha < \infty$ *is some deterministic real parameter. What is the probability of error now?*

Problem 9. (QAM Receiver) *Consider a transmitter which transmits waveforms of the form,*

$$s(t) \quad = \quad \begin{cases} s_1\sqrt{\frac{2}{T}}\cos 2\pi f_c t + s_2\sqrt{\frac{2}{T}}\sin 2\pi f_c t, & \text{for } 0 \le t \le T, \\ 0, & \text{otherwise,} \end{cases} \tag{3.16}$$

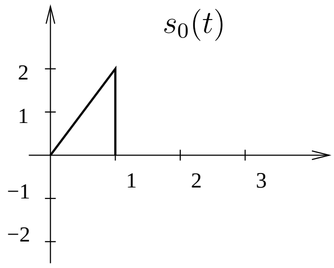
where $2f_cT \in \mathbb{Z}$ *and* $(s_1, s_2) \in \{(\sqrt{E}, \sqrt{E}), (-\sqrt{E}, \sqrt{E}), (-\sqrt{E}, -\sqrt{E}), (\sqrt{E}, -\sqrt{E})\}$ *with equal probability. The signal received at the receiver is corrupted by AWGN of power spectral density* $\frac{N_0}{2}$.

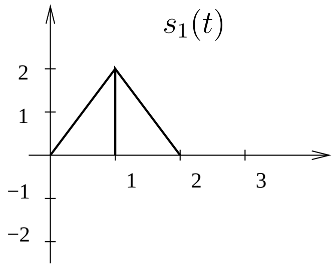
(a) *Specify the receiver for this transmission scheme.*

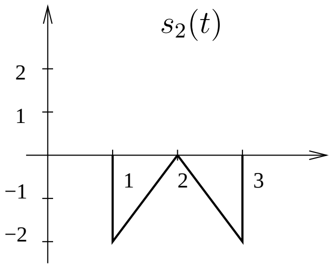
(b) *Draw the decoding regions and find the probability of error.*

Problem 10. (Gram-Schmidt Procedure on Waveforms: 1) *Consider the following functions* $s_0(t)$, $s_1(t)$ *and* $s_2(t)$.

(a) *Using the Gram-Schmidt procedure, determine a basis of the space spanned by* $\{s_0(t), s_1(t), s_2(t)\}$. *Denote the basis functions by* $\phi_0(t)$, $\phi_1(t)$ *and* $\phi_2(t)$.





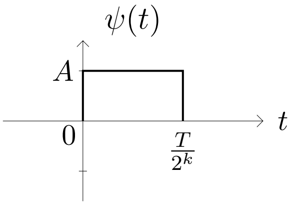


(b) *Let* $\boldsymbol{v}_1 = (3, -1, 1)^T$ *and* $\boldsymbol{v}_2 = (-1, 2, 3)^T$ *be two points in the space spanned by* $\{\phi_0(t), \phi_1(t), \phi_2(t)\}$. *What is their corresponding signal,* $v_1(t)$ *and* $v_2(t)$? *(You can simply draw a detailed graph.)*

(c) *Compute* $\int v_1(t)v_2(t)dt$.

PROBLEM 11. (Signaling Scheme Example) *Consider the following communication chain. We have* 2^k *possible hypotheses with* $k \in \mathbb{N}$ *to convey through a waveform channel. When hypothesis* i *is selected, the transmitted signal is* $s_i(t)$ *and the received signal is given by* $R(t) = s_i(t) + N(t)$, *where* $N(t)$ *denotes white Gaussian noise with double-sided power spectral density* $\frac{N_0}{2}$. *Assume that the transmitter uses the position of a pulse* $\psi(t)$ *in an interval* $[0, T]$, *in order to convey the desired hypothesis, i.e., to send hypothesis* i, *the transmitter sends the signal* $\psi_i(t) = \psi(t - \frac{iT}{2^k})$.

(a) *If the pulse is given by the waveform* $\psi(t)$ *depicted below. What is the value of* A *that gives us signals of energy equal to one as a function of* k *and* T?



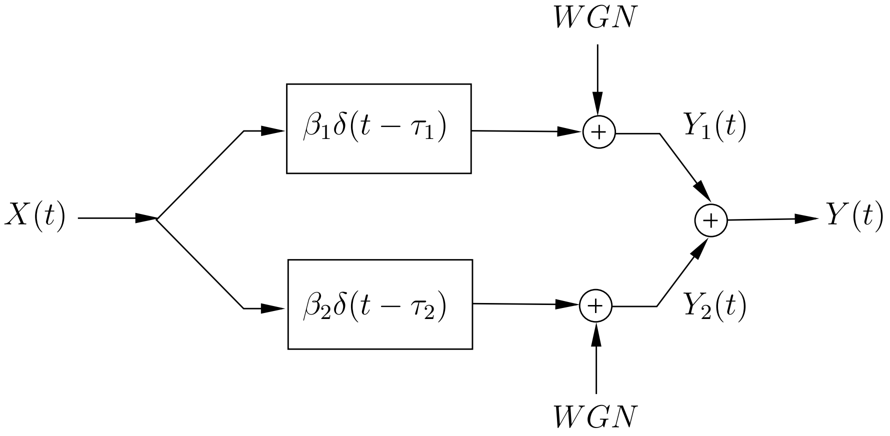
(b) *We want to transmit the hypothesis* $i = 3$ *followed by the hypothesis* $j = 2^k - 1$. *Plot the waveform you will see at the output of the transmitter, using the pulse given in the previous question.*

(c) *Sketch the optimal receiver.*
What is the minimum number of filters you need for the optimal receiver? Explain.

(d) *What is the major drawback of this signaling scheme? Explain.*

PROBLEM 12. (Two Receive Antennas) *Consider the following communication chain, where we have two possible hypotheses, $H \in \{0, 1\}$. Assume that $P_H(0) = P_H(1) = \frac{1}{2}$. The transmitter uses antipodal signaling. To transmit $H = 0$, the transmitter sends a unit energy pulse $p(t)$, and to transmit $H = 1$, it sends $-p(t)$. That is, the transmitted signal is $X(t) = \pm p(t)$. The observation consists of $Y_1(t)$ and $Y_2(t)$ as shown below. The signal along each "path" is an attenuated and delayed version of the transmitted signal $X(t)$. The noise is additive white Gaussian with double sided power spectral density $N_0/2$. Also, the noise added to the two observations is independent and independent of the data. The goal of the receiver is to decide which hypothesis was transmitted, based on its observation.*

We will look at two different scenarios: either the receiver has access to each individual signal $Y_1(t)$ and $Y_2(t)$, or the receiver has only access to the combined *observation $Y(t) = Y_1(t) + Y_2(t)$.*



a. *The case where the receiver has only access to the combined output $Y(t)$.*

1. *In this case, observe that we can write the received waveform as $\pm g(t) + Z(t)$. What are $g(t)$ and $Z(t)$ and what are the statistical properties of $Z(t)$?* Hint: Recall that $\int \delta(\tau - \tau_1)p(t - \tau)d\tau = p(t - \tau_1)$.
2. *What is the optimal receiver for this case? Your answer can be in the form of a block diagram that shows how to process $Y(t)$ or in the form of equations. In either case, specify how the decision is made between $\hat{H} = 0$ or $\hat{H} = 1$.*
3. *Assume that $\int p(t-\tau_1)p(t-\tau_2)dt = \gamma$, where $-1 \leq \gamma \leq 1$. Find the probability of error for this optimal receiver, express it in terms of the Q function, β_1, β_2, γ and $N_0/2$.*

b. *The case where the receiver has access to the individual observations $Y_1(t)$ and $Y_2(t)$.*

1. *Argue that the performance of the optimal receiver for this case can be no worse than that of the optimal receiver for part (a).*

2. *Compute the sufficient statistics* (Y_1, Y_2), *where* $Y_1 = \int Y_1(t)p(t-\tau_1)dt$ *and* $Y_2 = \int Y_2(t)p(t-\tau_2)dt$. *Show that this sufficient statistic* (Y_1, Y_2) *has the form* $(Y_1, Y_2) = (\beta_1 + Z_1, \beta_2 + Z_2)$ *under* $H = 0$, *and* $(-\beta_1 + Z_1, -\beta_2 + Z_2)$ *under* $H = 1$, *where* Z_1 *and* Z_2 *are independent zero-mean Gaussian random variables of variance* $N_0/2$.

3. *Using the LLR (Log-Likelihood Ratio), find the optimum decision rule for this case.* Hint: *It may help to draw the two hypotheses as points in* $\mathbb{R}^2$. *If we let* $V = (V_1, V_2)$ *be a Gaussian random vector of mean* $m = (m_1, m_2)$ *and covariance matrix* $\Sigma = \sigma^2 I$, *then its pdf is* $p_V(v_1, v_2) = \frac{1}{2\pi\sigma^2}\exp\left(-\frac{(v_1-m_1)^2}{2\sigma^2} - \frac{(v_2-m_2)^2}{2\sigma^2}\right)$.

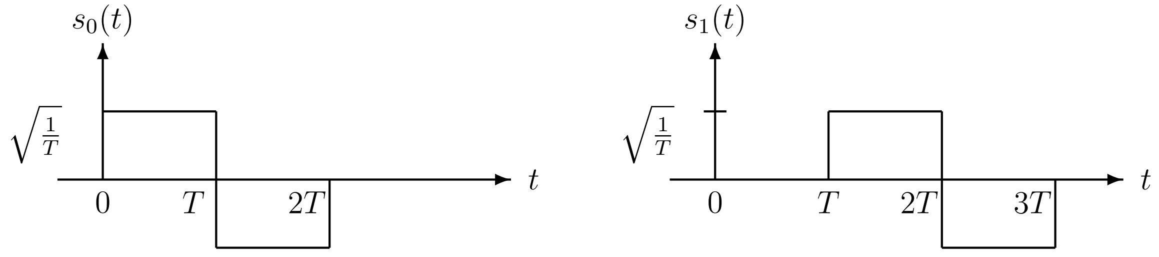
4. *What is the optimal receiver for this case? Your answer can be in the form of a block diagram that shows how to process* $Y_1(t)$ *and* $Y_2(t)$ *or in the form of equations. In either case, specify how the decision is made between* $\hat{H} = 0$ *or* $\hat{H} = 1$.

5. *Find the probability of error for this optimal receiver, express it in terms of the* Q *function,* β_1, β_2 *and* N_0.

c. *Comparison of the two cases*

1. *In the case of* $\beta_2 = 0$, *that is the second observation is solely noise, give the probability of error for both cases (a) and (b). What is the difference between them? Explain why.*

Problem 13. (Delayed Signals) *One of two signals shown in the figure below is transmitted over the additive white Gaussian noise channel. There is no bandwidth constraint and either signal is selected with probability* $1/2$.



(a) *Draw a block diagram of a maximum likelihood receiver. Be as specific as you can. Try to use the smallest possible number of filters and/or correlators.*

(b) *Determine the error probability in terms of the* Q*-function, assuming that the power spectral density of the noise is* $\frac{N_0}{2} = 5\ [\frac{W}{Hz}]$.

Problem 14. (Antenna Array) *Consider an L-element antenna array as shown in the figure below.*

Let $u(t)\beta_i$ be a complex-valued signal transmitted at antenna element i, $i = 1, 2, \ldots, L$ (according to some indexing which is irrelevant here) and let

$$v(t) = \sum_{i=1}^{L} u(t - \tau_D)\beta_i\alpha_i$$

(plus noise) be the sum-signal at the receiver antenna, where α_i is the path strength for the signal transmitted at antenna element i and τ_D is the (common) path delay.

(a) *Choose the vector $\beta = (\beta_1, \beta_2, \ldots, \beta_L)^T$ that maximizes the signal energy at the receiver, subject to the constraint $\|\beta\| = 1$. The signal energy is defined as $E_v = \int |v(t)|^2 dt$. Hint Use the Cauchy-Schwarz inequality: for any two vectors $\mathbf{a}$ and $\mathbf{b}$ in $\mathbb{C}^n$, $|\langle \mathbf{a}, \mathbf{b} \rangle|^2 \leq \|\mathbf{a}\|^2\|\mathbf{b}\|^2$ with equality iff $\mathbf{a}$ and $\mathbf{b}$ are linearly dependent.*

(b) *Let $u(t) = \sqrt{E_u}\phi(t)$ where $\phi(t)$ has unit energy. Determine the received signal power as a function of L when β is selected as in (a) and $\alpha = (\alpha, \alpha, \ldots, \alpha)^T$ for some complex number α.*

(c) *In the above problem the received energy grows monotonically with L while the transmit energy is constant. Does this violate energy conservation or some other fundamental low of physics? Hint: an antenna array is not an isotropic antenna (i.e. an antenna that sends the same energy in all directions).*

Problem 15. (Cioffi) *The signal set*

$$\begin{aligned} s_0(t) &= \operatorname{sinc}^2(t) \\ s_1(t) &= \sqrt{2}\operatorname{sinc}^2(t)\cos(4\pi t) \end{aligned}$$

is used to communicate across an AWGN channel of power spectral density $\frac{N_0}{2}$.

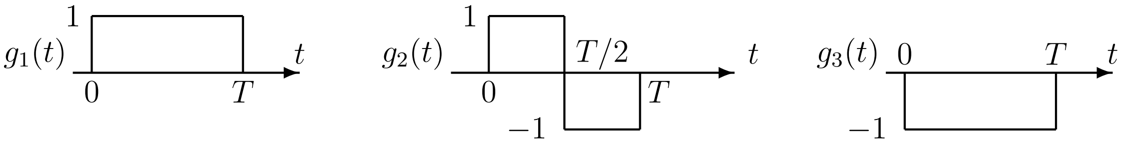
(a) *Find the Fourier transforms of the above signals and plot them.*

(b) *Sketch a block diagram of a ML receiver for the above signal set.*

(c) *Determine its error probability of your receiver assuming that $s_0(t)$ and $s_1(t)$ are equally likely.*

(d) If you keep the same receiver, but use $s_0(t)$ with probability $\frac{1}{3}$ and $s_1(t)$ with probability $\frac{2}{3}$, does the error probability increase, decrease, or remain the same? Justify your answer.

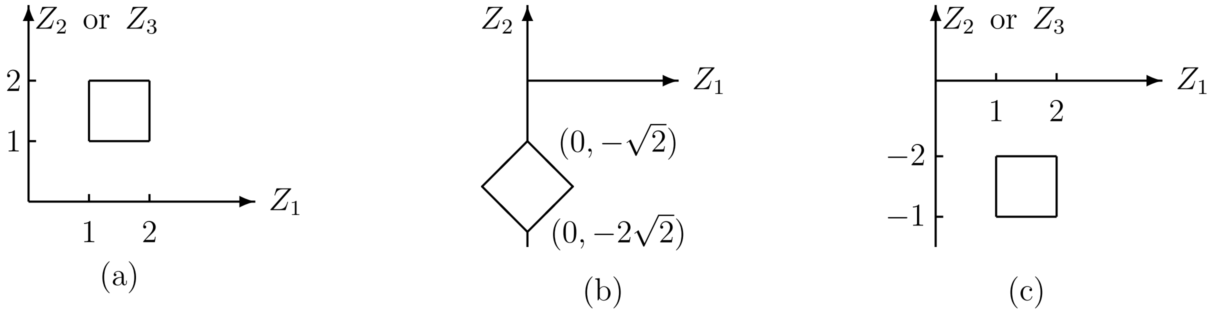
PROBLEM 16. (Sample Exam Question) *Let $N(t)$ be a zero-mean white Gaussian process of power spectral density $\frac{N_0}{2}$. Let $g_1(t)$, $g_2(t)$, and $g_3(t)$ be waveforms as shown in the following figure.*



(a) Determine the norm $\|g_i\|$, $i = 1, 2, 3$.

(b) Let Z_i be the projection of $N(t)$ onto $g_i(t)$. Write down the mathematical expression that describes this projection, i.e. how you obtain Z_i from $N(t)$ and $g_i(t)$.

(c) Describe the object Z_1, i.e. tell us everything you can say about it. Be as concise as you can.



(d) Are Z_1 and Z_2 independent? Justify your answer.

(e) (i) Describe the object $\boldsymbol{Z} = (Z_1, Z_2)$. (We are interested in what it is, not on how it is obtained.)

(ii) Find the probability P_a that $\boldsymbol{Z}$ lies in the square labeled (a) in the figure below.

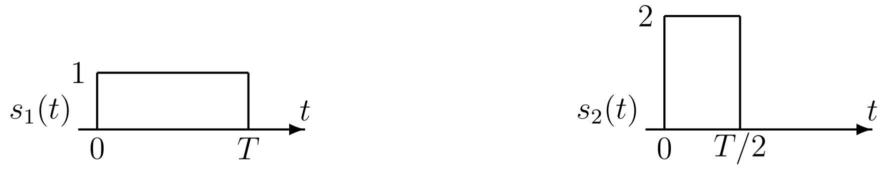
(iii) Find the probability P_b that $\boldsymbol{Z}$ lies in the square (b) of the same figure. Justify your answer.

(f) (i) Describe the object $\boldsymbol{W} = (Z_1, Z_3)$.

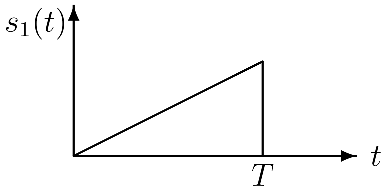
(ii) Find the probability Q_a that $\boldsymbol{W}$ lies in the square (a).

(iii) Find the probability Q_c that $\boldsymbol{W}$ lies in the square (c).

Problem 17. (Gram-Schmidt Procedure on Waveforms: 2) *Use the Gram Schmidt procedure to find an orthonormal basis for the vector space spanned by the functions shown below.*



Problem 18. (ML Receiver With Single Causal Filter) *You want to design a Maximum Likelihood (ML) receiver for a system that communicates an equiprobable binary hypothesis by means of the signals $s_1(t)$ and $s_2(t) = s_1(t - T_d)$, where $s_1(t)$ is shown in the figure and T_d is a fixed number assumed to be known at the receiver. The channel is the usual AWGN channel with noise power spectral density $N_0/2$. At the receiver front end you are allowed to use a single causal filter of impulse response $h(t)$ (*A causal filter is one whose impulse response is 0 for $t < 0$*).*



(a) Describe the $h(t)$ that you chose for your receiver.

(b) Sketch a block diagram of your receiver. Be specific about the sampling times.

(c) Assuming that $T_d > T$, determine the error probability for the receiver as a function of N_0 and E_s ($E_s = ||s_1(t)||^2$).

Problem 19. (Waveform Receiver)

Consider the signals $s_0(t)$ and $s_1(t)$ shown in the figure.

(a) Determine an orthonormal basis $\{\psi_0(t), \psi_1(t)\}$ for the space spanned by $\{s_0(t), s_1(t)\}$ and find the n-tuples of coefficients $\mathbf{s}_0$ and $\mathbf{s}_1$ that correspond to $s_0(t)$ and $s_1(t)$, respectively.

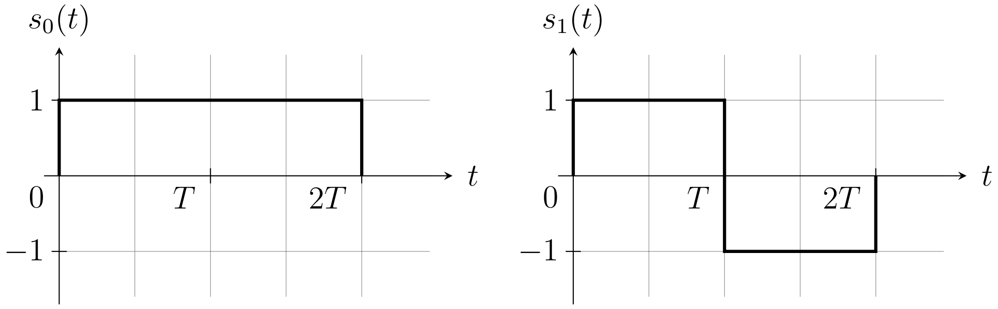


Figure 3.12: Signal waveforms

(b) *Let X be a uniformly distributed binary random variable that takes values in $\{0,1\}$. We want to communicate the value of X over an additive white Gaussian noise channel. When $X=0$, we send $S(t)=s_0(t)$, and when $X=1$, we send $S(t)=s_1(t)$. The received signal at the destination is*

$$Y(t)=S(t)+Z(t),$$

where $Z(t)$ is AWGN of power spectral density $\frac{N_0}{2}$.

(i) *Draw an optimal matched filter receiver for this case. Specifically say how the decision is made.*

(ii) *What is the output of the matched filter(s) when $X=0$ and the* noise variance is zero *($\frac{N_0}{2}=0$)?*

(iii) *Describe the output of the matched filter when $S(t)=0$ and the noise variance is $\frac{N_0}{2}>0$.*

(c) *Plot the $\mathbf{s}_0$ and $\mathbf{s}_1$ that you have found in part (**??**), and determine the error probability P_e of this scheme as a function of T and N_0.*

(d) *Find a suitable waveform $v(t)$, such that the new signals $\hat{s}_0(t)=s_0(t)-v(t)$ and $\hat{s}_1(t)=s_1(t)-v(t)$ have minimal energy and plot the resulting $\hat{s}_0(t)$ and $\hat{s}_1(t)$. Hint: you may first want to find $\mathbf{v}$, the n-tuple of coefficients that corresponds to $v(t)$.*

(e) *Compare $\hat{s}_0(t)$ and $\hat{s}_1(t)$ to $s_0(t)$ and $s_1(t)$, respectively, and comment on the part $v(t)$ that has been removed.*

Chapter 4

Signal Design Trade-Offs

4.1 Introduction

It is time to shift our focus to the transmitter and take a look at some of the options we have in terms of choosing the signal constellation. The goal of this chapter is to build up some intuition about the impact that those options have on the transmission rate, bandwidth, power, and error probability. Throughout this chapter we assume that the channel is the AWGN channel and that the receiver implements a MAP decision rule. Initially we will assume that all signals are used with the same probability in which case the MAP rule is a ML rule.

To put things into perspective, we mention from the outset that the problem of choosing a convenient signal constellation is not as clean-cut as the receiver design problem that has kept us busy until now. The reason is that the receiver design problem has a clear objective, namely to minimize the error probability, and an essentially unique solution, a MAP decision rule. In contrast, choosing a good signal constellation is making a tradeoff among conflicting objectives. Specifically, if we could we would choose a signal constellation that contains a very large number m of signals of very small duration T and very small bandwidth B. By making m sufficiently large and BT sufficiently small we could achieve any arbitrarily large communication rate $\frac{\log_2 m}{TB}$ (expressed in bits per second per Hz). In addition, if we could we would choose our signals so that they use very little energy (what about zero) and result in a very small error probability (why not zero). These are conflicting goals.

While we have already mentioned a few times that when we transmit a signal chosen from a constellation of m signals we are in essence transmitting the equivalent of $k = \log_2 m$ bits, we clarify this concept since it is essential for the sequel. So far we have implicitly considered *one-shot communication*, i.e., we have considered the problem of sending a single message in isolation. In practice we send several messages by using the same idea over and over. Specifically, if in the one-shot setting the message $H = i$ is mapped into the signal s_i, then for a sequence of messages $H_0, H_1, H_2, \dots = i_0, i_1, i_2 \dots$ we send $s_{i_0}(t)$ followed by $s_{i_1}(t-T_m)$ followed by $s_{i_2}(t-2T_m)$ etc, where T_m (m for message) is typical the smallest amount of time we have to wait to make $s_i(t)$ and $s_j(t-T_m)$ orthogonal for all i and j in $\{0,1,\dots,m-1\}$. Assuming that the probability of error P_e is negligible, the system consisting of the sender, the channel, and the receiver is equivalent to a pipe that carries m-ary symbols at a rate of $1/T_m$ [symbol/sec]. (Whether we call them messages or symbols is irrelevant. In single-shot transmission it makes sense to speak of a message being sent whereas in repeated transmissions it is natural to consider the message as being the whole sequence and individual components of the message as being symbols that take value in an m-letter alphabet.) It would be a significant restriction if this virtual pipe could be used only with sources that produce m-ary sequences. Fortunately this is not the case. To facilitate the discussion, assume that m is a power of 2, i.e., $m = 2^k$ for some integer k. Now if the source produces a binary sequence, the sender and the receiver can agree on a one-to-one map between the set $\{0,1\}^k$ and the set of messages $\{0,1,\dots,m-1\}$. This allows us to map every k bits of the source sequence into an m-ary symbol. The resulting transmission rate is $k/T_m = \log_2 m/T_m$ bits per second. The key is once again that with an m-ary alphabet each letter is equivalent to $\log_2 m$ bits.

The chapter is organized as follows. First we consider transformations that may be applied to a signal constellation without affecting the resulting probability of error. One such transformation consists of translating the entire signal constellation and we will see how to choose the translation to minimize the resulting average energy. We may picture the translation as being applied to the constellation of n-tuples that describes the original waveform constellation with respect to a fixed orthonormal basis. Such a translation is a special case of an isometry in $\mathbb{R}^n$ and any such isometry applied to a constellation of n-tuples will lead to a constellation that has the exact same error probability as the original (assuming the AWGN channel). A transformation that also keeps the error probability unchanged but can have more dramatic consequences on the time and frequency properties consists of keeping the original n-tuple constellation and changing the orthonormal basis. The transformation is also an isometry but this time applied directly to the waveform constellation rather than to the n-tuple constellation. Even though we did not emphasize this point of view, implicitly we did exactly this in Example 51. Such transformations allow us to vary the duration and/or the bandwidth occupied by the process produced by the transmitter. This raises the question about the possible time/bandwidth tradeoffs. The question is studied in Subsection 4.3. The chapter concludes with a number of representative examples of signal constellations intended to sharpen our intuition about the available tradeoffs.

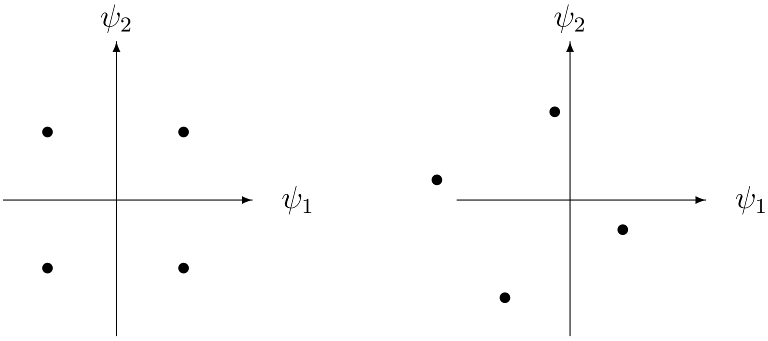
4.2 Isometric Transformations

An isometry in $\mathcal{L}_2$ (also called rigid motion) is a distance-preserving transformation $a : \mathcal{L}_2 \to \mathcal{L}_2$. Hence for any two vectors p, q in $\mathcal{L}_2$, the distance from p to q equals the distance from $a(p)$ to $a(q)$. Isometries can be defined in a similar way over a subspace $\mathcal{W}$ of $\mathcal{L}_2$ as well as over $\mathbb{R}^n$. In fact, once we fix an orthonormal basis for an n-dimensional subspace $\mathcal{W}$ of $\mathcal{L}_2$, any isometry of $\mathcal{W}$ corresponds to an isometry of $\mathbb{R}^n$ and vice-versa. Alternatively, there are isometries of $\mathcal{L}_2$ that map a subspace $\mathcal{W}$ to a different subspace $\mathcal{W}'$. We consider both, isometries within $\mathcal{W}$ and those from $\mathcal{W}$ to $\mathcal{W}'$.

4.2.1 Isometric Transformations within a subspace $\mathcal{W}$

We assume that we have a constellation $\mathcal{S}$ of waveforms that spans an n-dimensional subspace $\mathcal{W}$ of $\mathcal{L}_2$ and that we have fixed an orthonormal basis $\mathcal{B}$ for $\mathcal{W}$. The waveform constellation $\mathcal{S}$ and the orthonormal basis $\mathcal{B}$ lead to a corresponding n-tuple constellation $\boldsymbol{\mathcal{S}}$. If we apply an isometry of $\mathbb{R}^n$ to $\boldsymbol{\mathcal{S}}$ we obtain an n-tuple constellation $\boldsymbol{\mathcal{S}}'$ and the corresponding waveform constellation $\mathcal{S}'$. From the way we compute the error probability it should be clear that when the channel is the AWGN, the probability of error associated to $\boldsymbol{\mathcal{S}}$ is identical to that associated to $\boldsymbol{\mathcal{S}}'$. A proof of this rather intuitive fact is given in Appendix 4.A.

Example 52. *The composition of a translation and a rotation is an isometry. The figure below shows an original signal set and a translated and rotated copy. The probability of error is the same for both. The average energy is in general not the same.*



□

In the next subsection we see how to translate a constellation so as to minimize the average energy.

4.2.2 Energy-Minimizing Translation

Let $\boldsymbol{Y}$ be a zero-mean random vector in $\mathbb{R}^n$. It is immediate to verify that for any $\boldsymbol{b} \in \mathbb{R}^n$,

$$E\|\boldsymbol{Y} - \boldsymbol{b}\|^2 = E\|\boldsymbol{Y}\|^2 + \|\boldsymbol{b}\|^2 - 2E\langle \boldsymbol{Y}, \boldsymbol{b}\rangle = E\|\boldsymbol{Y}\|^2 + \|\boldsymbol{b}\|^2 \geq E\|\boldsymbol{Y}\|^2$$

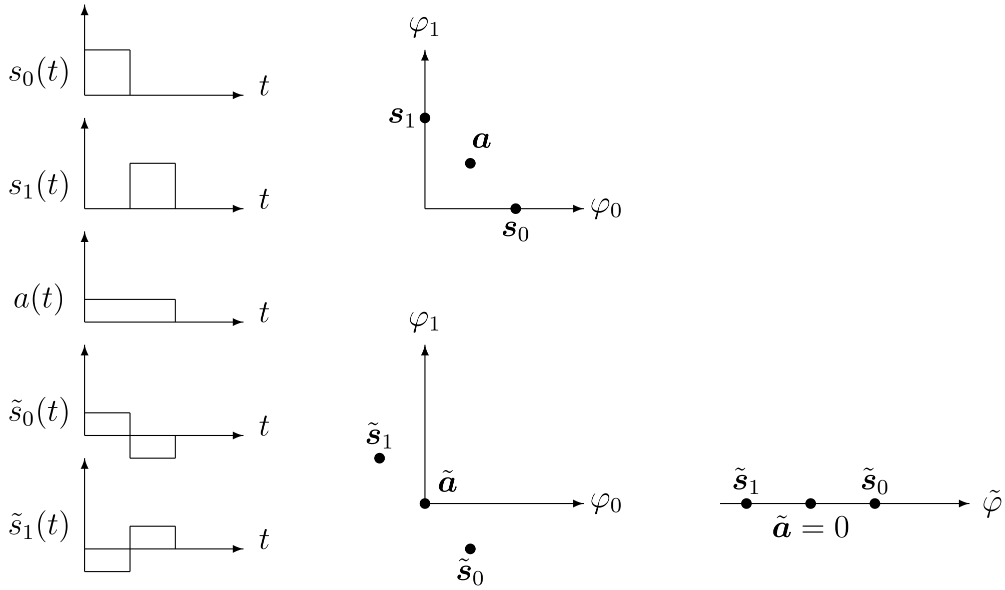


Figure 4.1: Example of isometric transformation to minimize the energy.

with equality iff $\boldsymbol{b} = 0$. This says that the expected squared norm is minimized when the random vector is zero-mean. Hence for a generic random vector $\boldsymbol{S} \in \mathbb{R}^n$ (not necessarily zero-mean), the translation vector $\boldsymbol{b} \in \mathbb{R}^n$ that minimizes the expected squared norm of $\boldsymbol{S} - \boldsymbol{b}$ is the mean $\boldsymbol{m} = E[\boldsymbol{S}]$.

The average energy $\mathcal{E}$ of a signal constellation $\{\boldsymbol{s}_0, \boldsymbol{s}_1, \dots, \boldsymbol{s}_{m-1}\}$ is defined as

$$\mathcal{E} = \sum_i P_H(i)\|\boldsymbol{s}_i\|^2.$$

Hence $\mathcal{E} = E\|\boldsymbol{S}\|^2$, where $\boldsymbol{S}$ is the random vector that takes value $\boldsymbol{s}_i$ with probability $P_H(i)$. The result of the previous paragraph says that we can reduce the energy (without affecting the error probability) by using the translated constellation $\{\boldsymbol{s}'_0, \boldsymbol{s}'_1, \dots, \boldsymbol{s}'_{m-1}\}$, where $\boldsymbol{s}'_i = \boldsymbol{s}_i - \boldsymbol{m}$, with

$$\boldsymbol{m} = \sum_i P_H(i)\boldsymbol{s}_i.$$

Example 53. *Let $s_0(t)$ and $s_1(t)$ be rectangular pulses with support $[0, T]$ and $[T, 2T]$, respectively, as shown on the left of Figure 4.1. Assuming that $P_H(0) = P_H(1) = \frac{1}{2}$, we calculate the centroid $a(t) = \frac{1}{2}s_0(t) + \frac{1}{2}s_1(t)$ and see that it is non-zero. Hence we can save energy by using instead $\tilde{s}_i(t) = s_i(t) - a(t)$, $i = 0, 1$. The result are two antipodal signals (see again the figure). On the right of Figure 4.1 we see the equivalent representation in the signal space, where φ_0 and φ_1 form an orthonormal basis for the 2-dimensional space spanned by s_0 and s_1 and $\tilde{\varphi}$ forms an orthonormal basis for the 1-dimensional space spanned by $\tilde{s}_0$ and $\tilde{s}_1$*

□

4.2.3 Isometric Transformations from $\mathcal{W}$ to $\mathcal{W}'$

Assume again a constellation $\mathcal{S}$ of waveforms that spans an n-dimensional subspace $\mathcal{W}$ of $\mathcal{L}_2$, an orthonormal basis $\mathcal{B}$ for $\mathcal{W}$, and the associated n-tuple constellation $\boldsymbol{\mathcal{S}}$. Let $\mathcal{B}'$ be the orthonormal basis of another n-dimensional subspace of $\mathcal{L}_2$. Together $\boldsymbol{\mathcal{S}}$ and $\mathcal{B}'$ specify a constellation $\mathcal{S}'$ that spans $\mathcal{W}'$. It is easy to see that corresponding vectors are related by an isometry. Indeed, if p maps into p' and q into q' then $\|p-q\| = \|\boldsymbol{p}-\boldsymbol{q}\| = \|p'-q'\|$. Once again, an example of this sort of transformation is implicit in Example 51. Notice that some of those constellations have finite support in the time domain and some have finite support in the frequency domain. Are we able to choose the duration T and the bandwidth B at will? That would be quite nice. Recall that the relevant parameters associated to a signal constellation are the average energy $\mathcal{E}$, the error probability P_e, the number k of bits carried by a signal (equivalently the size $m = 2^k$ of the signal constellation), and the time-bandwidth-product BT where for now B is informally defined as the frequency interval that contains most of the signal's energy and T as the time interval that contains most of the signal's energy. The ratio k/BT is the number of bits per second per Hz of bandwidth carried in average by a signal. (In this informal discussion the underlying assumption is that signals are correctly decoded at the receiver. If the signals are not correctly decoded then we can not claim that k bits of information are conveyed every time that we send a signal.) The class of transformations described in this subsection has no effect on the average energy, on the error probability, and on the number of bits carried by a signal. Hence a question of considerable practical interest is that of finding the transformation that minimizes BT for a fixed n. In the next section we take a look at the largest possible value of BT.

4.3 Time Bandwidth Product Vs Dimensionality

The goal of this section is to establish a relationship between n and BT. The reader may be able to see already that n can be made to grow at least as fast as linearly with BT (two examples will follow) but can it grow faster and, if not, what is the constant in front of BT?

Fist we need to define B and T rigorously. We are tempted to define the bandwidth of a *baseband* signal $s(t)$ to be B if the support of $s_{\mathcal{F}}(t)$ is $[-\frac{B}{2}, \frac{B}{2}]$. This definition is not useful in practice since all man-made signals $s(t)$ have finite support (in the time domain) and thus $s_{\mathcal{F}}(f)$ has infinite support.[1] A better definition of bandwidth for a baseband signal (but not the only one that makes sense) is to fix a number $\eta \in (0,1)$ and say that the baseband signal $s(t)$ has bandwidth B if B is the smallest number such that

$$\int_{-\frac{B}{2}}^{\frac{B}{2}} |s_{\mathcal{F}}(f)|^2 df = \|s\|^2(1-\eta).$$

[1] We define the support of a real or complex valued function $x : \mathcal{A} \to \mathcal{B}$ as the smallest interval $\mathcal{C} \subseteq \mathcal{A}$ such that $x(c) = 0$, for all $c \notin \mathcal{C}$.

In words, the baseband signal has bandwidth B if $[-\frac{B}{2}, \frac{B}{2}]$ is the smallest interval that contains *at least* $100(1-\eta)\%$ of the signal's power. The bandwidth changes if we change η . Reasonable values for η are $\eta = 0.1$ and $\eta = 0.01$. This definition has the property that allows us to relate time, bandwidth, and dimensionality in a rigorous way. If we let $\eta = \frac{1}{12}$ and define

$$\mathcal{L}_2(T_a, T_b, B_a, B_b) = \left\{ s(t) \in \mathcal{L}_2 : s(t) = 0, t \notin [T_a, T_b] \text{ and } \int_{B_a}^{B_b} |s_{\mathcal{F}}(f)|^2 df \geq \|s\|^2(1-\eta) \right\}$$

then one can show that the dimensionality of $\mathcal{L}_2(T_a, T_b, B_a, B_b)$ is

$$n = \lfloor TB + 1 \rfloor$$

where $B = |B_b - B_a|$ and $T = |T_b - T_a|$ (see Wozencraft & Jacobs for more on this). As T goes to infinity, we see that the number $\frac{n}{T}$ of dimensions per second goes to B. Moreover, if one changes the value of η, then the essentially linear relationship between $\frac{n}{T}$ and B remains (but the constant in front of B may be different than 1). Be aware that many authors would say that a frequency domain pulse that has most of its energy in the interval $[-B, B]$ has bandwidth B (not $2B$ as we have defined it). The rationale for neglecting the negative frequencies is that with a spectrum analyzer, which is an instrument to see measure and plot the spectrum of real-valued signals, we see only the positive frequencies. We prefer our definition since it applies also when $B_a \neq -B_b$.

Example 54. *(Orthogonality via frequency shifts) The Fourier transform of the rectangular pulse $p(t)$ that has unit amplitude and support $[-\frac{T}{2}, \frac{T}{2}]$ is $p_{\mathcal{F}}(f) = T sinc(fT)$ and $p_l(t) = p(t)\exp(j2\pi l \frac{t}{T})$ has Fourier transform $p_{\mathcal{F}}(f - \frac{l}{T})$. The set $\{p_l(t)\}_{l=0}^{n-1}$ consists of a collection of n orthogonal waveform of duration T. For simplicity, but also to make the point that the above result is not sensitive to the definition of bandwidth, in this example we let the bandwidth of $p(t)$ be $2/T$. This is the width of the main lobe and it is the η-bandwidth for some η. Then the n pulses fit in a the frequency interval $[-\frac{1}{T}, \frac{n}{T}]$, which has width $\frac{n+1}{T}$. We have constructed n orthogonal signals with time-bandwidth-product equal $n+1$. (Be aware that in this example T is the support of one pulse whereas in the expression $n = \lfloor TB + 1 \rfloor$ it is the with of the union of all supports.)* □

Example 55. *(Orthogonality via time shifts) Let $p(t)$ and its bandwidth be defined as in the previous example. The set $\{p_l(t - lT)\}_{l=0}^{n-1}$ is a collection of n orthogonal waveforms. Recall that the Fourier transform of p_l is the Fourier transform of p times $\exp(-j2\pi lTf)$. This multiplicative term of unit magnitude does not affect the energy spectral density which is the squared magnitude of the Fourier transform. Hence regardless of η, the frequency interval that contains the fraction $1-\eta$ of the energy is the same for all p_l. If we take B as the bandwidth occupied by the main lobe of the sinc we obtain $BT = 2n$. In this example BT is larger than in the previous example by a factor 2. One is tempted to guess that this is due to the fact that we are using real-valued signals but it is actually not so. In fact if we use sinc pulses rather than rectangular pulses then we also construct real-valued time-domain pulses and obtain the same time bandwidth product as in the previous example. In fact in doing so we are just swapping the time and frequency variables of the previous example.* □

4.4 Examples of Large Signal Constellations

The aim of this section is to sharpen our intuition by looking at a few examples of signal constellation that contain a large number m of signals. We are interested in exploring what happens to the probability of error when the number $k = \log m$ of bits carried by one signal becomes large. In doing so we will let the energy grow linearly with k so as to keep the energy per bit constant, which seems to be fair. The dimensionality of the signal space will be $n = 1$ for the first example (PAM) and $n = 2$ for the second (PSK). In the third example (bit-by-bit on a pulse train) n will be equal to k. In the final example—an instance of block orthogonal signaling—we will have $n = 2^k$. These examples will provide useful insight about the role played by the dimensionality n.

4.4.1 Keeping BT Fixed While Growing k

Example 56. (PAM) *In this example we consider Pulse Amplitude Modulation. Let m be a positive even integer, $\mathcal{H} = \{0, 1, \ldots, m-1\}$ be the message set, and for each $i \in \mathcal{H}$ let $\boldsymbol{s}_i$ be a distinct element of $\{\pm a, \pm 3a, \pm 5a, \ldots \pm (m-1)a\}$. Here a is a positive number that determines the average energy $\mathcal{E}$. The waveform associated to message i is*

$$s_i(t) = \boldsymbol{s}_i \psi(t),$$

where ψ is an arbitrary unit-energy waveform[2]*. The signal constellation and the receiver block diagram are shown in Figure 4.2 and 4.3, respectively. We can easily verify that the probability of error of PAM is*

$$P_e = (2 - \frac{2}{m}) Q(\frac{a}{\sigma}),$$

where $\sigma^2 = N_0/2$. As shown in one of the problems, the average energy of the above constellation when signals are uniformly distributed is $\mathcal{E} = a^2(m^2-1)/3$. Equating to $\mathcal{E} = k\mathcal{E}_b$, solving for a, and using the fact that $k = \log_2 m$ yields

$$a = \sqrt{\frac{3\mathcal{E}_b \log_2 m}{(m^2-1)}},$$

which goes to 0 as m goes to ∞. Hence P_e goes to 1 as m goes to ∞. The next example uses a two-dimensional constellation.

□

Example 57. *(PSK) In this example we consider Phase-Shift-Keying. Let T be a positive number and define*

$$s_i(t) = \sqrt{\frac{2\mathcal{E}}{T}} \cos(2\pi f_0 t + \frac{2\pi}{m} i) 1_{[0,T]}(t), \quad i = 0, 1, \ldots, m-1. \tag{4.1}$$

[2] We follow our convention and write $\boldsymbol{s}_i$ in bold even if in this case it is a scalar.

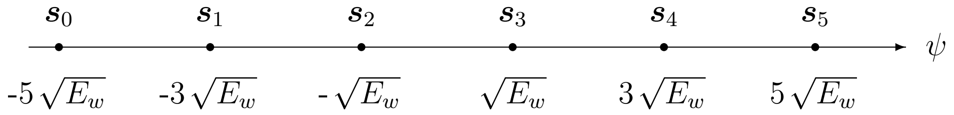


Figure 4.2: Signal Space Constellation for 6-ary PAM.

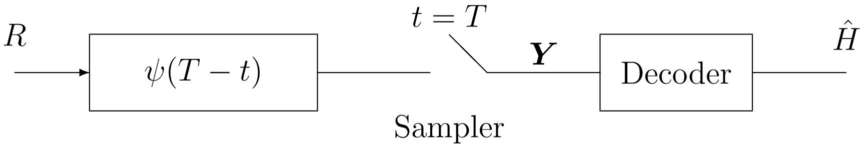


Figure 4.3: PAM Receiver

We assume $f_0 T = \frac{k}{2}$ *for some integer* k, *so that* $\|s_i\|^2 = \mathcal{E}$ *for all* i. *The signal space representation may be obtained by using the trigonometric equivalence* $\cos(\alpha+\beta) = \cos(\alpha)\cos(\beta) - \sin(\alpha)\sin(\beta)$ *to rewrite (4.1) as*

$$s_i(t) = s_{i,1}\psi_1(t) + s_{i,2}\psi_2(t),$$

where

$$\begin{aligned} s_{i1} &= \sqrt{\mathcal{E}}\cos\left(\tfrac{2\pi i}{m}\right), \quad \psi_1(t) = \sqrt{\tfrac{2}{T}}\cos(2\pi f_0 t)1_{[0,T]}(t), \\ s_{i2} &= \sqrt{\mathcal{E}}\sin\left(\tfrac{2\pi i}{m}\right), \quad \psi_2(t) = -\sqrt{\tfrac{2}{T}}\sin(2\pi f_0 t)1_{[0,T]}(t). \end{aligned}$$

Hence, the n*-tuple representation of the signals is*

$$\boldsymbol{s}_i = \sqrt{\mathcal{E}}\begin{pmatrix}\cos 2\pi i/m \\ \sin 2\pi i/m\end{pmatrix}.$$

In Example 15 we have already studied this constellation and derived the following lower bound to the error probability

$$P_e \geq 2Q\left(\sqrt{\frac{\mathcal{E}}{\sigma^2}}\sin\frac{\pi}{m}\right)\frac{m-1}{m},$$

where $\sigma^2 = \frac{N_0}{2}$ *is the variance of the noise in each coordinate.*

As in the previous example, let us see what happens as k *goes to infinity while* $\mathcal{E}_b$ *remains constant. Since* $\mathcal{E} = k\mathcal{E}_b$ *grows linearly with* k, *the circle that contains the signal points has radius* $\sqrt{\mathcal{E}} = \sqrt{k\mathcal{E}_b}$. *It's circumference grows with* $\sqrt{k}$ *while the number* $m = 2^k$ *of points on this circle grows exponentially with* k. *Hence the minimum distance between points goes to zero (indeed exponentially fast). As a consequence, the argument of the* Q *function that lowerbounds the probability of error for PSK goes to* 0 *and the probability of error goes to* 1. □

As they are, the signal constellations used in the above two examples are not suitable to transmit a large amount of data. The problem with the above two examples is that, as m grows, we are trying to pack more and more signal points into a space that also grows in size but does not grow fast enough. The space becomes "crowded" as m grows, meaning that the minimum distance becomes smaller, and the probability of error increases.

In the next example we try to do better. So far we have not made use of the fact that we expect to need more time to transmit more bits. In both of the above examples, the length T of the time interval used to communicate was constant. In the next example we let T grow linearly with the number of bits. This will free up a number of dimensions that grows linearly with k. (Recall that $n = BT$ is possible.)

4.4.2 Growing BT Linearly with k

Example 58. *(Bit by Bit on a Pulse Train) The idea is to transmit a signal of the form*

$$s_i(t) = \sum_{j=1}^{k} s_{i,j}\psi_j(t), \quad t \in \mathbb{R}, \tag{4.2}$$

and choose $\psi_j(t) = \psi(t - jT)$ for some waveform ψ that fulfills $\langle \psi_i, \psi_j \rangle = \delta_{ij}$. Assuming that it is indeed possible to find such a waveform, we obtain

$$s_i(t) = \sum_{j=1}^{k} s_{i,j}\psi(t - jT_s), \quad t \in \mathbb{R}. \tag{4.3}$$

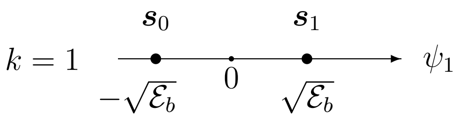
We let $m = 2^k$, so that to every message $i \in \mathcal{H} = \{0, 1, \ldots, m-1\}$ corresponds a unique binary sequence $(d_1, d_2, \ldots, d_k)$. It is convenient to see the elements of such binary sequences as elements of $\{\pm 1\}$ rather than $\{0, 1\}$. Let $(d_{i,1}, d_{i,2}, \ldots, d_{i,k})$ be the binary sequence that corresponds to message i and let the corresponding vector signal $\boldsymbol{s}_i = (s_{i,1}, s_{i,2}, \ldots, s_{i,k})^T$ be defined by

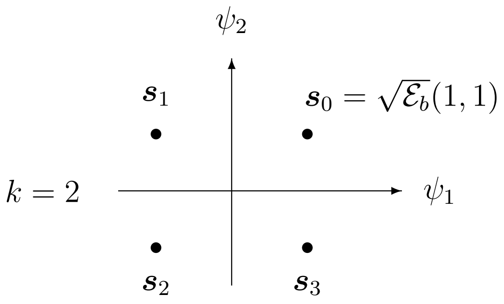
$$s_{i,j} = d_{i,j}\sqrt{\mathcal{E}_b}$$

where $\mathcal{E}_b = \frac{\mathcal{E}}{k}$ is the energy assigned to individual symbols. For reasons that should be obvious, the above signaling method will be called bit-by-bit on a pulse train.

There are various possible choices for ψ. Common choices are sinc pulses, rectangular pulses, and raised-cosine pulses (to be defined later). We will see how to choose ψ in Chapter 5.

To gain insight in the operation of the receiver and to determine the error probability, it is always a good idea to try to picture the signal constellation. In this case $\boldsymbol{s}_0, \ldots, \boldsymbol{s}_{m-1}$ are the vertices of a k-dimensional hypercube as shown in the figures below for $k = 1, 2$.





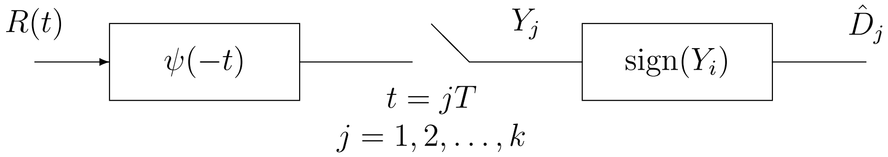
From the picture we immediately see what the decoding regions of a ML decoder are, but let us proceed analytically and find a ML decoding rule that works for any k. *The ML receiver decides that the constellation point used by the sender is one of the* $\boldsymbol{s} = (s_1, s_2, \ldots, s_k) \in \{\pm\sqrt{\mathcal{E}_b}\}^k$ *that maximizes* $\langle \boldsymbol{y}, \boldsymbol{s} \rangle - \frac{\|\boldsymbol{s}\|^2}{2}$. *Since* $\|\boldsymbol{s}\|^2$ *is the same for all constellation points, the previous expression is maximized iff* $\langle \boldsymbol{y}, \boldsymbol{s} \rangle = \sum y_j s_j$ *is maximized. The maximum is achieved with* $s_j = \operatorname{sign}(y_j)\sqrt{\mathcal{E}_b}$ *where*

$$\operatorname{sign}(y) = \begin{cases} 1, & y \geq 0 \\ -1, & y < 0. \end{cases}$$

The corresponding bit sequence is

$$\hat{d}_j = \operatorname{sign}(y_j).$$

The next figure shows the block diagram of our ML receiver. Notice that we need only one matched filter to do the k *projections. This is one of the reasons why we choose* $\psi_i(t) = \psi(t - iT_s)$. *Other reasons will be discussed in the next chapter.*



We now compute the error probability. As usual, we first compute the error probability conditioned on the event $\boldsymbol{S} = \boldsymbol{s} = (s_1, \ldots, s_k)$ *for some arbitrary constellation point* $\boldsymbol{s}$. *From the geometry of the signal constellation, we expect that the error probability will not depend on* $\boldsymbol{s}$. *If* s_j *is positive,* $Y_j = \sqrt{\mathcal{E}_b} + Z_j$ *and* $\hat{D}_j$ *will be correct iff* $Z_j \geq -\sqrt{\mathcal{E}_b}$. *This happens with probability* $1 - Q(\frac{\sqrt{\mathcal{E}_b}}{\sigma})$. *Reasoning similarly, you should verify that the probability of error is the same if* s_j *is negative. Now let* C_j *be the event that the decoder makes the correct decision about the* j *th bit. The probability of* C_j *depends only on* Z_j. *The independence of the noise components implies the independence of* C_1, C_2, …, C_k. *Thus, the probability that all* k *bits are decoded correctly when* $\boldsymbol{S} = \boldsymbol{s}_i$ *is*

$$P_c(i) = \left[1 - Q\left(\frac{\sqrt{\mathcal{E}_b}}{\sigma}\right)\right]^k.$$

Since this probability does not depend on i, $P_c = P_c(i)$.

Notice that $P_c \to 0$ *as* $k \to \infty$. *However, the probability that a specific symbol (bit) be decoded incorrectly is* $Q(\frac{\sqrt{\mathcal{E}_b}}{\sigma})$. *This is constant with respect to* k.

While in this example we have chosen to transmit a single bit per dimension, we could have transmitted instead some small number of bits per dimension by means of one of the methods discussed in the previous two examples. In that case we would have called the signaling scheme symbol by symbol on a pulse train. *Symbol by symbol on a pulse train will come up often in the remainder of this course. In fact it is the basis for most digital communication systems.*

□

The following question seems natural at this point: Is it possible to avoid that $P_c \to 0$ as $k \to \infty$? The next example shows that it is indeed possible.

4.4.3 Growing BT Exponentially With k

Example 59. *(Block Orthogonal Signaling) Let* $n = m = 2^k$, *pick* n *orthonormal waveforms* $\psi_1, \dots, \psi_n$ *and define* $s_1, \dots, s_m$ *to be*

$$s_i = \sqrt{\mathcal{E}}\psi_i.$$

This is called block orthogonal signaling. The name stems from the fact that one collects a block of k *bits and maps them into one of* 2^k *orthogonal waveforms. (In a real-world application* k *is a positive integer but for the purpose of giving specific examples with* $m = 3$ *we will not force* k *to be integer.) Notice that* $\|s_i\| = \sqrt{\mathcal{E}}$ *for all* i.

There are many ways to choose the 2^k *waveforms* ψ_i. *One way is to choose* $\psi_i(t) = \psi(t - iT)$ *for some normalized pulse* ψ *such that* $\psi(t-iT)$ *and* $\psi(t-jT)$ *are orthogonal when* $i \neq j$. *An example is*

$$\psi(t) = \sqrt{\frac{1}{T}} 1_{[0,T]}(t).$$

Notice that the requirement for ψ *is the same as in bit-by-bit on a pulse train, but now we need* 2^k *rather than* k *shifted versions. For obvious reasons this signaling method is sometimes called pulse position modulation.*

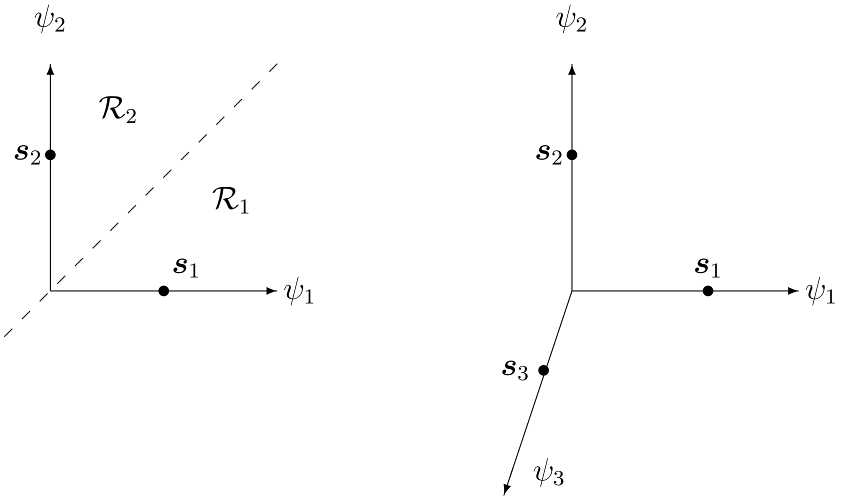
Another possibility is to choose

$$s_i(t) = \sqrt{\frac{2\mathcal{E}}{T}} \cos(2\pi f_i t) 1_{[0,T]}(t). \tag{4.4}$$

This is called m*-FSK (*m*-ary frequency shift keying). If we choose* $f_i T = k_i/2$ *for some* *integer k_i such that $k_i \neq k_j$ if $i \neq j$ then*

$$\begin{aligned}\langle s_i, s_j \rangle &= \frac{2\mathcal{E}}{T}\int_0^T \cos(2\pi f_i t)\cos(2\pi f_j t)dt \\ &= \frac{2\mathcal{E}}{T}\int_0^T \left[\frac{1}{2}\cos[2\pi(f_i+f_j)t] + \frac{1}{2}\cos[2\pi(f_i-f_j)t]\right] dt \\ &= \mathcal{E}\delta_{ij}\end{aligned}$$

as desired.



When $m \geq 3$, it is not easy to visualize the decoding regions. However we can proceed analytically using the fact that $\boldsymbol{s}_i$ is 0 everywhere except at position i where it is $\sqrt{\mathcal{E}}$. Hence,

$$\begin{aligned}\hat{H}_{ML}(\boldsymbol{y}) &= \arg\max_i \langle \boldsymbol{y}, \boldsymbol{s}_i \rangle - \frac{\mathcal{E}}{2} \\ &= \arg\max_i \langle \boldsymbol{y}, \boldsymbol{s}_i \rangle \\ &= \arg\max_i y_i.\end{aligned}$$

To compute (or bound) the error probability, we start as usual with a fixed $\boldsymbol{s}_i$. We pick $i = 1$. When $H = 1$,

$$Y_j = \begin{cases} Z_j & \textit{if } j \neq 1, \\ \sqrt{\mathcal{E}} + Z_j & \textit{if } j = 1. \end{cases}$$

Then

$$P_c(1) = Pr\{Y_1 > Z_2, Y_1 > Z_3, \ldots, Y_1, > Z_m | H = 1\}.$$

To evaluate the right side, we start by conditioning on $Y_1 = \alpha$, where $\alpha \in \mathbb{R}$ is an arbitrary number

$$Pr\{c|H = 1, Y_1 = \alpha\} = Pr\{\alpha > Z_2, \ldots, \alpha > Z_m\} = \left[1 - Q\left(\frac{\alpha}{\sqrt{N_0/2}}\right)\right]^{m-1},$$

and then remove the conditioning on Y_1,

$$\begin{aligned}
P_c(1) &= \int_{-\infty}^{\infty} f_{Y_1|H}(\alpha|1)\left[1-Q\left(\frac{\alpha}{\sqrt{N_0/2}}\right)\right]^{m-1} d\alpha \\
&= \int_{-\infty}^{\infty} \frac{1}{\sqrt{\pi N_0}} e^{-\frac{(\alpha-\sqrt{\mathcal{E}})^2}{N_0}} \left[1-Q\left(\frac{\alpha}{\sqrt{N_0/2}}\right)\right]^{m-1} d\alpha,
\end{aligned}$$

where we used the fact that when $H=1$, $Y_1 \sim \mathcal{N}(\sqrt{\mathcal{E}}, \frac{N_0}{2})$. *The above expression for* $P_c(1)$ *cannot be simplified further but one can evaluate it numerically. By symmetry,*

$$P_c = P_c(1) = P_c(i)$$

for all i.

The union bound is especially useful when the signal set $\{\boldsymbol{s}_1, \ldots, \boldsymbol{s}_m\}$ *is completely symmetric, like for orthogonal signals. In this case:*

$$\begin{aligned}
P_e = P_e(i) &\le (m-1)Q\left(\frac{d}{2\sigma}\right) \\
&= (m-1)Q\left(\sqrt{\frac{\mathcal{E}}{N_0}}\right) \\
&< 2^k \exp\left[-\frac{\mathcal{E}}{2N_0}\right] \\
&= \exp\left[-k\left(\frac{\mathcal{E}/k}{2N_0} - \ln 2\right)\right],
\end{aligned}$$

where we used $\sigma^2 = \frac{N_0}{2}$ *and*

$$d^2 = \|\boldsymbol{s}_i - \boldsymbol{s}_j\|^2 = \|\boldsymbol{s}_i\|^2 + \|\boldsymbol{s}_j\|^2 - 2\langle \boldsymbol{s}_i, \boldsymbol{s}_j\rangle = \|\boldsymbol{s}_i\|^2 + \|\boldsymbol{s}_j\|^2 = 2\mathcal{E}.$$

(The above is Pythagora's Theorem.)

If we let $\mathcal{E} = \mathcal{E}_b k$, *meaning that we let the signal's energy grow linearly with the number of bits as in bit-by-bit on a pulse train, then we obtain*

$$P_e < e^{-k(\frac{\mathcal{E}_b}{2N_0} - \ln 2)}.$$

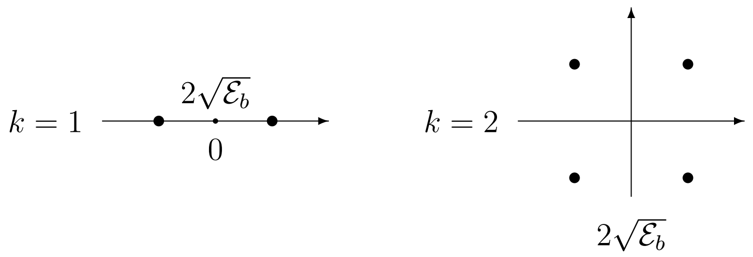
Now $P_e \to 0$ *as* $k \to \infty$, *provided that* $\frac{\mathcal{E}_b}{N_0} > 2\ln 2$. *(*$2\ln 2$ *is approximately* 1.39.*)*

□

4.5 Bit By Bit Versus Block Orthogonal

In the previous two examples we have let the number of dimensions n increase linearly and exponentially with k, respectively. In both cases we kept the energy per bit $\mathcal{E}_b$ fixed, and have let the signal energy $\mathcal{E} = k\mathcal{E}_b$ grow linearly with k. Let us compare the two cases.

In bit-by-bit on a pulse train the bandwidth is constant (we have not proved this yet, but this is consistent with the asymptotic limit $B = n/T$ seen in Section 4.3 applied with $T = nT_s$) and the signal duration increased linearly with k, which is quite natural. The drawback of bit-by-bit on a pulse train was found to be the fact that the probability of error goes to 1 as k goes to infinity. The union bound is a useful tool to understand why this happens. Let us use it to bound the probability of error when $H = i$. The union bound has one term for each alternative j. The dominating terms in the bound are those that correspond to signals $\boldsymbol{s}_j$ that are closest to $\boldsymbol{s}_i$. There are k closest neighbors, obtained by changing $\boldsymbol{s}_i$ in exactly one component, and each of them is at distance $2\sqrt{\mathcal{E}_b}$ from $\boldsymbol{s}_i$ (see the figure below). As k increases, the number of dominant terms goes up and so does the probability of error.



Let us now consider block orthogonal signaling. Since the dimensionality of the space it occupies grows exponentially with k, the expression $n = BT$ tells us that either the time or the bandwidth has to grow exponentially. This is a significant drawback. Using the bound

$$Q\left(\frac{d}{2\sigma}\right) \leq \frac{1}{2}\exp\left[\frac{d^2}{8\sigma^2}\right] = \frac{1}{2}\exp\left[-\frac{k\mathcal{E}_b}{2N_0}\right]$$

we see that the probability that the noise carries a signal closer to a specific neighbor goes down as $\exp\left(-\frac{k\mathcal{E}_b}{2N_0}\right)$. There are $2^k - 1 = e^{k\ln 2} - 1$ nearest neighbors (all alternative signals are nearest neighbors). For $\frac{\mathcal{E}_b}{2N_0} > k\ln 2$, the growth in distance is the dominating factor and the probability of error goes to 0. For $\frac{\mathcal{E}_b}{2N_0} < k\ln 2$ the number of neighbors is the dominating factor and the probability of error goes to 1.

Notice that the bit error probability P_b must satisfy $\frac{P_e}{k} \leq P_b \leq P_e$. The lower bound holds with equality if every block error results in a single bit error, whereas the upper bound holds with equality if a block error causes all bits to be decoded incorrectly. This expression guarantees that the bit error probability of block orthogonal signaling goes to 0 as $k \to \infty$ and provides further insight as to why it is possible to have the bit error probability be constant while the block error probability goes to 1 as in the case of bit-by-bit on a pulse train.

Do we want P_e to be small or are we happy with P_b small? It depends. If we are sending a file that contains a computer program, every single bit of the file has to be received correctly in order for the transmission to be successful. In this case we clearly want P_e to be small. On the other hand, there are sources that are more tolerant to occasional errors. This is the case of a digitized voice signal. For voice, it is sufficient to have P_b small.

4.6 Conclusion

We have discussed some of the trade-offs between the number of transmitted bits, the signal epoch, the bandwidth, the signal's energy, and the error probability. We have seen that, rather surprisingly, it is possible to transmit an increasing number k of bits at a fixed energy per bit $\mathcal{E}_b$ and make the probability that even a single bit is decoded incorrectly go to zero as k increases. However, the scheme we used to prove this has the undesirable property of requiring an exponential growth of the time bandwidth product. Ideally we would like to make the probability of error go to zero with a scheme similar to bit by bit on a pulse train. Is it possible? The answer is yes and the technique to do so is coding. We will give an example of coding in Chapter 6.

In this Chapter we have looked at the relationship between k, T, B, $\mathcal{E}$ and P_e by considering specific signaling methods. Information theory is a field that looks at these and similar communication problems from a more fundamental point of view that holds for every signaling method. A main result of information theory is the famous formula

$$C = B \log_2 \left(1 + \frac{P}{N_0 B} \right) \quad [\frac{\text{bits}}{\text{sec}}],$$

where B [Hz] is the bandwidth, N_0 the power spectral density of the additive white Gaussian noise, P the signal power, and C the transmission rate in bits/sec. Proving that one can transmit at rates arbitrarily close to C and achieve an arbitrarily small probability of error is a main result of information theory. Information theory also shows that at rates above C one can not reduce the probability of error below a certain value.

Appendix 4.A Isometries Do Not Affect the Probability of Error

Let

$$g(\gamma) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{\gamma^2}{2\sigma^2} \right), \ \gamma \in \mathbb{R}$$

so that for $\boldsymbol{Z} \sim \mathcal{N}(0, \sigma^2 I_n)$ we can write $f_{\boldsymbol{Z}}(\boldsymbol{z}) = g(\|\boldsymbol{z}\|)$. Then for any isometry $a : \mathbb{R}^n \to \mathbb{R}^n$ we have

$$\begin{aligned} P_c(i) &= Pr\{\boldsymbol{Y} \in \mathcal{R}_i | \boldsymbol{S} = \boldsymbol{s}_i\} \\ &= \int_{\boldsymbol{y} \in \mathcal{R}_i} g(\|\boldsymbol{y} - \boldsymbol{s}_i\|) d\boldsymbol{y} \\ &\overset{(a)}{=} \int_{\boldsymbol{y} \in \mathcal{R}_i} g(\|a(\boldsymbol{y}) - a(\boldsymbol{s}_i)\|) d\boldsymbol{y} \\ &\overset{(b)}{=} \int_{a(\boldsymbol{y}) \in a(\mathcal{R}_i)} g(\|a(\boldsymbol{y}) - a(\boldsymbol{s}_i)\|) d\boldsymbol{y} \\ &\overset{(c)}{=} \int_{\boldsymbol{\alpha} \in a(\mathcal{R}_i)} g(\|\boldsymbol{\alpha} - a(\boldsymbol{s}_i)\|) d\boldsymbol{\alpha} = Pr\{\boldsymbol{Y} \in a(\mathcal{R}_i) | \boldsymbol{S} = a(\boldsymbol{s}_i)\}, \end{aligned}$$

where in (a) we used the distance preserving property of an isometry, in (b) we used the fact that $y \in \mathcal{R}_i$ iff $a(\boldsymbol{y}) \in a(\mathcal{R}_i)$, and in (c) we made the change of variable $\boldsymbol{\alpha} = a(\boldsymbol{y})$ and used the fact that the Jacobian of an isometry is ± 1. The last line is the probability of decoding correctly when the transmitter sends $a(\boldsymbol{s}_i)$ and the corresponding decoding region is $a(\mathcal{R}_i)$.

Appendix 4.B Problems

PROBLEM 1. (Orthogonal Signal Sets) *Consider the following situation: A signal set* $\{s_j(t)\}_{j=0}^{m-1}$ *has the property that all signals have the same energy* $\mathcal{E}_s$ *and that they are mutually orthogonal:*

$$\langle s_i, s_j \rangle = \mathcal{E}_s \delta_{ij}. \tag{4.5}$$

Assume also that all signals are equally likely. The goal is to transform this signal set into a minimum-energy signal set $\{s_j^*(t)\}_{j=0}^{m-1}$. *It will prove useful to also introduce the unit-energy signals* $\phi_j(t)$ *such that* $s_j(t) = \sqrt{\mathcal{E}_s}\phi_j(t)$.

(a) Find the minimum-energy signal set $\{s_j^*(t)\}_{j=0}^{m-1}$.

(b) What is the dimension of span$\{s_0^*(t), \dots, s_{m-1}^*(t)\}$*? For* $m = 3$, *sketch* $\{s_j(t)\}_{j=0}^{m-1}$ *and the corresponding minimum-energy signal set.*

(c) What is the average energy per symbol if $\{s_j^*(t)\}_{j=0}^{m-1}$ *is used? What are the savings in energy (compared to when* $\{s_j(t)\}_{j=0}^{m-1}$ *is used) as a function of* m*?*

PROBLEM 2. (Antipodal Signaling and Rayleigh Fading) *Suppose that we use antipodal signaling (i.e* $s_0(t) = -s_1(t)$*). When the energy per symbol is* $\mathcal{E}_b$ *and the power spectral density of the additive white Gaussian noise in the channel is* $N_0/2$, *then we know that the average probability of error is*

$$Pr\{e\} = Q\left(\sqrt{\frac{\mathcal{E}_b}{N_0/2}}\right). \tag{4.6}$$

In mobile communications, one of the dominating effects is fading. *A simple model for fading is the following: Let the channel attenuate the signal by a random variable* A. *Specifically, if* $\boldsymbol{s}_i$ *is transmitted, the received signal is* $Y = A\boldsymbol{s}_i + N$. *The probability density function of* A *depends on the particular channel that is to be modeled.*[3] *Suppose* A *assumes the value* a. *Also assume that the receiver knows the value of* A *(but the sender does not). From the receiver point of view this is as if there is no fading and the transmitter uses the signals* $as_0(t)$ *and* $-as_0(t)$. *Hence,*

$$Pr\{e|A = a\} = Q\left(\sqrt{\frac{a^2\mathcal{E}_b}{N_0/2}}\right). \tag{4.7}$$

The average probability of error can thus be computed by taking the expectation over the random variable A, *i.e.*

$$Pr\{e\} = E_A[Pr\{e|A\}] \tag{4.8}$$

[3] In a more realistic model, not only the amplitude, but also the phase of the channel transfer function is a random variable.

An interesting, yet simple model is to take A to be a Rayleigh random variable, i.e.

$$f_A(a) \;=\; \begin{cases} 2ae^{-a^2}, & \text{if } a \geq 0, \\ 0, & \text{otherwise..} \end{cases} \tag{4.9}$$

This type of fading, which can be justified especially for wireless communications, is called Rayleigh fading.

(a) Compute the average probability of error for antipodal signaling subject to Rayleigh fading.

(b) Comment on the difference between Eqn. (4.6) (the average error probability without fading) and your answer in the previous question (the average error probability with Rayleigh fading). Is it significant? For an average error probability $Pr\{e\} = 10^{-5}$, find the necessary $\mathcal{E}_b/N_0$ for both cases.

Problem 3. (Root-Mean Square Bandwidth)

(a) The root-mean square (rms) bandwidth of a low-pass signal $g(t)$ of finite energy is defined by

$$W_{rms} = \left[\frac{\int_{-\infty}^{\infty} f^2 |G(f)|^2 df}{\int_{-\infty}^{\infty} |G(f)|^2 df}\right]^{1/2}$$

where $|G(f)|^2|$ is the energy spectral density of the signal. Correspondingly, the root mean-square (rms) duration of the signal is defined by

$$T_{rms} = \left[\frac{\int_{-\infty}^{\infty} t^2 |g(t)|^2 dt}{\int_{-\infty}^{\infty} |g(t)|^2 dt}\right]^{1/2}.$$

Using these definitions and assuming that $|g(t)| \to 0$ faster than $1/\sqrt{|t|}$ as $|t| \to \infty$, show that

$$T_{rms} W_{rms} \geq \frac{1}{4\pi}.$$

Hint: Use Schwarz's inequality

$$\left\{\int_{-\infty}^{\infty} [g_1^*(t) g_2(t) + g_1(t) g_2^*(t)] dt\right\}^2 \leq 4 \int_{-\infty}^{\infty} |g_1(t)|^2 dt \int_{-\infty}^{\infty} |g_2(t)|^2 dt$$

in which we set

$$g_1(t) = tg(t)$$

and

$$g_2(t) = \frac{dg(t)}{dt}.$$

(b) *Consider a Gaussian pulse defined by*

$$g(t) = \exp(-\pi t^2).$$

Show that for this signal, the equality

$$T_{rms} W_{rms} = \frac{1}{4\pi}$$

can be reached. Hint:

$$\exp(-\pi t^2) \overset{\mathcal{F}}{\longleftrightarrow} \exp(-\pi f^2).$$

Problem 4. (Minimum Energy for Orthogonal Signaling) *Let $H \in \{1, \dots, m\}$ be uniformly distributed and consider the communication problem described by:*

$$H = i : \qquad \boldsymbol{Y} = s_i + Z, \quad Z \sim \mathcal{N}(0, \sigma^2 I_m),$$

where $s_1, \dots, s_m$, $s_i \in \mathbb{R}^m$, is a set of constant-energy orthogonal signals. Without loss of generality we assume

$$s_i = \sqrt{\mathcal{E}} e_i,$$

where e_i is the ith unit vector in $\mathbb{R}^m$, i.e., the vector that contains 1 at position i and 0 elsewhere, and $\mathcal{E}$ is some positive constant.

(a) *Describe the statistic of Y_j (the jth component of $\boldsymbol{Y}$) for $j = 1, \dots, m$ given that $H = 1$.*

(b) *Consider a suboptimal receiver that uses a threshold $t = \alpha\sqrt{\mathcal{E}}$ where $0 < \alpha < 1$. The receiver declares $\hat{H} = i$ if i is the* only *integer such that $Y_i \geq t$. If there is no such i or there is more than one index i for which $Y_i \geq t$, the receiver declares that it can't decide. This will be viewed as an error.*

Let $E_i = \{Y_i \geq t\}$, $E_i^c = \{Y_i < t\}$, and describe, in words, the meaning of the event

$$E_1 \cap E_2^c \cap E_3^c \cap \cdots \cap E_m^c.$$

(c) *Find an upper bound to the probability that the above event* does not *occur when $H = 1$. Express your result using the Q function.*

(d) *Now we let $\mathcal{E}$ and $\ln m$ go to ∞ while keeping their ratio constant, namely $\mathcal{E} = \mathcal{E}_b \ln m \log_2 e$. (Here $\mathcal{E}_b$ is the energy per transmitted bit.) Find the smallest value of $\mathcal{E}_b/\sigma^2$ (according to your bound) for which the error probability goes to zero as $\mathcal{E}$ goes to ∞. Hint: Use $m - 1 < m = \exp(\ln m)$ and $Q(x) < \frac{1}{2}\exp(-\frac{x^2}{2})$.*

PROBLEM 5. (Pulse Amplitude Modulated Signals) *Consider using the signal set*

$$s_i(t) = s_i\phi(t), \quad i = 0, 1, \dots, m-1,$$

where $\phi(t)$ *is a unit-energy waveform,* $s_i \in \{\pm\frac{d}{2}, \pm\frac{3}{2}d, \dots, \pm\frac{m-1}{2}d\}$*, and* $m \geq 2$ *is an even integer.*

(a) *Assuming that all signals are equally likely, determine the average energy* $\mathcal{E}_s$ *as a function of* m*. Hint:* $\sum_{i=0}^{n} i^2 = \frac{n}{6} + \frac{n^2}{2} + \frac{n^3}{3}$*. Note: If you prefer you may determine an approximation of the average energy by assuming that* $S(t) = S\phi(t)$ *and* S *is a continuous random variable which is uniformly distributed in the interval* $[-\frac{m}{2}d, \frac{m}{2}d]$*.*

(b) *Draw a block diagram for the ML receiver, assuming that the channel is AWGN with power spectral density* $\frac{N_0}{2}$*.*

(c) *Give an expression for the error probability.*

(d) *For large values of* m*, the probability of error is essentially independent of* m *but the energy is not. Let* k *be the number of bits you send every time you transmit* $s_i(t)$ *for some* i*, and rewrite* $\mathcal{E}_s$ *as a function of* k*. For large values of* k*, how does the energy behaves when* k *increases by 1?*

PROBLEM 6. (Exact Energy of Pulse Amplitude Modulation) *In this problem you will compute the average energy* $\mathcal{E}(m)$ *of* m*-ary PAM. Throughout the problem,* m *is an arbitrary* positive even integer.

(a) *Let* U *and* V *be a two uniformly distributed discrete random variables that take values in* $\mathcal{U} = \{1, 3, \dots, (m-1)\}$ *and* $\mathcal{V} = \{\pm 1, \pm 3, \dots, \pm(m-1)\}$*, respectively. Argue (preferably in a rigorous mathematical way) that* $E[U^2] = E[V^2]$*.*

(b) *Let*

$$g(m) = \sum_{i \in \mathcal{U}} i^2.$$

The difference $g(m+2) - g(m)$ *is a polynomial in* m *of degree 2. Find this polynomial* $p(m)$*. For later use, notice that the relationship* $g(m+2) - g(m) = p(m)$ *holds also for* $m = 0$ *if we define* $g(0) = 0$*. Let us do that.*

(c) *Even though we are interested in evaluating* $g(\cdot)$ *only at positive even integers* m*, our aim is to find a function* $g : \mathbb{R} \to \mathbb{R}$ *defined over* $\mathbb{R}$*. Assuming that such a function exists and that it has second derivative, take the second derivative on both sides of* $g(m+2) - g(m) = p(m)$ *and find a function* $g''(m)$ *that fulfills the resulting recursion. Then integrate twice and find a general expression for* $g(m)$*. It will depend on two parameters introduced by the integration.*

(d) *If you could not solve (c), you may continue assuming that* $g(m)$ *has the general form* $g(m) = \frac{1}{6}m^3 + am + b$ *for some real valued* a *and* b. *Determine* $g(0)$ *and* $g(2)$ *directly from the definition of* $g(m)$ *given in question (b) and use those values to determine* a *and* b.

(e) *Express* $E\left[V^2\right]$ *in terms of the expression you have found for* $g(m)$ *and verify if for* $m = 2, 4, 6$. *Hint: Recall that* $E\left[V^2\right] = E\left[U^2\right]$.

(f) *More generally, let* S *be uniformly distributed in* $\{\pm d, \pm 3d, \ldots, \pm(m-1)d\}$ *where* d *is an arbitrary positive number and define* $\mathcal{E}(d, m) = E\left[S^2\right]$. *Use your results found thus far to determine a simple expression for* $\mathcal{E}(d, m)$.

(g) *Let* T *be uniformly distributed in* $[-md, md]$. *Computing* $E\left[T^2\right]$ *is straightforward, and one expects* $E[S^2]$ *to be close to* $E[T^2]$ *when* m *is large. Determine* $E\left[T^2\right]$ *and compare the result obtained via this continuous approximation to the exact value of* $E\left[S^2\right]$.

Chapter 5

Controlling the Spectrum

5.1 Introduction

In many applications, notably cellular communications, the power spectral density of the transmitted signal has to fit a certain frequency-domain mask. This restriction is meant to limit the amount of interference that a user can cause to users of adjacent bands. There are also situations when a restriction is selfimposed. For instance, if the channel attenuates certain frequencies more than others or the power spectral density of the noise is stronger at certain frequencies, then the channel is not equally good at all frequencies and by shaping the power spectral density of the transmitted signal so as to put more power there where the channel is good one can minimize the total transmit power for a given performance. This is done according to a technique called *water filling*. For these reasons we are interested in knowing how to shape the power spectral density of the signal produced by the transmitter. Throughout this chapter we consider the framework of Fig. 5.1, where the noise is white and Gaussian with power spectral density $\frac{N_0}{2}$ and $\{\psi(t-jT)\}_{j=-\infty}^{\infty}$ forms an orthonormal set. These assumptions guarantee many desirable properties, in particular that $\{s_j\}_{j=-\infty}^{\infty}$ is the sequence of coefficients of the orthonormal expansion of $s(t)$ with respect to the orthonormal basis $\{\psi(t-jT)\}_{j=-\infty}^{\infty}$ and that Y_j is the output of a discrete-time AWGN channel with input s_j and noise variance $\sigma^2 = \frac{N_0}{2}$.

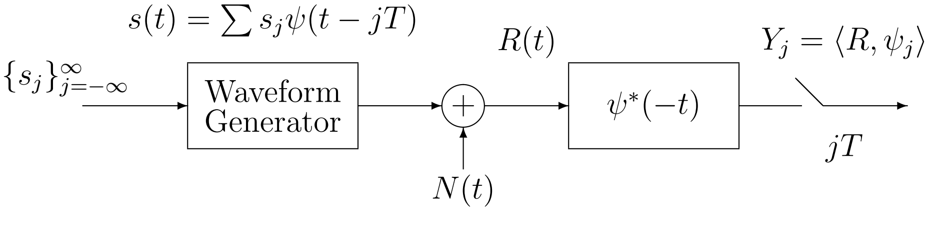

$s(t) = \sum s_j\psi(t-jT)$
$\{s_j\}_{j=-\infty}^{\infty}$
Waveform Generator
$R(t)$
$\psi^*(-t)$
$Y_j = \langle R, \psi_j \rangle$
jT
$N(t)$

Figure 5.1: Framework assumed in the current chapter.

The chapter is organized as follows. In Section 5.2 we consider a special and idealized case that consists in requiring that the power spectral density of the transmitted signal vanishes outside a frequency interval of the form $[-\frac{B}{2}, \frac{B}{2}]$. Even though such a strict restriction is not realistic in practice, we start with that case since it is quite instructive. In Section 5.3 we derive the expression for the power spectral density of the transmitted signal when the symbol sequence can be modeled as a discrete-time wide-sense-stationary process. We will see that when the symbols are uncorrelated—a condition often fulfilled in practice—the spectrum is proportional to $|\psi_{\mathcal{F}}|^2(f)$. In Section 5.4 we derive the necessary and sufficient condition on $|\psi_{\mathcal{F}}|^2(f)$ so that $\{\psi(t-jT)\}_{j=-\infty}^{\infty}$ forms an orthonormal sequence. Together sections 5.3 and 5.4 will give us the knowledge we need to tell which spectra are achievable within our framework and how to design the pulse $\psi(t)$ to achieve that spectrum.

5.2 The Ideal Lowpass Case

As a start, it is instructive to assume that the spectrum of the transmitted signal has to vanish outside a frequency interval $[-\frac{B}{2}, \frac{B}{2}]$ for some $B > 0$. This would be the case if the channel contained an ideal filter such as in Figure 5.2 where the filter frequency response is

$$h_{\mathcal{F}}(f) = \begin{cases} 1, & |f| \le \frac{B}{2} \\ 0, & \text{otherwise.} \end{cases}$$

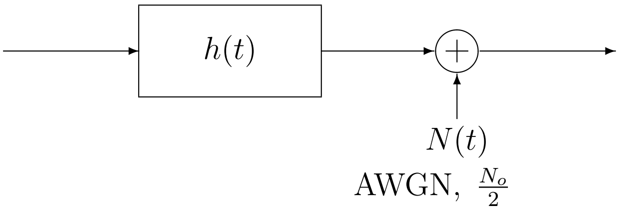


Figure 5.2: Lowpass channel model.

For years people have modeled the telephone line that way with $\frac{B}{2} = 4$ [KHz]. The sampling theorem is the right tool to deal with this situation.

Theorem 60. *(Sampling Theorem) Let* $\{s(t) : t \in \mathbb{R}\} \in \mathcal{L}_2$ *be such that* $s_{\mathcal{F}}(f) = 0$ *for* $f \notin [-\frac{B}{2}, \frac{B}{2}]$. *Then for all* $t \in \mathbb{R}$, $s(t)$ *is specified by the sequence* $\{s(nT)\}_{-\infty}^{\infty}$ *of samples and the parameter* T, *provided that* $T \le \frac{1}{B}$. *Specifically,*

$$s(t) = \sum_{n=-\infty}^{\infty} s(nT) \operatorname{sinc}\left(\frac{t}{T} - n\right) \tag{5.1}$$

where $\operatorname{sinc}(t) = \frac{\sin(\pi t)}{\pi t}$. □

For a proof of the sampling theorem see Appendix 5.A. In the same appendix we have also reviewed Fourier series since they are a useful tool to prove the sampling theorem and they will be useful later in this chapter.

The sinc pulse (used in the statement of the sampling theorem) is not normalized to unit energy. Notice that if we normalize the sinc pulse, namely define $\psi(t) = \frac{1}{\sqrt{T}} \operatorname{sinc}(\frac{t}{T})$, then $\{\psi(t-jT)\}_{j=-\infty}^{\infty}$ forms an orthonormal set. Thus (5.1) can be rewritten as

$$s(t) = \sum_{j=-\infty}^{\infty} s_j \psi(t-jT), \qquad \psi(t) = \frac{1}{\sqrt{T}} \operatorname{sinc}(\frac{t}{T}), \tag{5.2}$$

where $s_j = s(jT)\sqrt{T}$. This highlights the way we should think about the sampling theorem: a signal that fulfills the condition of the sampling theorem is one that lives in the inner product space spanned by $\{\psi(t-jT)\}_{j=-\infty}^{\infty}$ and when we sample such a signal we obtain (up to a scaling factor) the coefficients of its orthonormal expansion with respect to the orthonormal basis $\{\psi(t-jT)\}_{j=-\infty}^{\infty}$.

Now let us go back to our communication problem. We have just shown that any signal $s(t)$ that has no energy outside the frequency range $[-\frac{B}{2}, \frac{B}{2}]$ can be generated by the transmitter of Fig. 5.1. The channel in Fig. 5.1 does not contain the lowpass filter but this is immaterial since, by design, the lowpass filter is transparent to the transmitter output signal. Hence the receiver front end shown in Fig. 5.1 produces a sufficient statistic whether or not the channel contains the filter.

It is interesting to observe that the sampling theorem is somewhat used backwards in the diagram of Figure 5.1. Normally one starts with a signal from which one takes samples to represent the signal. In the setup of Figure 5.1 we start with a sequence of symbols produced by the *encoder* and we use them as the samples of the desired signal. Hence at the sender we are using the reconstruction part of the sampling theorem. The sampling is done by the receiver front end of Figure 5.1. In fact the filter with impulse response $\psi(-t)$ is an ideal lowpass filter that removes every frequency component outside $[-\frac{B}{2}, \frac{B}{2}]$. Thus $\{Y_j\}_{j=-\infty}^{\infty}$ is the sequence of samples of the bandlimited signal at the output of the receiver front-end filter.

From the input to the output of the block diagram of Figure 5.1 we see the discrete-time Gaussian channel depicted in Figure 5.3 and studied in Chapter 2. The channel takes and delivers a new symbol every T seconds.

5.3 Power Spectral Density

Even though we have not proved this, you may guess from the sampling theorem that the transmitter described in the previous section produces a strictly rectangular spectrum. This is true provided some condition (that we now derive) on the symbol sequence $\{s_j\}_{j=-\infty}^{\infty}$.

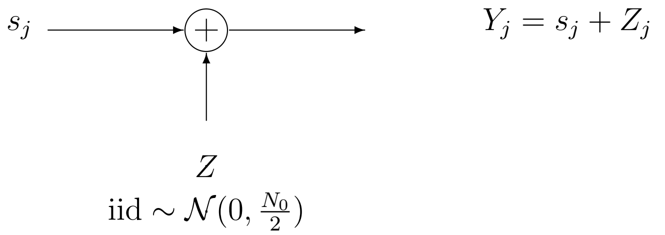


Figure 5.3: Equivalent discrete time channel.

Our aim is to be more general and not be limited to using sinc pulses. The question addressed in the current section is: how does the power spectral density of the transmitted signal relate to the pulse?

In order for the question to make sense, the transmitter output must be a wide-sense stationary process—the only processes for which the power spectral density is defined. As we will see, this is the case for any process of the form

$$X(t) \quad = \quad \sum_{i=-\infty}^{\infty} X_i \xi(t - iT - \Theta), \tag{5.3}$$

where $\{X_j\}_{j=-\infty}^{\infty}$ is a wide-sense stationary discrete-time process and Θ is a random dither (or delay) modeled as a uniformly distributed random variable taking value in the interval $[0, T)$. The pulse $\xi(t)$ may be any unit-energy pulse (not necessarily orthogonal to its shifts by multiples of T).

The first step to determine the power spectral density is to compute the autocorrelation. First define

$$R_X[i] = E[X_{j+i}X_j^*] \quad \text{and} \quad R_\xi(\tau) = \int_{-\infty}^{\infty} \xi(\alpha + \tau)\xi^*(\alpha) d\alpha. \tag{5.4}$$

Now we may compute the autocorrelation

$$\begin{aligned}
R_X(t+\tau,t) &= E[X(t+\tau)X^*(t)] \\
&= E\big[\sum_{i=-\infty}^{\infty} X_i\xi(t+\tau-iT-\Theta)\sum_{j=-\infty}^{\infty} X_j^*\xi^*(t-jT-\Theta)\big] \\
&= E\big[\sum_{i=-\infty}^{\infty}\sum_{j=-\infty}^{\infty} X_iX_j^*\xi(t+\tau-iT-\Theta)\xi^*(t-jT-\Theta)\big] \\
&= \sum_{i=-\infty}^{\infty}\sum_{j=-\infty}^{\infty} E[X_iX_j^*]E[\xi(t+\tau-iT-\Theta)\xi^*(t-jT-\Theta)] \\
&= \sum_{i=-\infty}^{\infty}\sum_{j=-\infty}^{\infty} R_X[i-j]E[\xi(t+\tau-iT-\Theta)\xi^*(t-jT-\Theta)] \\
&= \sum_{k=-\infty}^{\infty} R_X[k]\sum_{i=-\infty}^{\infty}\frac{1}{T}\int_0^T \xi(t+\tau-iT-\theta)\xi^*(t-iT+kT-\theta)d\theta \\
&= \sum_{k=-\infty}^{\infty} R_X[k]\frac{1}{T}\int_{-\infty}^{\infty}\xi(t+\tau-\theta)\xi^*(t+kT-\theta)d\theta.
\end{aligned}$$

Hence

$$R_X(\tau) = \sum_{k=-\infty}^{\infty} R_X[k]\frac{1}{T}R_\xi(\tau-kT), \tag{5.5}$$

where, with a slight abuse of notation, we have written $R_X(\tau)$ instead of $R_X(t+\tau,t)$ to emphasize that $R_X(t+\tau,t)$ depends only on the difference τ between the first and the second variable. It is straightforward to verify that $E[X(t)]$ does not depend on t either. Hence $X(t)$ is a wide-sense stationary process.

The power spectral density S_X is the Fourier transform of R_X. Hence,

$$S_X(f) = \frac{|\xi_{\mathcal{F}}(f)|^2}{T}\sum_k R_X[k]\exp(-j2\pi kfT). \tag{5.6}$$

In the above expression we used the fact that the Fourier transform of $R_\xi(\tau)$ is $|\xi_{\mathcal{F}}(f)|^2$. This follows from Parseval's relationship, namely

$$R_\xi(\tau) = \int_{-\infty}^{\infty}\xi(\alpha+\tau)\xi^*(\alpha)d\alpha = \int_{-\infty}^{\infty}\xi_{\mathcal{F}}(f)\xi_{\mathcal{F}}^*(f)\exp(j2\pi\tau f)df.$$

The last term says indeed that $R_\xi(\tau)$ is the Fourier inverse of $|\xi_{\mathcal{F}}(f)|^2$. Notice also that the summation in (5.6) is the discrete-time Fourier transform of $\{R_X[k]\}_{k=-\infty}^{\infty}$ evaluated at fT.

In many cases of interest $\{X_i\}_{i=-\infty}^{\infty}$ is a sequence of uncorrelated random variables. Then $R_X[k] = \mathcal{E}\delta_k$ where $\mathcal{E} = E[|X_j|^2]$ and the formulas simplify to

$$R_X(\tau) = \mathcal{E}\frac{R_\xi(\tau)}{T} \tag{5.7}$$

$$S_X(f) = \mathcal{E}\frac{|\xi_{\mathcal{F}}(f)|^2}{T}. \tag{5.8}$$

Example 61. *When $\xi(t) = \sqrt{1/T}\,\mathrm{sinc}(\frac{t}{T})$ and $R_X[k] = \mathcal{E}\delta_k$, the spectrum is $S_X(f) = \mathcal{E}1_{[-\frac{B}{2},\frac{B}{2}]}(f)$, where $B = \frac{1}{T}$. By integrating the power spectral density we obtain the power $B\mathcal{E} = \frac{\mathcal{E}}{T}$. This is consistent with our expectation: When we use the pulse $\mathrm{sinc}(\frac{t}{T})$ we expect to obtain a spectrum which is flat for all frequencies in $[-\frac{B}{2},\frac{B}{2}]$ and vanishes outside this interval. The energy per symbol is $\mathcal{E}$. Hence the power is $\frac{\mathcal{E}}{T}$.* □

5.4 Generalization Using Nyquist Pulses

To simplify the discussion let us assume that the stochastic process that models the symbol sequence is uncorrelated. Then the power spectral density of the transmitter output process is given by (5.8). Unfortunately we are not free to choose $|\xi_{\mathcal{F}}(f)|^2$, since we are limited to those choices for which $\{\xi(t-jT)\}_{j=-\infty}^{\infty}$ forms an orthonormal sequence. The goal of this section is to derive a necessary and sufficient condition on $|\xi_{\mathcal{F}}(f)|^2$ in order for $\{\xi(t-jT)\}_{j=-\infty}^{\infty}$ to form an orthonormal sequence. To remind ourself of the orthonormal condition we revert to our original notation and use $\psi(t)$ to represent the pulse.

Our aim is a frequency-domain characterization of the property

$$\int_{-\infty}^{\infty} \psi(t-nT)\psi^*(t)dt = \delta_n. \tag{5.9}$$

The form of the left hand side suggests using Parseval's relationship. Following that lead we obtain

$$\begin{aligned}
\delta_n = \int_{-\infty}^{\infty} \psi(t-nT)\psi^*(t)dt &= \int_{-\infty}^{\infty} \psi_{\mathcal{F}}(f)\psi_{\mathcal{F}}^*(f)e^{-j2\pi nTf}df \\
&= \int_{-\infty}^{\infty} |\psi_{\mathcal{F}}|^2(f)e^{-j2\pi nTf}df \\
&\overset{(a)}{=} \int_{-\frac{1}{2T}}^{\frac{1}{2T}} \sum_{k\in\mathbb{Z}} |\psi_{\mathcal{F}}|^2(f-\frac{k}{T})e^{-j2\pi nTf}df \\
&\overset{(b)}{=} \int_{-\frac{1}{2T}}^{\frac{1}{2T}} g(f)e^{-j2\pi nTf}df,
\end{aligned}$$

where in (a) we used the fact that for an arbitrary function $u : \mathbb{R} \to \mathbb{R}$ and an arbitrary positive value a,

$$\int_{-\infty}^{\infty} u(x)dx = \int_{-\frac{a}{2}}^{\frac{a}{2}} \sum_{i=-\infty}^{\infty} u(x+ia)dx,$$

as well as the fact that $e^{-j2\pi nT(f-\frac{k}{T})} = e^{-j2\pi nTf}$, and in (b) we have defined

$$g(f) = \sum_{k\in\mathbb{Z}} |\psi_{\mathcal{F}}|^2(f+\frac{k}{T}).$$

Notice that g is a periodic function of period $1/T$ and the right side of (b) above is $1/T$ times the nth Fourier series coefficient A_n of the periodic function g. Thus the above chain of equalities establishes that $A_0 = T$ and $A_n = 0$ for $n \neq 0$. These are the Fourier series coefficients of a constant function of value T. Due to the uniqueness of the Fourier series expansion we conclude that $g(f) = T$ for all values of f. Due to the periodicity of g, this is the case if and only if g is constant in any interval of length $1/T$. We have proved the following theorem:

Theorem 62. *(Nyquist) . A waveform $\psi(t)$ is orthonormal to each shift $\psi(t-nT)$ if and only if*

$$\sum_{k=-\infty}^{\infty} |\psi_{\mathcal{F}}(f+\frac{k}{T})|^2 = T \quad \textit{for all } f \textit{ in some interval of length } \frac{1}{T}. \tag{5.10}$$

□

Waveforms that fulfill Nyquist theorem are called Nyquist pulses. A few comments are in order:

(a) Often we are interested in Nyquist pulses that have small bandwidth, between $1/2T$ and $1/T$. For pulses that are strictly bandlimited to $1/T$ or less, the Nyquist criterion is satisfied if and only if $|\psi_{\mathcal{F}}(\frac{1}{2T}-\epsilon)|^2 + |\psi_{\mathcal{F}}(-\frac{1}{2T}-\epsilon)|^2 = T$ for $\epsilon \in [-\frac{1}{2T}, \frac{1}{2T}]$ (See the picture below). If we assume (as we do) that $\psi(t)$ is real-valued, then $|\psi_{\mathcal{F}}(-f)|^2 = |\psi_{\mathcal{F}}(f)|^2$. In this case the above relationship is equivalent to

$$|\psi_{\mathcal{F}}(\frac{1}{2T}-\epsilon)|^2 + |\psi_{\mathcal{F}}(\frac{1}{2T}+\epsilon)|^2 = T, \qquad \epsilon \in [0, \frac{1}{2T}].$$

This means that $|\psi_{\mathcal{F}}(\frac{1}{2T})|^2 = \frac{T}{2}$ and the amount by which $|\psi_{\mathcal{F}}(f)|^2$ increases when we go from $f = \frac{1}{2T}$ to $f = \frac{1}{2T} - \epsilon$ equals the decrease when we go from $f = \frac{1}{2T}$ to $f = \frac{1}{2T} + \epsilon$. An example is given in Figure 5.4.

(b) The sinc pulse is just a special case of a Nyquist pulse. It has the smallest possible bandwidth, namely $1/2T$ [Hz], among all pulses that satisfy Nyquist criterion for a given T. (Draw a picture if this is not clear to you).

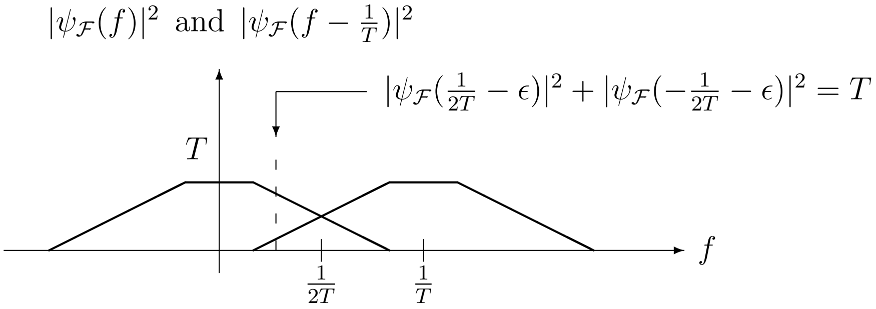

$|\psi_{\mathcal{F}}(f)|^2$ and $|\psi_{\mathcal{F}}(f-\frac{1}{T})|^2$
$|\psi_{\mathcal{F}}(\frac{1}{2T}-\epsilon)|^2+|\psi_{\mathcal{F}}(-\frac{1}{2T}-\epsilon)|^2=T$
T
f
$\frac{1}{2T}$
$\frac{1}{T}$

Figure 5.4: Nyquist condition for pulses $\psi_{\mathcal{F}}(f)$ that have support within $[-\frac{1}{T},\frac{1}{T}]$.

(c) Nyquist criterion is a condition expressed in the frequency domain. It is equivalent to the time domain condition (5.9). Hence if one asks you to "verify that $\psi(t)$ fulfills Nyquist criterion" it does not mean that you have to take the Fourier transform of ψ and then check that $\psi_{\mathcal{F}}$ fulfills (5.10). It may be easier to check if ψ fulfills the time-domain condition (5.9).

(d) Any pulse $\psi(t)$ that satisfies

$$|\psi_{\mathcal{F}}(f)|^2=\begin{cases}T, & |f|\le\frac{1-\beta}{2T}\\ \frac{T}{2}\left(1+\cos\left[\frac{\pi T}{\beta}\left(|f|-\frac{1-\beta}{2T}\right)\right]\right), & \frac{1-\beta}{2T}<|f|<\frac{1+\beta}{2T}\\ 0, & \text{otherwise}\end{cases}$$

for some $\beta\in(0,1)$ fulfills Nyquist criterion. Such a pulse is called *raised-cosine pulse.* (See Figure 5.7 (top) for a raised-cosine pulse with $\beta=\frac{1}{2}$.) Using the relationship $\cos^2\frac{\alpha}{2}=\frac{1}{2}(1+\cos\alpha)$, we can immediately verify that the following $\psi_{\mathcal{F}}(f)$ satisfies the above relationship

$$\psi_{\mathcal{F}}(f)=\begin{cases}\sqrt{T}, & |f|\le\frac{1-\beta}{2T}\\ \sqrt{T}\cos\frac{\pi T}{2\beta}(|f|-\frac{1-\beta}{2T}), & \frac{1-\beta}{2T}<|f|\le\frac{1+\beta}{2T}\\ 0, & \text{otherwise.}\end{cases}$$

$\psi_{\mathcal{F}}(f)$ is called square-root raised-cosine pulse[1]. The inverse Fourier transform of $\psi_{\mathcal{F}}(f)$, derived in Appendix 5.E, is

$$\psi(t)=\frac{4\beta}{\pi\sqrt{T}}\frac{\cos\left((1+\beta)\pi\frac{t}{T}\right)+\frac{(1-\beta)\pi}{4\beta}\operatorname{sinc}\left((1-\beta)\frac{t}{T}\right)}{1-\left(4\beta\frac{t}{T}\right)^2},$$

where $\operatorname{sinc}(x)=\frac{\sin(\pi x)}{\pi x}$. Notice that when $t=\pm\frac{T}{4\beta}$, both the numerator and the denominator of the above expression become 0. One can show that the limit as t approaches $\pm\frac{T}{4\beta}$ is $\beta\sin\left[\frac{(1+\beta)\pi}{4\beta}\right]+\frac{2\beta}{\pi}\sin\left[\frac{(1-\beta)\pi}{4\beta}\right]$. The pulse $\psi(t)$ is plotted in Figure 5.7 (bottom) for $\beta=\frac{1}{2}$.

[1] In is common practice to use the name square-root raised-cosine pulse to refer to the inverse Fourrier transform of $\psi_{\mathcal{F}}(f)$, i.e., to the time-domain waveform $\psi(t)$.

(e) We have derived Nyquist criterion inspired by what we have done in Section 5.2. However, Nyquist criterion is not limited to lowpass signals. If ψ fulfills Nyquist criterion and has bandpass characteristic then it will give rise to a bandpass signal $s(t)$.

5.5 Summary and Conclusion

The communication problem, as we see it in this course, consists of signaling to the recipient the message chosen by the sender. The message is one out of m possible messages. For each message there is a unique signal used as a proxi to communicate the message across the channel.

Regardless of how we pick the m signals, which are assumed to be finite-energy and known to the sender and the receiver, there exists an orthonormal basis $\psi_1 \dots \psi_n$ and a constellation of points $\boldsymbol{s}_0, \dots, \boldsymbol{s}_{m-1}$ in $\mathbb{C}^n$ such that

$$s_i(t) = \sum_{j=1}^{n} s_{ij}\psi_j(t), \quad i = 0, \dots, m-1. \tag{5.11}$$

A minimum-probability-of-error receiver that observes the received signal R may decide which message was signaled based on the sufficient statistic $\boldsymbol{Y} = (Y_1, \dots, Y_n)^T \in \mathbb{C}^n$, where $Y_j = \langle R, \psi_j \rangle$.

It is up to us to decide if we want to start by choosing the m waveforms s_i, $i = 0, \dots, m-1$ and then, if we so choose, use the Gram Schmidt procedure to find an orthonormal basis $\psi_1 \dots \psi_n$ and the associated constellation of n-tuples $\boldsymbol{s}_0 \dots, \boldsymbol{s}_{m-1}$, or if we want to start with an arbitrary orthonormal basis $\psi_1 \dots \psi_n$ and a selection of m n-tuples $\boldsymbol{s}_0 \dots, \boldsymbol{s}_{m-1}$ and let the signaling waveforms be obtained through (5.11). The latter approach has the advantage of decoupling design choices that can be made independently and with different objectives in mind: The orthonormal basis affects the duration and the bandwidth of the signals whereas the n-tuples of coefficients affect the transmit power and the probability of error.

In Chapter 4 we have already come across the idea of letting $\psi_1, \dots \psi_n$ be obtained from a single pulse ψ by the assignment $\psi_i(t) = \psi(t - iT)$. In that occasion our motivation was to give an example of a signaling scheme for which the number of dimensions occupied by the signal space grew linearly with the number k of transmitted bits. Implicit was the hope that we could let the dimensionality grow by letting the bandwidth be constant and letting the signal epoch grow linearly with k. In the present chapter we went further in at least two ways. In Section 5.3 we have seen how exactly the power spectral density of the transmitted signal depends on $|\psi_{\mathcal{F}}(f)|^2$ and in Section 5.4 we have derived the condition that $|\psi_{\mathcal{F}}(f)|^2$ has to satisfy so that $\{\psi(t - iT)\}_{i=-\infty}^{\infty}$ be an orthonormal set.

Another very important consequence of choosing $\psi_i(t) = \psi(t - iT)$ is the simplicity of the receiver front end: all the projections can be done with a single matched filter of impulse response $\psi^*(t_0 - t)$ for an arbitrary t_0. Specifically, $Y_i = \langle R, \psi_i \rangle$ is the filter output at time $t_0 + iT$.

Appendix 5.A Fourier Series

We briefly review the Fourier series focusing on the big picture and on how to remember things.

Let $f(x)$ be a periodic function, $x \in \mathbb{R}$. It has period p if $f(x) = f(x+p)$ for all $x \in \mathbb{R}$. Its fundamental period is the smallest such p. We are using the "physically unbiased" variable x instead of t (which usually represents time) since we want to emphasize that we are dealing with a general periodic function, not necessarily a function of time.

The periodic function $f(x)$ can be represented as a linear combination of complex exponentials of the form $e^{j2\pi \frac{x}{p} i}$. These are all the complex exponentials that have period p. Hence

$$f(x) = \sum_{i \in \mathbb{Z}} A_i \, e^{j2\pi \frac{x}{p} i} \tag{5.12}$$

for some sequence of coefficients $\dots A_{-1}, A_0, A_1, \dots$ with value in $\mathbb{C}$. This says that a function of fundamental period p may be written as a linear combination of all the complex exponentials of period p.

Two functions of fundamental period p are identical iff they coincide over a period. Hence to check if a given series of coefficients $\dots A_{-1}, A_0, A_1, \dots$ is the correct series, it is sufficient to verify that

$$f(x) 1_{[-\frac{p}{2}, \frac{p}{2}]}(x) = \sum_{i \in \mathbb{Z}} \sqrt{p} A_i \frac{e^{j\frac{2\pi}{p} xi}}{\sqrt{p}} 1_{[-\frac{p}{2}, \frac{p}{2}]}(x),$$

where we have multiplied and divided by $\sqrt{p}$ to make $\phi_i(x) = \frac{e^{j\frac{2\pi}{p} xi}}{\sqrt{p}} 1_{[-\frac{p}{2}, \frac{p}{2}]}(x)$, $i \in \mathbb{Z}$, an orthonormal basis. Hence the right side of the above expression is an orthonormal expansion. The coefficients of an orthonormal expansion are always found in the same way. Specifically, the ith coefficient $\sqrt{p} A_i$ is the result of $\langle f, \phi_i \rangle$. Hence,

$$A_i = \frac{1}{p} \int_{-\frac{p}{2}}^{\frac{p}{2}} f(x) e^{-j\frac{2\pi}{p} xi} dx. \tag{5.13}$$

We hope that this will make it easier for you to remember (or re-derive) (5.12) and (5.13).

Appendix 5.B Sampling Theorem: Fourier Series Proof

As an example of the utility of this relationship we derive the sampling theorem. Recall that the sampling theorem states that any $\mathcal{L}_2$ function $s(t)$ such that $s_{\mathcal{F}}(f) = 0$ for $f \notin [-\frac{B}{2}, \frac{B}{2}]$ may be written as

$$s(t) = \sum_k s(kT) \ \text{sinc}\left(\frac{t - kT}{T} \right)$$

where $T = \frac{1}{B}$.

Proof of the sampling theorem: By assumption, $s_{\mathcal{F}}(f) = 0$, $f \notin [-\frac{B}{2}, \frac{B}{2}]$. Hence there is a one-to-one relationship between $s_{\mathcal{F}}(f)$ and its periodic extension defined as $\tilde{s}_{\mathcal{F}}(f) = \sum_n s_{\mathcal{F}}(f - n/T)$. In fact

$$s_{\mathcal{F}}(f) = \tilde{s}_{\mathcal{F}}(f) 1_{[-\frac{B}{2}, \frac{B}{2}]}(f).$$

The periodic extension may be written as a Fourier series. This means that $\tilde{s}_{\mathcal{F}}(f)$ can be described by a sequence of numbers. Thus $s_{\mathcal{F}}(f)$ as well as $s(t)$ can be described by the same sequence of numbers. To determine the relationship between $s(t)$ and those numbers we write

$$s_{\mathcal{F}}(f) = \tilde{s}_{\mathcal{F}}(f) 1_{[-\frac{B}{2}, \frac{B}{2}]}(f) = \sum_k A_k e^{+j\frac{2\pi}{B}fk} 1_{[-\frac{B}{2}, \frac{B}{2}]}(f)$$

and take the Fourier transform on both sides using

$$1_{[-\frac{B}{2}, \frac{B}{2}]}(f) \Leftrightarrow \frac{1}{T} \operatorname{sinc}(\frac{t}{T}), \quad T = \frac{1}{B},$$

and the time shift property of the Fourier transform

$$h(t - \tau) \Leftrightarrow h_{\mathcal{F}}(f) e^{-j2\pi f \tau},$$

to obtain

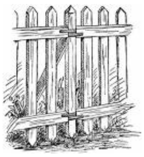
$$s(t) = \sum_k \frac{A_k}{T} \operatorname{sinc}\left(\frac{t + kT}{T}\right).$$

We still need to determine $\frac{A_k}{T}$. It is straightforward to determine A_k from its definition, but it is easier to observe that if we plug in $t = nT$ on both sides of the expression above we obtain $s(nT) = \frac{A_{-n}}{T}$. This completes the proof. To see that we may indeed obtain A_k from the definition (5.13) we write

$$A_k = T \int_{-\frac{1}{2T}}^{\frac{1}{2T}} \tilde{s}_{\mathcal{F}}(f) e^{-j\frac{2\pi}{B}kf} df = T \int_{-\infty}^{\infty} s_{\mathcal{F}}(f) e^{-j\frac{2\pi}{B}kf} df = Ts(-kT),$$

where the first equality is the definition of the Fourier coefficient A_k, the second uses the fact that $s_{\mathcal{F}}(f) = 0$ for $f \notin [-\frac{B}{2}, \frac{B}{2}]$, and the third is the inverse Fourier transform evaluated at $t = -kT$. The above sequence of equalities show that to have the desirable property that A_k is a sample of $s(t)$ it is key that the support of $s_{\mathcal{F}}(f)$ be confined to one period of $\tilde{s}_{\mathcal{F}}(f)$. One could imagine the support of $s_{\mathcal{F}}(f)$ be such that it is the union of disjoint intervals placed in such a way that there is a one to one relationship between $s_{\mathcal{F}}(f)$ and $\tilde{s}_{\mathcal{F}}(f)$. In that case it is still true $s(t)$ can be described by a sequence of numbers (the Fourier series coefficients of $\tilde{s}_{\mathcal{F}}(f)$) but those numbers are not necessarily the samples of $s(t)$. □

Appendix 5.C The Picket Fence Miracle



The T-spaced picket fence is the train of Dirac delta functions

$$\sum_{n=-\infty}^{\infty} \delta(x-nT).$$

The picket fence miracle refers to the fact that the Fourier transform of a picket fence is again a (scaled) picket fence. Specifically,

$$\mathcal{F}\left[\sum_{n=-\infty}^{\infty} \delta(t-nT)\right] = \frac{1}{T}\sum_{n=-\infty}^{\infty} \delta(f-\frac{n}{T}).$$

To prove the above relationship, we expand the periodic function $\sum \delta(t-nT)$ as a Fourier series, namely

$$\sum_{n=-\infty}^{\infty} \delta(t-nT) = \frac{1}{T}\sum_{n=-\infty}^{\infty} e^{j2\pi\frac{t}{T}n}.$$

Taking the Fourier transform on both sides yields

$$\mathcal{F}\left[\sum_{n=-\infty}^{\infty} \delta(t-nT)\right] = \frac{1}{T}\sum_{n=-\infty}^{\infty} \delta\left(f-\frac{n}{T}\right)$$

which is what we wanted to prove.

It is convenient to have a notation for the picket fence. Thus we define[2]

$$E_T(x) = \sum_{n=-\infty}^{\infty} \delta(x-nT).$$

Using this notation, the relationship that we have just proved may be written as

$$\mathcal{F}\left[E_T(t)\right] = \frac{1}{T}E_{\frac{1}{T}}(f).$$

[2] The choice of the letter E is suggested by the fact that it looks like a picket fence when rotated by 90 degrees.

Appendix 5.D Sampling Theorem: Picket Fence Proof

In this Appendix we give a somewhat less rigorous but more pictorial proof of the sampling theorem. The proof makes use of the picket fence miracle.

Let $s(t)$ be the signal of interest and let $\{s(nT)\}_{n=-\infty}^{\infty}$ be its samples taken every T seconds. From the samples we can form the signal

$$s|(t) = \sum_{n=-\infty}^{\infty} s(nT)\delta(t-nT).$$

($s|$ is just a name for a function.) Notice that

$$s|(t) = s(t)E_T(t).$$

Taking the Fourier transform on both sides yields

$$\mathcal{F}[s|(t)] = s_{\mathcal{F}}(t) * \left(\frac{1}{T}E_{\frac{1}{T}}(f)\right) = \frac{1}{T}\sum_{n-\infty}^{\infty} s_{\mathcal{F}}\left(f - \frac{n}{T}\right).$$

Hence the Fourier transform of $s|(t)$ is the superposition of $\frac{1}{T}s_{\mathcal{F}}(f)$ with all of its shifts by multiples of $\frac{1}{T}$, as shown in Fig. 5.5.

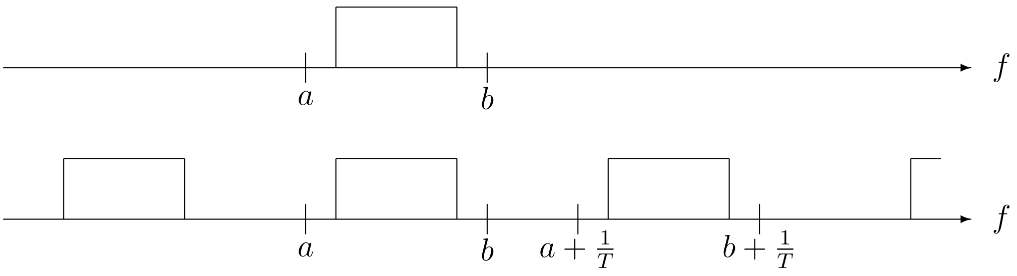


Figure 5.5: The Fourier transform of $s(t)$ and that of $s|(t) = \sum s(nT)\delta(t-nT)$.

From Fig. 5.5 it is obvious that we can reconstruct the original signal $s(t)$ by filtering $s|(t)$ with a filter that passes $s_{\mathcal{F}}(f)$ and blocks $\sum_{n\neq-\infty}^{\infty} s_{\mathcal{F}}\left(f-\frac{n}{T}\right)$. Such a filter exists if, like in the figure, the support of $s_{\mathcal{F}}(f)$ lies in an interval $\mathcal{I}(a,b)$ of width smaller than $\frac{1}{T}$. We do not want to assume that the receiver knows the support of each individual $s_{\mathcal{F}}(f)$, but we can assume that it knows that $s_{\mathcal{F}}(f)$ belongs to a class of signals that have support contained in some interval $\mathcal{I} = (a,b)$ for some real numbers $a < b$. (See Fig. 5.5). If this is the case and we know that $b - a < \frac{1}{T}$ then we can reconstruct $s(t)$ from $s|(t)$ by filtering the latter with any filer of impulse response $h(t)$ that satisfies

$$h_{\mathcal{F}}(f) = \begin{cases} T, & f \in \mathcal{I} \\ 0, & f \in \bigcup_{n\neq 0} \mathcal{I} + \frac{n}{T}, \end{cases}$$

where by $\mathcal{I}+\frac{n}{T}$ we mean $(a+\frac{n}{T}, b+\frac{n}{T})$. Hence

$$s_{\mathcal{F}}(f)=\left(\sum_{n=-\infty}^{\infty} s_{\mathcal{F}}\left(f-\frac{n}{T}\right)\right)h_{\mathcal{F}}(f),$$

and, after taking the inverse Fourier transform on both sides, we obtain the *reconstruction* (also called *interpolation*) formula

$$s(t)=\sum_{n=-\infty}^{\infty} s(nT)h(t-nT).$$

We summarize the various relationships in the following diagram that holds for a fixed T. The diagram says that the samples of a signal $s(t)$ are in one-to-one relationship with $\tilde{s}_{\mathcal{F}}(f)$. From the latter we can reconstruct $s_{\mathcal{F}}(f)$ if there is a single $s_{\mathcal{F}}(f)$ that could have lead to $\tilde{s}_{\mathcal{F}}(f)$, which is the case if the condition of the sampling theorem is satisfied, i.e., if $b-a<\frac{1}{T}$. The star on an arrow is meant to remind us that the map in that direction is unique if the condition of the sapling theorem is met.

$$\begin{tikzcd}
 & \{s(nT)\}_{n=-\infty}^{\infty} \arrow[d, Leftrightarrow] & \\
s(t) \arrow[r, Rightarrow, shift right=1] & s|(t) \arrow[l, Rightarrow, shift right=1, "(*)"'] \arrow[d, Leftrightarrow] & =\sum_{n=-\infty}^{\infty} s(n)\delta(t-nT) \\
s_{\mathcal{F}}(f) \arrow[r, Rightarrow, shift right=1] & \tilde{s}_{\mathcal{F}}(f) \arrow[l, Rightarrow, shift right=1, "(**)"'] & =\sum s_{\mathcal{F}}(f-\frac{n}{T})
\end{tikzcd}$$

To emphasize the importance of knowing $\mathcal{I}$, observe that if $s(t)\in\mathcal{L}_2(\mathcal{I})$ then $s(t)e^{-j\frac{2\pi}{T}t}\in\mathcal{L}_2(\mathcal{I}+\frac{1}{T})$. These are different signals that, when sampled at multiples of nT, lead to the same sample sequence.

It is clear that the condition $b-a<\frac{1}{T}$ is not only sufficient but also necessary to guarantee that no two signals such that the support of their Fourier transform lies in $\mathcal{I}$ do not lead to the same sample sequence. Fig. 5.6 gives an example of two distinct signals that lead to the same sequence of samples. This example also shows that the condition for the sampling theorem to hold is *not* that $s_{\mathcal{F}}(f)$ and $s_{\mathcal{F}}(f-\frac{n}{T})$ do not overlap when $n\neq 0$. This condition is satisfied by the example of Fig. 5.6 yet the map that sends $s_{\mathcal{F}}(f)$ to $\tilde{s}_{\mathcal{F}}(f)$ is not invertible. Invertibility depends on the domain of the map, i.e. the length of $\mathcal{I}$, and not just on how the map acts on a single element of the domain.

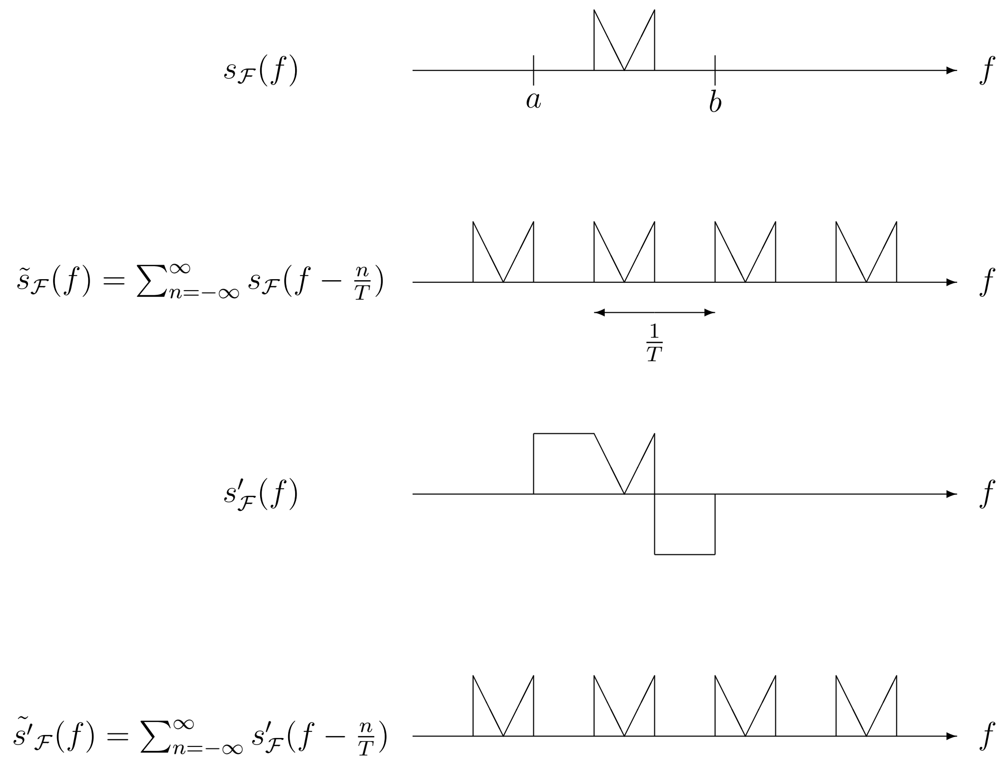

$s_{\mathcal{F}}(f)$
f
a
b
$\tilde{s}_{\mathcal{F}}(f) = \sum_{n=-\infty}^{\infty} s_{\mathcal{F}}(f - \frac{n}{T})$
f
$\frac{1}{T}$
$s'_{\mathcal{F}}(f)$
f
$\tilde{s'}_{\mathcal{F}}(f) = \sum_{n=-\infty}^{\infty} s'_{\mathcal{F}}(f - \frac{n}{T})$
f

Figure 5.6: Example of two distinct signals $s(t)$ and $s'(t)$ such that $s_{\mathcal{F}}(t)$ and $s'_{\mathcal{F}}(t)$ have domain in $\mathcal{I}$ and such that $\tilde{s}_{\mathcal{F}}(t) = \tilde{s'}_{\mathcal{F}}(t)$. When we sample the two signals at $t = nT$, n integer, we obtain the same sequence of samples.

Appendix 5.E Square-Root Raised-Cosine Pulse

We derive the inverse Fourier transform of the square-root raised-cosine pulse

$$\psi_{\mathcal{F}}(f) = \begin{cases} \sqrt{T}, & |f| \leq \frac{1-\beta}{2T} \\ \sqrt{T}\cos\frac{\pi T}{2\beta}(|f| - \frac{1-\beta}{2T}), & \frac{1-\beta}{2T} < |f| \leq \frac{1+\beta}{2T} \\ 0, & \text{otherwise.} \end{cases}$$

Write $\psi_{\mathcal{F}}(f) = a_{\mathcal{F}}(f) + b_{\mathcal{F}}(f)$ where $a_{\mathcal{F}}(f) = \sqrt{T}1_{[-\frac{1-\beta}{2T}, \frac{1-\beta}{2T}]}(f)$ is the central piece of the square-root raised-cosine pulse and $b_{\mathcal{F}}(f)$ accounts for the two square-root raised-cosine edges.

The inverse Fourier transform of $a_{\mathcal{F}}(f)$ is

$$a(t) = \frac{\sqrt{T}}{\pi t}\sin\left(\frac{\pi(1-\beta)t}{T}\right).$$

Write $b_{\mathcal{F}}(f) = b_{\mathcal{F}}^{-}(f) + b_{\mathcal{F}}^{+}(f)$, where $b_{\mathcal{F}}^{+}(f) = b_{\mathcal{F}}(f)$ for $f \geq 0$ and zero otherwise. Let $c_{\mathcal{F}}(f) = b_{\mathcal{F}}^{+}(f + \frac{1}{2T})$. Specifically,

$$c_{\mathcal{F}}(f) = \sqrt{T}\cos\left[\frac{\pi T}{2\beta}\left(f + \frac{\beta}{2T}\right)\right] 1_{[-\frac{\beta}{2T}, \frac{\beta}{2T}]}(f).$$

The inverse Fourier transform $c(t)$ is

$$c(t) = \frac{\beta}{2\sqrt{T}}\left[e^{-j\frac{\pi}{4}}\operatorname{sinc}\left(\frac{t\beta}{T} - \frac{1}{4}\right) + e^{j\frac{\pi}{4}}\operatorname{sinc}(\frac{t\beta}{T} + \frac{1}{4})\right].$$

Now we may use the relationship $b(t) = 2\Re\{c(t)e^{j2\pi\frac{1}{2T}t}\}$ to obtain

$$b(t) = \frac{\beta}{\sqrt{T}}\left[\operatorname{sinc}\left(\frac{t\beta}{T} - \frac{1}{4}\right)\cos\left(\frac{\pi t}{T} - \frac{\pi}{4}\right) + \operatorname{sinc}\left(\frac{t\beta}{T} + \frac{1}{4}\right)\cos\left(\frac{\pi t}{T} + \frac{\pi}{4}\right)\right].$$

After some manipulations of $\psi(f) = a(t) + b(t)$ we obtain the desired expression

$$\psi(t) = \frac{4\beta}{\pi\sqrt{T}}\frac{\cos\left((1+\beta)\pi\frac{t}{T}\right) + \frac{(1-\beta)\pi}{4\beta}\operatorname{sinc}\left((1-\beta)\frac{t}{T}\right)}{1 - \left(4\beta\frac{t}{T}\right)^2}.$$

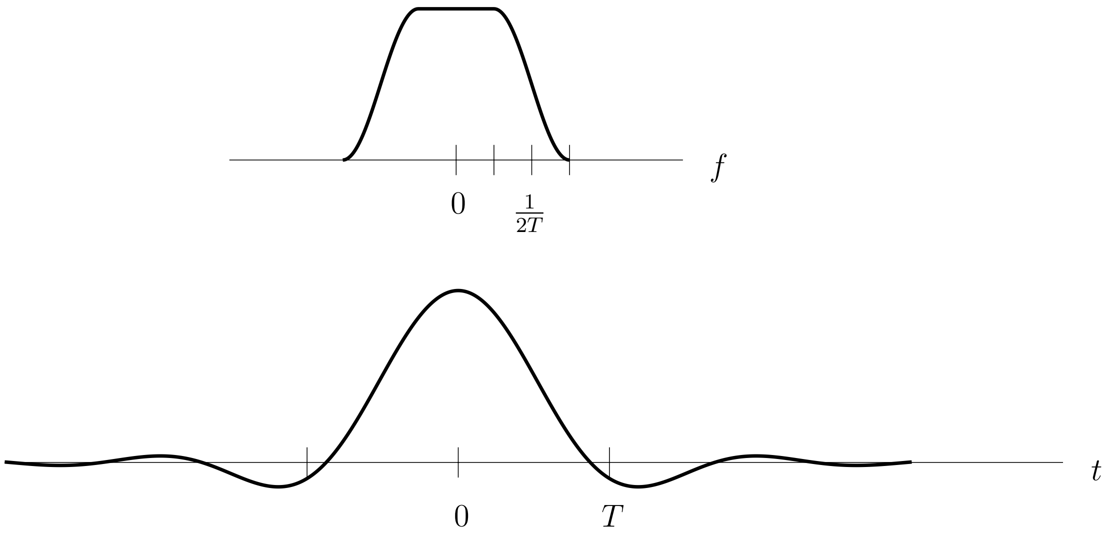


Figure 5.7: Raised cosine pulse $|\psi_{\mathcal{F}}(f)|^2$ with $\beta = \frac{1}{2}$ (top) and inverse Fourier transform $\psi(t)$ of the corresponding square-root raised-cosine pulse.

Appendix 5.F Problems

Problem 1. (Eye Diagram) *Consider the transmitted signal,* $S(t) = \sum_i X_i \psi(t - iT)$ *where* $\{X_i\}_{-\infty}^{\infty}$, $X_i \in \{\pm 1\}$, *is an i.i.d. sequence of random variables and* $\{\psi(t-iT)\}_{i=-\infty}^{\infty}$ *forms an orthonormal set. Let* $Y(t)$ *be the matched filter output at the receiver. In this MATLAB exercise we will try to see how crucial it is to sample at* $t = iT$ *as opposed to* $t = iT + \epsilon$. *Towards that goal we plot the so-called eye diagram. An eye diagram is the plot of* $Y(t+iT)$ *versus* $t \in [-\frac{T}{2}, \frac{T}{2}]$, *plotted on top of each other for each* $i = 0 \cdots K-1$, *for some integer* K. *Thus at* $t = 0$ *we see the superposition of several matched filter outputs when we output at the correct instant in time. At* $t = \epsilon$ *we see the superpositoin of several matched filter outputs when we sample at* $t = iT + \epsilon$.

(a) Assuming $K = 100$, $T = 1$ *and 10 samples per time period* T, *plot the eye diagrams when,*

(i) $\psi(t)$ *is a raised cosine with* $\alpha = 1$.

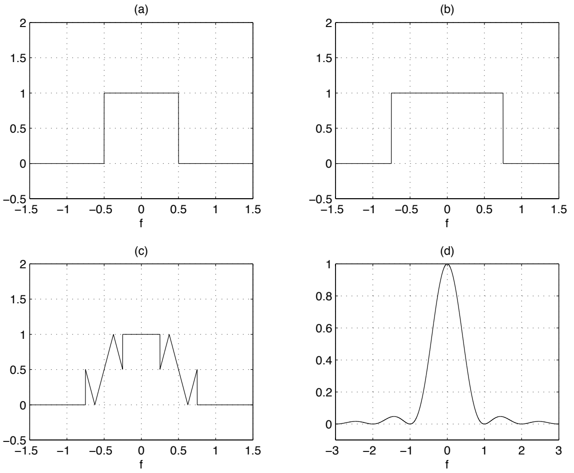
(ii) $\psi(t)$ *is a raised cosine with* $\alpha = \frac{1}{2}$.

(iii) $\psi(t)$ *is a raised cosine with* $\alpha = 0$. *This is a* sinc.

(b) From the plotted eye diagrams what can you say about the cruciality of the sampling points with respect to α?

Problem 2. (Nyquist Criterion)

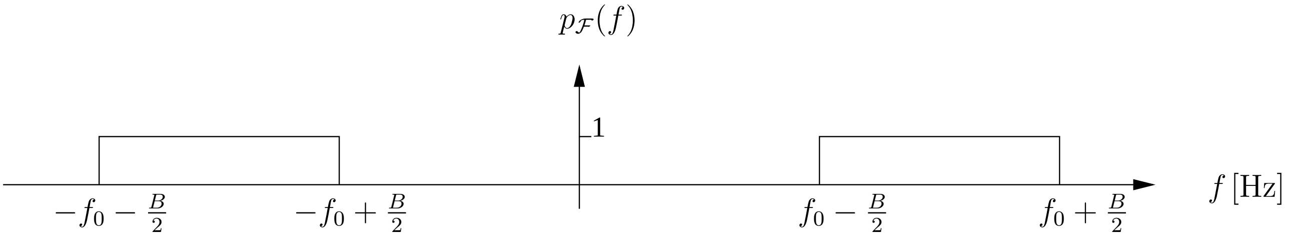
(a) *Consider the following* $|\theta_{\mathcal{F}}(f)|^2$. *The unit on the frequency axis is* $1/T$ *and the unit on the vertical axis is* T. *Which ones correspond to Nyquist pulses* $\theta(t)$ *for symbol rate* $1/T$*?* Note: *Figure* (d) *shows a* sinc^2 *function.*



(b) *Design a (non-trivial) Nyquist pulse yourself.*

(c) *Sketch the block diagram of a binary communication system that employs Nyquist pulses. Write out the formula for the signal after the matched filter. Explain the advantages of using Nyquist pulses.*

Problem 3. (Bandpass Nyquist Pulses) *Consider a pulse* $p(t)$ *defined via its Fourier transform* $p_{\mathcal{F}}(f)$ *as follows:*



(a) *What is the expression for* $p(t)$*?*

(b) *Determine the constant* c *so that* $\psi(t) = cp(t)$ *has unit energy.*

(c) *Assume that* $f_0 - \frac{B}{2} = B$ *and consider the infinite set of functions* $\cdots$, $\psi(t+T)$, $\psi(t)$, $\psi(t-T)$, $\psi(t-2T)$, $\cdots$. *Do they form an orthonormal set for* $T = \frac{1}{2B}$*? (Explain).*

(d) *Determine all possible values of* $f_0 - \frac{B}{2}$ *so that* $\cdots$, $\psi(t+T)$, $\psi(t)$, $\psi(t-T)$, $\psi(t-2T)$, $\cdots$ *forms an orthonormal set.*

Problem 4. (More on Nyquist Criterion) *Consider transmitting*

$$S(t) = \sum_{i=-\infty}^{\infty} X_i \psi(t - iT)$$

across an AWGN channel, where $\psi(t)$ is a Nyquist pulse. We know that an optimal thing to do at the receiver front end is to send the received signal $R(t)$ through the filter with impulse response $\psi^*(-t)$ and sample the filter output $Y(t)$ at time $t = iT$.

(a) *Show that, in absence of noise, the filter output $Y(iT)$ equals X_i.*

(b) *Now assume that you transmit $S(t) = \sum_{i=-\infty}^{\infty} X_i p(t-iT)$ and let the received signal through a filter of real-valued impulse response $q(t)$. You would like to retain the property that, in absence of noise, the filter output at time $t = iT$ be X_i. Show that this is equivalent to*

$$\int_{-\infty}^{\infty} p(kT + t) q(-t) dt \quad = \quad \delta(k).$$

(c) *Show that the equivalent condition in the frequency domain is*

$$\sum_{l=-\infty}^{\infty} p_{\mathcal{F}}(f - \frac{l}{T}) q^*_{\mathcal{F}}(f - \frac{l}{T}) \quad = \quad T \qquad \text{for} - \frac{1}{2T} \le f \le \frac{1}{2T}.$$

Problem 5. (Mixed Questions)

(a) *Consider the signal $x(t) = \cos(2\pi t) \left(\frac{\sin(\pi t)}{\pi t}\right)^2$. Assume that we sample $x(t)$ with sampling period T. What is the maximum T that guarantees signal recovery?*

(b) *Consider the three signals $s_1(t) = 1, s_2(t) = \cos(2\pi t), s_3(t) = \sin^2(\pi t)$, for $0 \le t \le 1$. What is the dimension of the signal space spanned by $\{s_1(t), s_2(t), s_3(t)\}$?*

(c) *You are given a pulse $p(t)$ with spectrum $p_{\mathcal{F}}(f) = T(1 - |f|T)$, $0 \le |f| \le \frac{1}{T}$. What is the value of $\int p(t)p(t - 3T)dt$?*

Chapter 6

Convolutional Coding and Viterbi Decoding

In Chapter 5 we have considered signals of the form $s(t) = \sum s_j \psi(t - jT)$, and we have focused on the Fourier domain characterization of those pulses $\psi(t)$ for which $\{\psi(t - jT)\}_{j=-\infty}^{\infty}$ forms an orthonormal set. In this chapter we focus on how to generate the symbol sequence s_j, $j = 1, 2, \ldots, n$. The sequence will be produced by a convolutional encoder. Some of the ideas presented in this chapter are best explained by means of a specific example. We do that at the expense of generality, but the reader should have no difficulty in applying the same techniques to other convolutional encoders.

The receiver will implement the Viterbi algorithm (VA)—a neat and clever way to decode efficiently in many circumstances. To analyze the bit-error probability we will introduce a few new tools, notably detour flow graphs and generating functions.

The signals that we will construct will have several desirable properties: The transmitter and the receiver adapt in a natural way to the number k of bits that need to be communicated; the duration of the transmission grows linearly with the number of bits; the symbol sequence is uncorrelated, implying that the power spectral density of the transmitted signal is $\mathcal{E}_s \frac{|\psi_{\mathcal{F}}(f)|^2}{T}$; the encoding complexity is linear in k and so is the complexity of the maximum likelihood receiver; for the same energy per bit, the bit error probability is much smaller than that of bit by bit on a pulse train. The convolutional encoder studied in this chapter produces two binary symbols per source bit, which implies that the bit rate [bits/sec] is half that of bit by bit on a pulse train. This is the price we pay to reduce the bit error probability.

6.1 The Transmitter

Like in bit by bit on a pulse train, the transmitted signal has the form

$$s(t) = \sum_{j=1}^{n} s_j \psi(t - jT),$$

with

$$\begin{aligned} s_j &= x_j \sqrt{\mathcal{E}_s} \\ x_j &\in \{\pm 1\}. \end{aligned}$$

However, x_j is now the jth element of the encoder output sequence rather than being the jth bit produced by the source.

Notice that there is a slight change in notation with respect to earlier chapters. Up until now the transmitted signal $s(t)$ had an index i, mainly since our starting point was the signal constellation specified by a list $s_1(t), \ldots, s_{m-1}(t)$ of generic signals. (It was a result of Chapter 3 to show that those signals can always be written in the form $s_i(t) = \sum s_{i,j} \psi_j(t)$.) In this chapter we start with k bits that are mapped to an $n$tuple $\boldsymbol{s} = (s_1, \ldots, s_n)$ that leads to $s(t)$. The fact that there are 2^k such signals is implied by the setup and adding an index i to $s(t)$ seems redundant. In some cases, like in the next section, it will be useful to have such an index. We will use the index as needed. Hopefully the going back and forth between the two notations will not create any confusion.

The k bits produced by the source will be denoted by $(d_1, d_2, \ldots, d_k)$, $d_j \in \{\pm 1\}$. At regular intervals there is one source bit entering the encoder and two binary symbols coming out. During the jth interval, $j = 1, 2, \ldots, k$, the encoder produces x_{2j-1}, x_{2j} according to the *encoding map*

$$\begin{aligned} x_{2j-1} &= d_j d_{j-2} \\ x_{2j} &= d_j d_{j-1} d_{j-2}, \end{aligned}$$

with $d_{-1} = d_0 = 1$ set as initial condition. The convolutional encoder is depicted in Figure 6.1. Notice that the encoder output has length $n = 2k$. The following example shows a source output sequence of length $k = 5$ and the corresponding encoder output sequence of length $n = 10$.

d_j	1	-1	-1	1	1
x_{2j-1}, x_{2j}	$1, 1$	$-1, -1$	$-1, 1$	$-1, 1$	$-1, -1$
j	1	2	3	4	5

Since the $n = 2k$ encoder output symbols are determined by the k input bits, only 2^k of the 2^n n-length binary sequences are codewords. This means that we are using only a fraction $2^{k-n} = 2^{-k}$ of all possible n-length binary sequences. Intuitively, we are giving

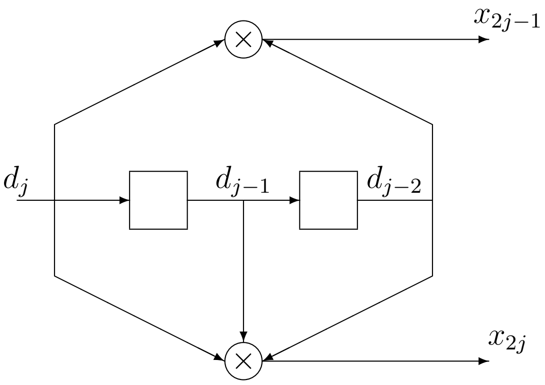


Figure 6.1: Rate $\frac{1}{2}$ convolutional encoder.

up a factor 2 in the bit rate to make the signal space much less crowded, hoping that this will significantly reduce the probability of error. There are several ways to describe a convolutional encoder. We have already seen that we can specify it by the encoding map and by the *encoding circuit* of Figure 6.1 . A third way, which will turn out to be useful in determining the error probability, is via the *state diagram* of Figure 6.2. The diagram describes a *finite state machine*. The state of the convolutional encoder is what the encoder needs to know about past inputs so that together with the current input it can determine the current output. For the convolutional encoder of Figure 6.1 the state at time j may be defined to be d_{j-1}, d_{j-2}. Hence we have 4 states represented by a box in the state diagram. As the diagram shows, there are two possible transitions from each state. The input symbol d_j decides which of the two possible transitions is taken at time j. Transitions are labeled by $d_j | x_{2j-1}, x_{2j}$. Throughout the chapter we assume that the encoder is in state $1, 1$ when the first symbol enters the encoder.

6.2 The Receiver

Let $\|s_i\|^2 = \sum_{j=1}^{n} \mathcal{E}_s x_{i,j}^2 = n\mathcal{E}_s$ be the signal's energy (the same for all signals).

A maximum likelihood (ML) decoder decides for one of the i that maximizes

$$\langle r, s_i \rangle - \frac{n\mathcal{E}_s}{2},$$

where r is the received signal and the second term is irrelevant since it does not depend on i.

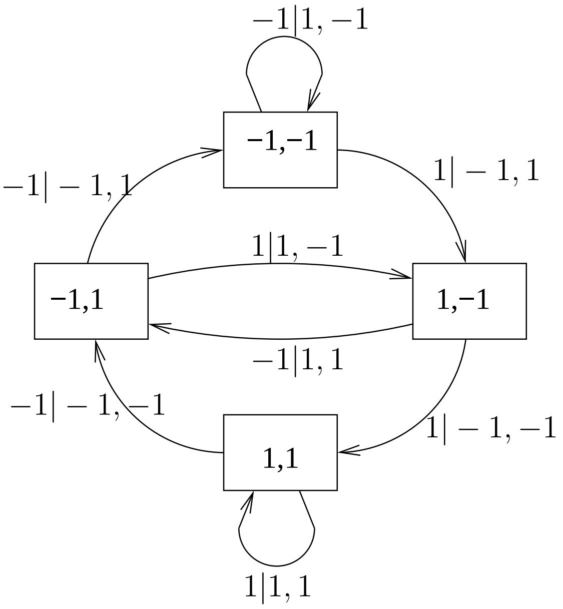


Figure 6.2: State diagram description of the convolutional encoder.

Hence a ML decoder picks a sequence $s_{i,1}, \dots, s_{i,n}$ that maximizes

$$\begin{aligned}
&\int r(t) \sum_{j=1}^{n} s_{i,j} \psi^*(t - jT) dt \\
&= \sum_{j=1}^{n} s_{i,j} \int r(t) \psi^*(t - jT) dt \\
&= \sum_{j=1}^{n} s_{i,j} y_j \\
&= \sqrt{\mathcal{E}_s} \sum_{j=1}^{n} x_{i,j} y_j \\
&= \sqrt{\mathcal{E}_s} \langle \boldsymbol{x}_i, \boldsymbol{y} \rangle
\end{aligned}$$

where we have defined

$$y_j = \int r(t) \psi^*(t - jT) dt.$$

Recall that y_j is the output at time jT of the filter with impulse response $\psi^*(-t)$ and input $r(t)$.

To find the $\boldsymbol{x}_i$ that maximizes $\langle \boldsymbol{x}_i, \boldsymbol{y} \rangle$, one could in principle compute $\langle \boldsymbol{x}, \boldsymbol{y} \rangle$ for all 2^k sequences that can be produced by the encoder. This *brute-force* approach would be quite unpractical. For instance, if $k = 100$ (which is a relatively modest value for k), $2^k = (2^{10})^{10}$ which is approximately $(10^3)^{10} = 10^{30}$. Using this approximation, a VLSI chip that makes 10^9 inner products per second takes 10^{21} seconds to check all possibilities. This is roughly $4\,10^{13}$ years. The universe is only $2\,10^{10}$ years old!

What we need is a method that finds the maximum $\langle \boldsymbol{x}, \boldsymbol{y} \rangle$ by making a number of operations that grows linearly (as opposed to exponentially) in k. By cleverly making use of the encoder structure we will see that this can be done. The result is the Viterbi algorithm.

To describe the Viterbi algorithm (VA) we introduce a fourth way of describing a convolutional encoder, namely the *trellis*. The trellis is an unfolded transition diagram that keeps track of the passage of time. It is obtained by making as many replicas of the states as there are input symbols. For our example, if we assume that we start at state $1, 1$, that we encode $k = 5$ source bits and then feed the encoder with $1, 1$ to make it go back to the initial state and thus be ready for the next transmission, we obtain the trellis description shown on the top of Figure 6.3. There is a one to one correspondence between a message $i \in \{0, 1, \ldots, 2^k - 1\}$, an encoder input sequence $\boldsymbol{d}_i$, an encoder output sequence $\boldsymbol{x}_i$, and a path (or state sequence) from the initial state $(1, 1)_0$ (very left state) to the final state $(1, 1)_{k+2}$ (very right state) of the trellis. Hence in the discussion that follows we may refer to a path by means of an input sequence, an output sequence or a sequence of states.

The trellis on the top of Figure 6.3 has edges labeled by the corresponding output symbols. To decode using the Viterbi algorithm we replace the label of each edge with the edge metric (also called branch metric) defined as follows. Let $\Gamma = \{(1, 1), (1, -1), (-1, 1), (-1, -1)\}$ be the state space. If there is an edge that connects state $\alpha \in \Gamma$ at depth $j - 1$ to state $\beta \in \Gamma$ at depth j, we let the *edge metric* $\mu_{j-1,j}(\alpha, \beta)$ be

$$\mu_{j-1,j}(\alpha, \beta) = x_{2j-1} y_{2j-1} + x_{2j} y_{2j},$$

wherex_{2j-1}, x_{2j} is the encoder output of the corresponding edge. If there is no such edge we let $\mu_{j-1,j}(\alpha, \beta) = -\infty$. Notice that $\mu_{j-1,j}(\alpha, \beta)$ is the jth term in $\langle \boldsymbol{x}, \boldsymbol{y} \rangle$ for any path that goes through state α at depth $j-1$ and state β at depth j. Hence $\langle \boldsymbol{x}, \boldsymbol{y} \rangle$ is obtained by adding the edge metrics along the path specified by $\boldsymbol{x}$. The second subfigure of Figure 6.3 shows the trellis labeled with the edge metric that corresponds to the hypothetical $\boldsymbol{y} = (1, 3), (-2, 1), (4, -1), (5, 5), (-3, -3), (1, -6), (2, -4)$, where parentheses have been inserted to facilitate parsing.

The *path metric* is the sum of the edge metrics taken along the edges of a path. A *longest path* from state $1, 1$ at depth $j = 0$, denoted $(1, 1)_0$, to a state α at depth j, denoted α_j, is one of the paths that has the largest path metric. The Viterbi algorithm works by constructing, for each j, a list of the longest paths to the states at depth j. The following observation is key to understand the Viterbi algorithm. If $\mathit{path} * \alpha_{j-1} * \beta_j$ is a longest path to state β of depth j, where $\mathit{path} \in \Gamma^{j-2}$ and $*$ denotes concatenation, then $\mathit{path} * \alpha_{j-1}$ must be a longest path to state α of depth $j-1$, for if another path, say $\mathit{alternatepath} * \alpha_{j-1}$ were shorter for some $\mathit{alternatepath} \in \Gamma^{j-2}$, then $\mathit{alternatepath} * \alpha_{j-1} * \beta_j$ would be shorter than $\mathit{path} * \alpha_{j-1} * \beta_j$. So the longest depth j path to a state can be obtained by checking the extension of the longest depth $(j - 1)$ paths by one branch.

The following notation is useful for the formal description of the Viterbi algorithm. Let $\mu_j(\alpha)$ be the metric of a longest path to state α_j and let $B_j(\alpha) \in \{\pm 1\}^j$ be the encoder input sequence that corresponds to this path. We call $B_j(\alpha) \in \{\pm 1\}^j$ the *survivor* since it is the only paths through state α_j that will be extended. (Paths through α_j that have smaller metric have no chance of extending into a maximum likelihood path). For each state the Viterbi algorithms computes two things, a survivor and its metric. The formal algorithm follows, where $B(\beta, \alpha)$ is the encoder input that corresponds to the transition from state β to state α if there is such a transition and is undefined otherwise.

```
1.  Initially set $\mu_0(1,1) = 0$, $\mu_0(\alpha) = -\infty$ for all $\alpha \neq (1,1)$,
    $B_0(1,1) = \emptyset$, and $j = 1$.

2.  For each $\alpha \in \Gamma$, find one of the $\beta$ for which
    $\mu_{j-1}(\alpha) + \mu_{j-1,j}(\beta, \alpha)$ is a maximum.  Then set

        $\mu_j(\alpha) \leftarrow \mu_{j-1}(\alpha) + \mu_{j-1,j}(\beta, \alpha),$
        $B_j(\alpha) \leftarrow B_{j-1}(\alpha) * B(\beta, \alpha).$

3.  If $j = k + 2$, output the first $k$ bits of $B_j(1,1)$ and
    stop.  Otherwise increment $j$ by one and go to Step 2.
```

The reader should have no difficulty verifying (by induction on j) that $\mu_j(\alpha)$ as computed by Viterbi's algorithm is indeed the metric of a longest path from $(1,1)_0$ to state α at depth j and that $B_j(\alpha)$ is the encoder input sequence associated to it.

The third subfigure of Figure 6.3 shows the computation that one does to mimic the Viterbi algorithm by hand. Each state α at depth j is labeled with a survivor's metric $\mu_j(\alpha)$. We keep track of a survivor to state α_j by "pruning" all the other edges to that state. (In this example there is only one other edge.) A "pruned" edge is one that has been dashed. Once we have reached state $(1,1)_{k+2}$ we can backtrack and follow the only surviving path (bottom figure). To make it possible to write down the source sequence of a path without requiring additional labels, we have positioned the states in such a way that the upper edge out of a state is taken when the input is -1 and the lower edge when the input is 1.

The complexity of the Viterbi algorithm is linear in the number of trellis vertices. It is also linear in the number k of source bits. Recall that the brute force approach had complexity exponential in k. The saving of the Viterbi algorithm comes from not having to compute the metric of non-survivors. When we prune an edge at depth j we are in fact eliminating 2^{k-j} possible extension of that edge. The brute force approach computes the metric of those extensions but not the Viterbi algorithm.

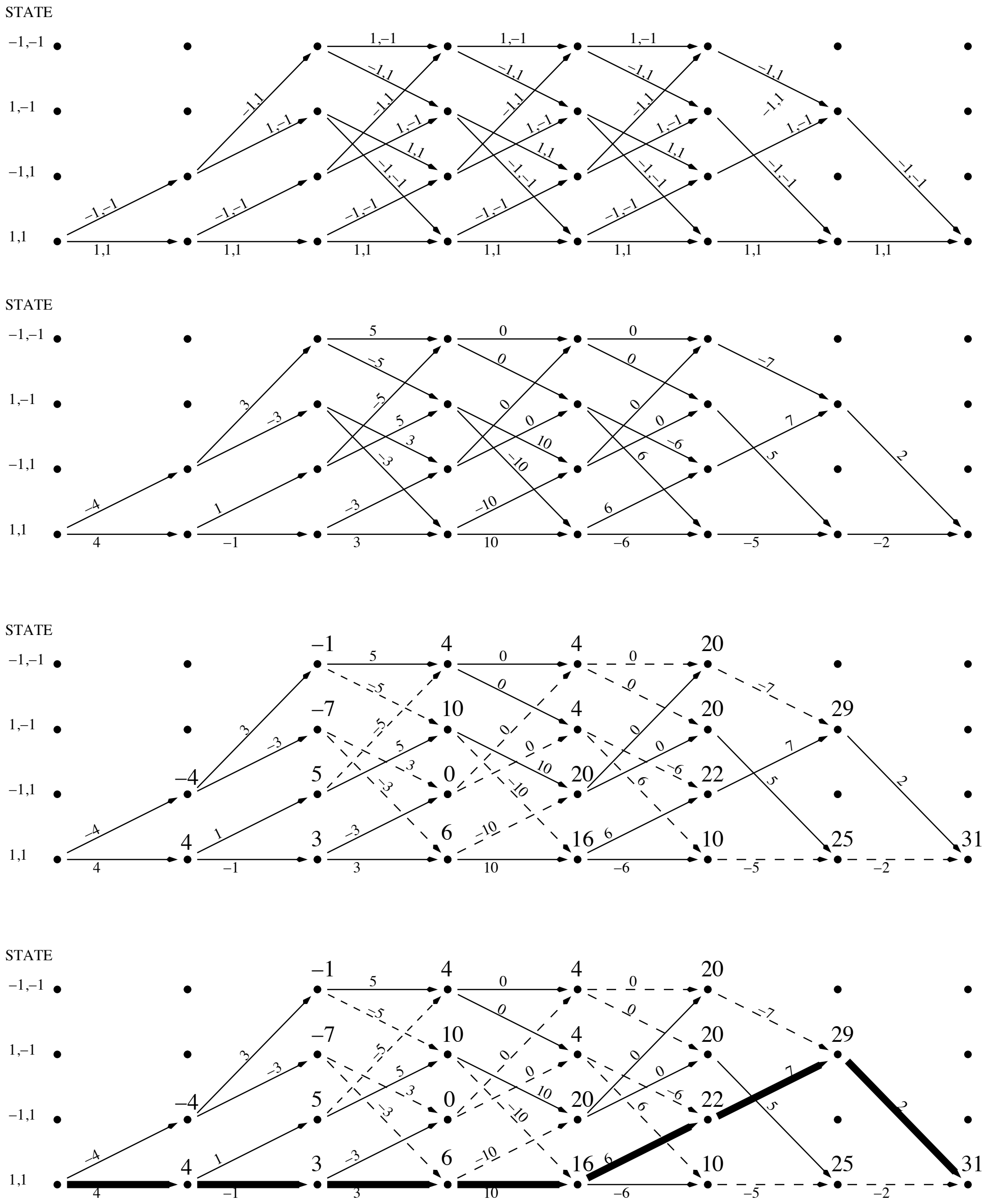


Figure 6.3: The Viterbi algorithm. Top figure: Trellis representing the encoder. The upper edge leaving a state corresponds to source symbol -1, the lower edge to source symbol 1. Edges are labeled with the corresponding output symbols; Second figure: Edges have been re-labeled with the edge metric corresponding to the received sequence $(1,3),(-2,1),(4,-1),(5,5),(-3,-3),(1,-6),(2,-4)$ (parentheses have been inserted to facilitate parsing); Third figure: Each state has been labeled with the metric of a survivor to that state and non-surviving edges are pruned (dashed); Fourth figure: Tracing back from the end we find the decoded path (bold). It corresponds to the source sequence $1,1,1,1,-1,1,1$.

6.3 Bit-Error Probability

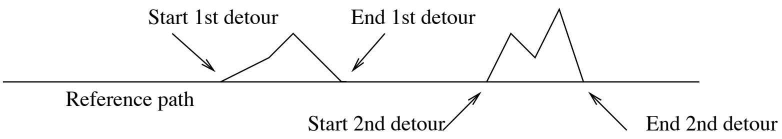
We assume that the initial state is $1,1$ and we transmit a number k of bits.

As we have done so far, we determine (an upper bound to) the probability of error by conditioning on a transmitted signal. It will turn out that our expression does not depend on the transmitted signal.

Each signal that can be produced by the transmitter corresponds to a path in the trellis. The path we are conditioning on when we compute (or when we bound) the bit error probability will be called the *reference* path.

The task of the decoder is to find (one of) the paths in the trellis that has the largest $\langle \boldsymbol{x}, \boldsymbol{y} \rangle$ where $\boldsymbol{x}$ is the encoder output that corresponds to that path. If the decoder does not come up with the correct path it is because it chooses a path that contains one or more detour(s).

Detours (with respect to the reference path) are trellis path segments that share with the reference path only the starting and the ending state.[1] (See the figure below.)



Errors are produced when the decoder follows a detour and that is the only way to produce errors. To compute the bit error probability we study the random process produced by the decoder when it chooses the maximum likelihood trellis path. Each such path is either the correct path or else it brakes down in some number of detours. To the path selected by the decoder we associate a sequence $w_0, w_1, \ldots, w_{k-1}$ defined as follows. If there is a detour that starts at depth j we let w_j be the number of bit errors produced by *that* detour. In all other cases we let $w_j = 0$. Then $\frac{1}{k}\sum_{j=0}^{k-1} w_j$ is the fraction of errors produced by the decoder. Over the ensemble of all possible noise processes, w_j becomes a random variable W_j and the bit error probability is

$$P_b \triangleq E\frac{1}{k}\left[\sum_{j=0}^{k-1} W_j\right].$$

In the next section we learn everything we need to know about detours to be able to upper bound the above expression.

[1] For an analogy, the reader may think of the trellis as a road map, of the reference path as of an intended road for a journey, and the path selected by the decoder as of the actual road that was taken during the journey. Due to constructions, occasionally the actual path deviates from the intended path to merge again with it at some later point. A detour is the chunk of road from the deviation to the merge.

6.3.1 Counting Detours

In this subsection we consider infinite trellises, i.e., trellises that are extended to infinity on both sides. Each path in such a trellis corresponds to an infinite input sequence $\boldsymbol{d} = \dots d_{-1}, d_0, d_1, d_2, \dots$ and infinite output sequence $\boldsymbol{x} = \dots x_{-1}, x_0, x_1, x_2, \dots$. These are sequences that belong to $\mathcal{D}^* = \{\pm 1\}^\infty$.

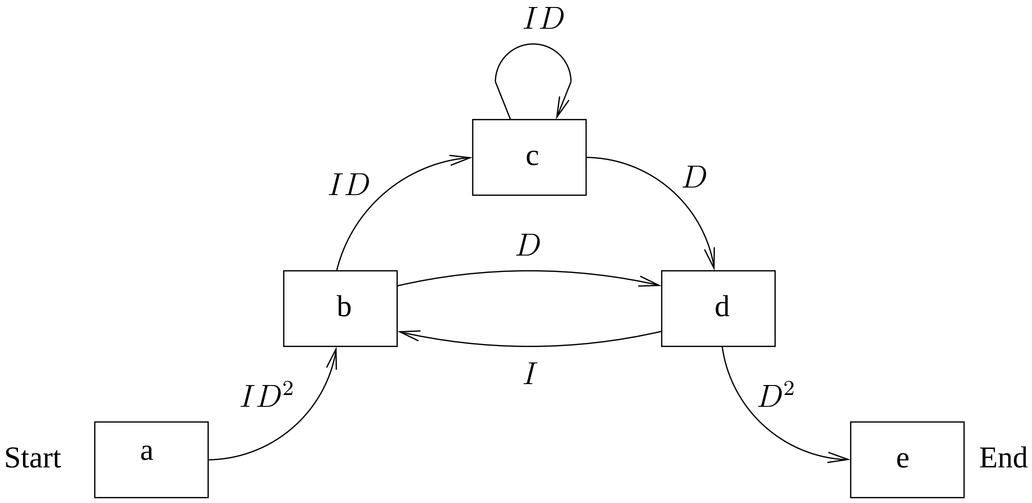
Let $\boldsymbol{d}$ and $\boldsymbol{x}$ be the input and output sequence, respectively, associated to the reference path and let $\tilde{\boldsymbol{d}}$ and $\tilde{\boldsymbol{x}}$, also in $\mathcal{D}^*$, be an alternative input and the associated output. The reference and the alternative path give rise to one or more detours. To each such detour we can associate two numbers, namely the *input distance* i and the *output distance* d. The input distance is the number of position in which the two input sequences differ over the course of the detour. The output distance counts the discrepancies of the output sequences over the course of the detour.

Example 63. *Using the top subfigure of Figure 6.3 as an example, if we take the all one path (the bottom path) as the reference and consider the detour that splits at depth* 0 *and merges back with the reference path at depth* 3, *we see that this detour has input distance* $i = 1$ *and output distance* $d = 5$.

The question that we address in this subsection is: for any given reference path $\boldsymbol{x}$ and integer j representing a trellis depth, what is the number $a(i,d)$ of detours that start at depth j and have input distance i and output distance d with respect to the reference path? We will see that this number depends neither on the reference path nor on j.

Example 64. *Using again the top subfigure of Figure 6.3 we can verify by inspection that for each reference path and each positive integer* j *there is a single detour that starts at depth* j *and has parameters* $i = 1$ *and* $d = 5$. *Thus* $a(1,5) = 1$. *We can also verify that* $a(2,5) = 0$ *and* $a(2,6) = 2$. □

We start by assuming that the reference path is the *all-one* path. This is the path generated by the all-one source symbols. The corresponding encoder output sequence also consists of all ones. For every j, there is a one-to-one correspondence between a detour that starts at depth j and a path between state a and states e of the following *detour flow graph* obtained from the state diagram by removing the self-loop of state $(1,1)$ and splitting this state into a starting state denoted by a and an ending state denoted by e.



The label I^iD^d, (i and d nonnegative integers), on an edge of the detour flow graph indicates that the input and output distances increase by i and d, respectively, when the detour takes that edge.

Now we show how to determine $a(i,d)$. In terms of the detour flow graph, $a(i,d)$ is the number of paths between a and e that have path label I^iD^d, where the label of a path is the product of all labels along that path. We will actually determine the *generating function* $T(I,D)$ of $a(i,d)$ defined as

$$T(I,D) = \sum_{i,d} I^iD^d a(i,d).$$

The letters I and D in the above expression should be seen as "place holders" without any physical meaning. It is like describing a set of coefficients $a_0, a_1, \ldots, a_{n-1}$ by means of the polynomial $p(x) = a_0 + a_1x + \ldots + a_{n-1}x^{n-1}$. To determine $T(I,D)$ we introduce auxiliary generating functions, one for each intermediate state of the detour flow graph, namely

$$T_b(I,D) = \sum_{i,d} I^iD^d a_b(i,d) \tag{6.1}$$

$$T_c(I,D) = \sum_{i,d} I^iD^d a_c(i,d) \tag{6.2}$$

$$T_d(I,D) = \sum_{i,d} I^iD^d a_d(i,d), \tag{6.3}$$

where in the fist line we have defined $a_b(i,d)$ as the number of paths in the detour flow graph that start at state a, end at state b, and have path label I^iD^d. Similarly, for $x = c, d$, $a_x(i,d)$ is the number of paths in the detour flow graph that start at state a, end at state x, and have path label I^iD^d. (In (6.3), the d of a_d is a fixed label not to be confused with the summation variable.)

From the detour flow graph we see that the following relationships holds. (To simplify the notation we are writing T_x instead of $T_x(I, D)$, $x = b, c, d$ and write T instead of $T(I, D)$)

$$\begin{aligned}
T_b &= ID^2 + T_d\, I \\
T_c &= T_b\, ID + T_c\, ID \\
T_d &= T_b\, D + T_c\, D \\
T &= T_d\, D^2.
\end{aligned}$$

The above system may be solved for T by pure formal manipulations. (Like solving a system of equations). The result is

$$T(I, D) = \frac{ID^5}{1 - 2ID}.$$

As we will see shortly, the generating function $T(I, D)$ of $a(i, d)$ is actually more useful than $a(i, d)$ itself. However, to show that one can indeed obtain $a(i, d)$ from $T(I, D)$ we use the expansion $\frac{1}{1-x} = 1 + x + x^2 + x^3 + \cdots$ to write

$$\begin{aligned}
T(I, D) = \frac{ID^5}{1 - 2ID} &= ID^5(1 + 2ID + (2ID)^2 + (2ID)^3 + \cdots \\
&= ID^5 + 2I^2D^6 + 2^2I^3D^7 + 2^3I^4D^8 + \cdots
\end{aligned}$$

This means that there is one path with parameters $d = 5$, $i = 1$, that there are two paths with $d = 6$, $i = 2$, etc. The general expression for $i = 1, 2, \ldots$ is

$$a(i, d) = \begin{cases} 2^{i-1}, & d = i + 4 \\ 0, & \text{otherwise.} \end{cases}$$

The correctness of this expression may be verified by means of the detour flow graph.

In this final paragraph we argue that $a(i, d)$ depends neither on the reference path, assumed so far to be the all-one path, nor on the starting depth. The reader willing to accept this fact may skip to the next section. To prove the claim, pick an arbitrary reference path described by the corresponding input sequence, say $\bar{\boldsymbol{d}} \in \mathcal{D}^*$. Let f be the input/output map, i.e., $f(\bar{\boldsymbol{d}})$ is the encoder output resulting from the input $\bar{\boldsymbol{d}}$. It is not hard to verify (see Problem 6) that f is a linear map in the sense that if $\boldsymbol{d} \in \mathcal{D}^*$ is an arbitrary input sequence then $f(\boldsymbol{d}\bar{\boldsymbol{d}}) = f(\boldsymbol{d})(\bar{\boldsymbol{d}})$, where products are understood componentwise. Let $\boldsymbol{e} \in \mathcal{D}^*$ be such that the path specified by the input $\bar{\boldsymbol{d}}\boldsymbol{e}$ has a detour that starts at depth j and has input distance i with respect to the reference path specified by $\bar{\boldsymbol{d}}$. Notice that the positions of $\boldsymbol{e}$ that contain -1 are precisely the positions where $\bar{\boldsymbol{d}}\boldsymbol{e}$ differs from $\bar{\boldsymbol{d}}$ and those positions determine j and i. Hence j and i are determined by $\boldsymbol{e}$ and are independent of $\bar{\boldsymbol{d}}$. Also the output distance d depends only on $\boldsymbol{e}$. Indeed it depends on the positions where $f(\bar{\boldsymbol{d}}\boldsymbol{e})$ differs from $f(\bar{\boldsymbol{d}})$ which are the positions where $f(\bar{\boldsymbol{d}}\boldsymbol{e})f(\bar{\boldsymbol{d}})$ is -1. Due to linearity, $f(\bar{\boldsymbol{d}}\boldsymbol{e})f(\bar{\boldsymbol{d}}) = f(\bar{\boldsymbol{d}})f(\boldsymbol{e})f(\bar{\boldsymbol{d}}) = f(\boldsymbol{e})$. We conclude that, independently of the reference path, the set of $\boldsymbol{e} \in \mathcal{D}^*$ that identify in the above way those detours that start at j and have input distance i and output distance d is the same regardless of the reference path. The size of that set is $a(i, d)$.

6.3.2 Upper Bound to P_b

We are now ready for the final step, namely the derivation of an upper bound to the bit-error probability. We recapitulate.

Fix an arbitrary encoder input sequence, let $\boldsymbol{x} = x_1, x_2 \ldots, x_n$ be the corresponding encoder output sequence and $\boldsymbol{s} = \sqrt{\mathcal{E}_s}\boldsymbol{x}$ be the vector signal. The waveform signal is

$$s(t) = \sum_{j=1}^{n} s_j \psi(t - jT).$$

We transmit this signal over an AWGN channel with power spectral density $N_0/2$. Let $r(t) = s(t) + z(t)$ be the received signal (where $z(t)$ is a sample path of the noise process $Z(t)$) and let

$$\boldsymbol{y} = (y_1, \ldots, y_n)^T, \quad y_i = \langle r, \psi_i \rangle$$

be a sufficient statistic.

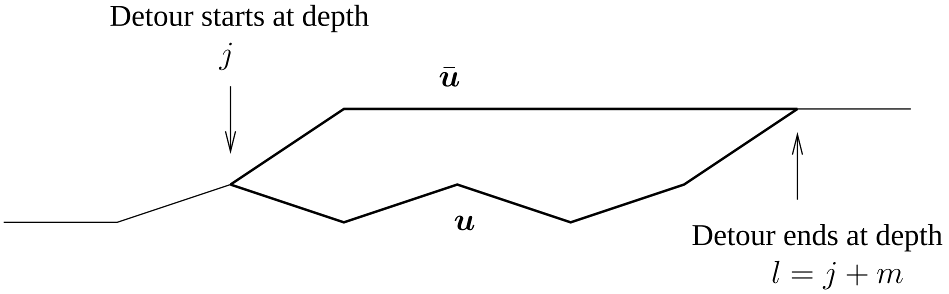
The Viterbi algorithm labels each edge in the trellis with the corresponding edge metric and finds the path through the trellis with the largest path metric. An edge from depth $j-1$ to j with output symbols x_{2j-1}, x_{2j} is assigned the edge metric $y_{2j-1}x_{2j-1} + y_{2j}x_{2j}$.

The ML path selected by the Viterbi decoder may contain several detours. Let w_j, $j = 0, 1, \ldots, k-1$, be the number of bit errors made on a detour that *begins* at depth j. If at depth j the VD is on the correct path or if it follows a detour started earlier then $w_j = 0$. Let W_j be the corresponding random variable (over all possible noise realizations).

For the path selected by the VD the total number of incorrect bits is $\sum_{j=0}^{k-1} w_j$ and $\frac{1}{k}\sum_{j=0}^{k-1} w_j$ is the fraction of errors with respect to the k source bits. Hence we define the bit-error probability

$$P_b \triangleq E\frac{1}{k}\left[\sum_{j=0}^{k-1} W_j\right] = \frac{1}{k}\sum_{j=0}^{k-1} EW_j. \tag{6.4}$$

Let us now focus on a detour. If it starts at depth j and ends at depth $l = j + m$, then the corresponding encoder-output symbols form a $2m$ tuple $\bar{\boldsymbol{u}} \in \{\pm 1\}^{2m}$. Let $\boldsymbol{u} = (x_{2j+1}, \ldots, x_{2l}) \in \{\pm 1\}^{2m}$ be the corresponding sub-sequence of the *reference* path and $\boldsymbol{\rho} = (y_{2j+1}, \ldots, y_{2l})$ the corresponding channel output subsequence.



Let d be the Hamming distance $d_H(\boldsymbol{u}, \bar{\boldsymbol{u}})$ between $\boldsymbol{u}$ and $\bar{\boldsymbol{u}}$. The Euclidean distance between the corresponding waveforms is the distance between $\sqrt{\mathcal{E}_s}\boldsymbol{u}$ and $\sqrt{\mathcal{E}_s}\bar{\boldsymbol{u}}$ which is $d_E = 2\sqrt{\mathcal{E}_s d}$.

A necessary (but not sufficient) condition for the Viterbi decoder to take the detour under consideration is that the path metric along the detour be equal or larger that of the corresponding segment along the reference path, i.e.,

$$\langle \boldsymbol{\rho}, \sqrt{\mathcal{E}_s}\boldsymbol{u}\rangle \leq \langle \boldsymbol{\rho}, \sqrt{\mathcal{E}_s}\bar{\boldsymbol{u}}\rangle.$$

This condition is satisfied iff

$$\|\boldsymbol{\rho} - \sqrt{\mathcal{E}_s}\boldsymbol{u}\|^2 \geq \|\boldsymbol{\rho} - \sqrt{\mathcal{E}_s}\bar{\boldsymbol{u}}\|^2.$$

The probability of the above event is the probability that a ML receiver for the discrete-time AWGN channel makes an error when the correct vector signal is $\sqrt{\mathcal{E}_s}\boldsymbol{u}$ and the alternative signal is $\sqrt{\mathcal{E}_s}\bar{\boldsymbol{u}}$. This probability is

$$Q\left(\frac{d_E}{2\sigma}\right) = Q\left(\sqrt{\frac{2\mathcal{E}_s d}{N_0}}\right) \leq \exp\left\{-\frac{\mathcal{E}_s d}{N_0}\right\} = z^d, \tag{6.5}$$

where $\sigma^2 = \frac{N_0}{2}$, $d_E = \|\sqrt{\mathcal{E}_s}\boldsymbol{u} - \sqrt{\mathcal{E}_s}\bar{\boldsymbol{u}}\| = 2\sqrt{\mathcal{E}_s d}$, and we have defined $z = \exp\left\{-\frac{\mathcal{E}_s}{N_0}\right\}$.

We are ready for the final steps towards upper bounding P_b.

$$EW_j = \sum_{\text{all detours } h} i(h)\pi(h)$$

where the sum is over all detours that start at depth j with respect to the all-one sequence, $\pi(h)$ stands for the probability that the detour h is taken, and $i(h)$ for the input distance between detour h and the all-one path. Using $\pi(h) \leq z^{d(h)}$ where $d(h)$ stands for the output distance between the detour h and the all-one path we obtain

$$\begin{aligned} EW_j &\leq \sum_{\text{all detours } h} i(h) z^{d(h)} \\ &= \sum_{i=1}^{k}\sum_{d=1}^{2k} i z^d \tilde{a}(i,d) \\ &\leq \sum_{i=1}^{\infty}\sum_{d=0}^{\infty} i z^d a(i,d), \end{aligned}$$

where in the second line we have grouped the terms of the sum that have the same i and d and used $\tilde{a}(i,d)$ to denote the number of such terms. Notice that $\tilde{a}(i,d)$ pertains to the finite trellis and may be smaller than the corresponding number $a(i,d)$ associated to the semi-infinite trellis. This justifies the last line's inequality. Using the relationship

$$\sum_{i=1}^{\infty} i f(i) = \frac{\partial}{\partial I}\sum_{i=0}^{\infty} I^i f(i)\Big|_{I=1},$$

which holds for any function f, we may write

$$\begin{aligned} EW_j &\leq \frac{\partial}{\partial I} \sum_{i=1}^{\infty} \sum_{d=0}^{\infty} I^i z^d a(i,d) \bigg|_{I=1} \\ &= \frac{\partial}{\partial I} T(I,D) \bigg|_{I=1, D=z}. \end{aligned}$$

We may plug into (6.4) and use the fact that the bound on EW_j does not depend on j to obtain

$$P_b = \frac{1}{k} \sum_{j=0}^{k-1} EW_j \leq \frac{\partial}{\partial I} T(I,D) \bigg|_{I=1, D=z}. \tag{6.6}$$

In our specific example $T(D,I) = \frac{ID^5}{1-2ID}$ and $\frac{\partial T}{\partial I} = \frac{D^5}{(1-2ID)^2}$. Thus

$$P_b \leq \frac{z^5}{(1-2z)^2},$$

where $z = \exp\{-\frac{\mathcal{E}_s}{N_0}\}$ and $\mathcal{E}_s = \frac{\mathcal{E}_b}{2}$ (we are transmitting two channel symbols per information bit).

Notice that in (6.6) the channels characteristic is summarized by the parameter z. Specifically, z^d is an upper bound to the probability that a maximum likelihood receiver makes a decoding error when the choice is between two binary codewords of Hamming distance d. As shown in Problem 29(ii) of Chapter 2, we may use the Bhattacharyya bound to determine z for any binary-input discrete memoryless channel. For such a channel,

$$z = \sum_y \sqrt{P(y|a)P(y|b)}$$

where a and b are the two letters of the input alphabet and y runs over all elements of the output alphabet. We have also assumed that in one tick the encoder takes $k_0 = 1$ source symbols and outputs $n_0 = 2$ channel symbols. It is straightforward to generalize the above derivation to any k_0 and any n_0 in $\mathbb{N}_+$. Details are left as an exercise.

6.4 Concluding Remarks

What have we done and how does it compare to what we have done before? It is convenient to think of the bit by bit on a pulse train as our starting point.

$$s(t) = \sum_{i=0}^{n-1} s_i \psi(t - iT)$$

$$s_i \in \{\pm \sqrt{\mathcal{E}_s}\}.$$

The relevant design choices for this system are (the subscripts s and b stand for symbol and bit, respectively):

$$
\begin{aligned}
&R_b = R_s = \frac{1}{T} \quad \text{bit rate}\\
&\mathcal{E}_b = \mathcal{E}_s \quad \text{energy per bit}\\
&P_b = Q\left(\frac{\sqrt{\mathcal{E}_s}}{\sigma}\right), \quad \text{bit-error probability.}
\end{aligned}
$$

Using $\sigma = \sqrt{\frac{N_0}{2}}$ and the upper bound $Q(x) \le \exp\big\{-\frac{x^2}{2}\big\}$ we obtain

$$P_b \le \exp\Big\{-\frac{\mathcal{E}_b}{N_0}\Big\}$$

which will be useful in comparing with the coded case.

The added value of this chapter was to have an encoder between the source and the transmitter that implements bit by bit on a pulse train. The encoder trades the bit error probability P_b for the bit rate R_b. The new parameters are:

$$
\begin{aligned}
&R_b = \frac{R_s}{2} = \frac{1}{2T_s}\\
&\mathcal{E}_b = 2\mathcal{E}_s\\
&P_b \le \frac{z^5}{(1-2z)^2} \text{ where } z = \exp\Big\{-\frac{\mathcal{E}_b}{2N_0}\Big\}.
\end{aligned}
$$

As $\frac{\mathcal{E}_b}{2N_0}$ becomes large, the denominator of the above bound for P_b becomes essentially 1 and the bound decreases as z^5. The bound for the uncoded case is z^2. As already mentioned, the price for the decreased bit error rate is the fact that we are sending two symbols per bit.

In both cases the expression for the power spectral density is $\frac{\mathcal{E}_s}{T}|\psi_{\mathcal{F}}(f)|^2$ but since the coded case has twice as many symbols, the energy per symbol $\mathcal{E}_s$ of the coded case is halved. Hence the bandwidth is the same in both cases but not the power spectral density. Since coding reduces the bit-rate by a factor 2, the bandwidth efficiency, defined as the number of bits per second per Hz, is smaller by a factor of 2 in the coded case. With more powerful codes we can further decrease the bit error probability without affecting the bandwidth efficiency.

The reader may wonder how the code we have considered as an example compares to bit by bit on a pulse train with each bit sent out twice. In this case the jth bit is sent as $d_j\sqrt{\mathcal{E}_s}\psi(t-(2j-1)T) + d_j\sqrt{\mathcal{E}_s}\psi(t-2jT)$ which is the same as sending $d_j\sqrt{\mathcal{E}_b}\tilde{\psi}(t-2jT)$ with $\tilde{\psi}(t) = (\psi(t)+\psi(t-T))/\sqrt{2}$. Hence this scheme is again bit by bit on a pulse train with the new pulse $\tilde{\psi}(t)$ used every $2T$ seconds. We know that the bit error probability of bit by bit on a pulse train does not depend on the pulse used. Thus this scheme uses twice the bandwidth of bit by bit on a pulse train with no benefit in terms of bit error probability.

From a "high level" point of view, coding is about exploiting the advantages of working in a higher dimensional signal space rather than making multiple independent uses of a small number of dimensions. (Bit by bit on a pulse train uses one dimension at a time.) In n dimensions, we send some $\boldsymbol{s} \in \mathbb{R}^n$ and receive $\boldsymbol{Y} = \boldsymbol{s} + \boldsymbol{Z}$, where $\boldsymbol{Z} \sim \mathcal{N}(0, I_n\sigma^2)$. By the law of large numbers, $\sqrt{(\sum z_i^2)/n}$ goes to σ as n goes to infinity. This means that with probability approaching 1, the received $n$tuple $\boldsymbol{Y}$ will be in a thin shell of radius $\sqrt{n}\sigma$ around $\boldsymbol{s}$. This phenomenon is referred to as *sphere hardening*. As n becomes large, the space occupied by $\boldsymbol{Z}$ becomes more predictable and in fact it becomes a small fraction of the entire $\mathbb{R}^n$. Hence there is hope that one can find many vector signals that are distinguishable (with high probability) even after the Gaussian noise has been added. Information theory tells us that we can make the probability of error go to zero as n goes to infinity provided that we use fewer than $m = 2^{nC}$ signals, where $C = \frac{1}{2}\log_2(1 + \frac{\mathcal{E}_s}{\sigma^2})$. It also teaches us that the probability of error can not be made arbitrarily small if we use more than 2^{nC} signals. Since $(\log_2 m)/n$ is the number of bits per dimension that we are sending when we use m signals embedded in an n dimensional space, it quite appropriate to call C [bits/dimension] the *capacity* of the discrete-time additive white Gaussian noise channel. For the continuous-time AWGN channel of total bandwidth B, the channel capacity is

$$B \log_2 \left(1 + \frac{P}{BN_0}\right) \text{ [bits/sec]},$$

where P is the transmitted power. It can be achieved with signals of the form $s(t) = \sum_j s_j \psi(t - jT)$.

6.5 Problems

Problem 1. (Power Spectral Density) *Block-orthogonal signaling may be the simplest coding method that achieves $Pr\{e\} \to 0$ as $N \to \infty$ for a non-zero data rate. However, we have seen in class that the price to pay is that block-orthogonal signaling requires infinite bandwidth to make $Pr\{e\} \to 0$. This may be a small problem for one space explorer communicating to another; however, for terrestrial applications, there are always constraints on the bandwidth consumption. Therefore, in the examination of any coding method, an important issue is to compute its bandwidth consumption. Compute the bandwidth occupied by the rate$-1/2$ convolutional code studied in this chapter. The signal that is put onto the channel is given by*

$$X(t) = \sum_{i=-\infty}^{\infty} X_i \sqrt{E_s}\psi(t - iT_s), \tag{6.7}$$

where $\psi(t)$ is some unit-energy function of duration T_s and we assume that the trellis extends to infinity on both ends, but as usual we actually assume that the signal is the wide-sense stationary signal

$$\tilde{X}(t) = \sum_{i=-\infty}^{\infty} X_i \sqrt{E_s}\psi(t - iT_s - T_0), \tag{6.8}$$

where T_0 is a random delay which is uniformly distributed over the interval $[0, T_s)$.

(a) *Find the expectation $E[X_i X_j]$ for $i = j$, for $(i,j) = (2n, 2n+1)$ and for $(i,j) = (2n, 2n+2)$ for the convolutional code that was studied in class. Then give the autocorrelation function $R_X[i-j] = E[X_i X_j]$ for all i and j.* Hint: *Consider the infinite trellis of the code. Recall that the convolution code studied in the class can be defined as*

$$\begin{aligned} X_{2n} &= D_n D_{n-2} \\ X_{2n+1} &= D_n D_{n-1} D_{n-2} \end{aligned}$$

(b) *Find the autocorrelation function of the signal $\tilde{X}(t)$, that is*

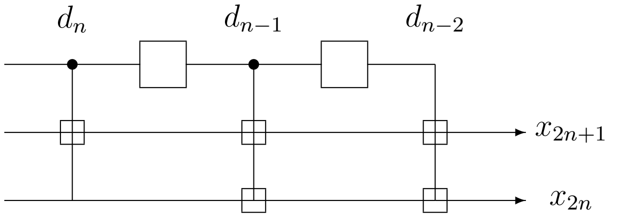
$$R_{\tilde{X}}(\tau) = E[\tilde{X}(t)\tilde{X}(t+\tau)] \tag{6.9}$$

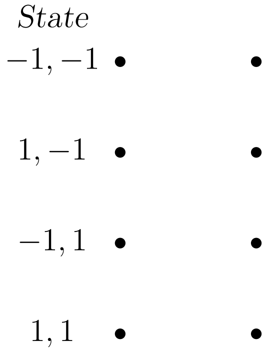
in terms of $R_X[k]$ and $R_\psi(\tau) = \frac{1}{T_s}\int_0^{T_s} \psi(t+\tau)\psi(t)dt$.

(c) *Give the expression of power spectral density of the signal $\tilde{X}(t)$.*

(d) *Find and plot the power spectral density that results when $\psi(t)$ is a rectangular pulse of width T_s centered at 0.*

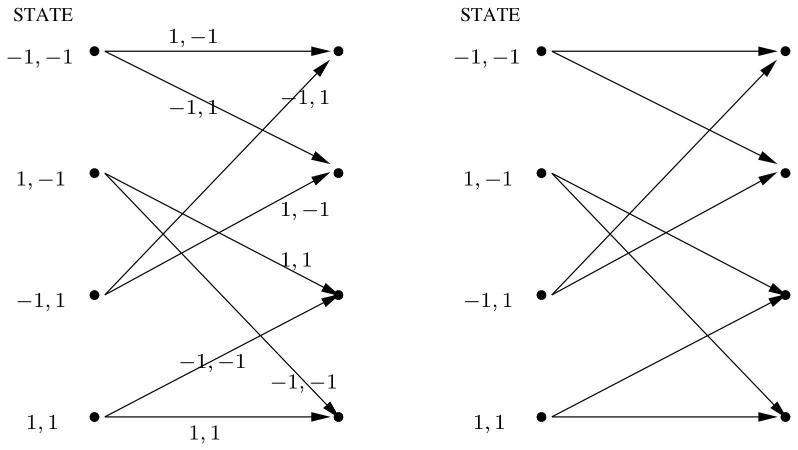
Problem 2. (Trellis Section) *For the convolutional encoder shown below on the left, fill in the section of the trellis shown below on the right, that is, find the correct arrows and label them with the corresponding output value pairs (x_{2n}, x_{2n+1}). The input sequence d_n takes values in $\{\pm 1\}$ and the outputs fulfill the relationships $x_{3n} = d_n d_{n-2}$; $x_{3n+1} = d_{n-1}d_{n-2}$; $x_{3n+2} = d_n d_{n-1} d_{n-2}$.*





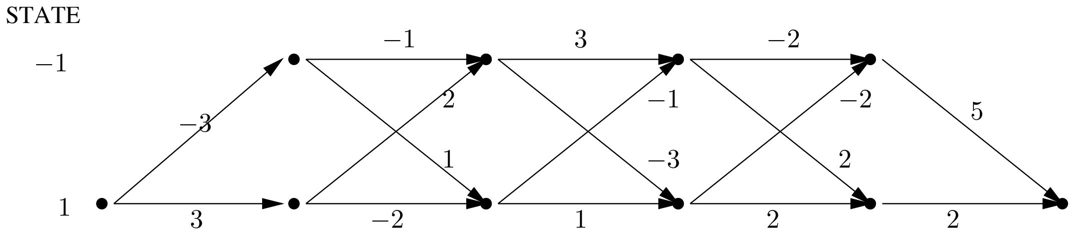
Problem 3. (Branch Metric) *Consider the convolutional code described by the trellis section below on the left. You may assume that each of the encoder output symbols*

(x_{2n}, x_{2n+1}), are mapped into orthogonal waveforms, $\phi_1(t)$ if $x_i = +1$ and $\phi_2(t)$ if $x_i = -1$. The waveforms are of equal energy E_s. At the receiver we perform matched filtering with the filters matched to $\phi_1(t)$ and $\phi_2(t)$. Suppose the output of the matched filter at time n are $(y_{2n}, y_{2n+1}) = (1, -2)$. Find the branch metric values to be used by the Viterbi algorithm and enter them into the trellis section on the right.



Problem 4. (Viterbi Algorithm)

In the trellis below, the received sequence has already been preprocessed. The labels on the branches of the trellis are the branch metric values. Find the maximum likelihood path.



Problem 5. (Intersymbol Interference) *An information sequence $\underline{U} = (U_1, U_2, \dots, U_5)$, $U_i \in \{0, 1\}$ is transmitted over a noisy intersymbol interference channel. The i th sample of the receiver-front-end filter (e.g. a filter matched to the pulse used by the sender)*

$$Y_i = S_i + Z_i,$$

where the noise Z_i forms an independent and identically distributed (i.i.d.) sequence of Gaussian random variables,

$$S_i = \sum_{j=0}^{\infty} U_{i-j} h_j, \qquad i = 1, 2, \dots$$

and

$$h_i = \begin{cases} 1, & i = 0 \\ -2, & i = 1 \\ 0, & \textit{otherwise.} \end{cases}$$

You may assume that $U_i = 0$ *for* $i \geq 6$ *and* $i \leq 0$.

(a) *Rewrite* S_i *in a form that explicitly shows by which symbols of the information sequence it is affected.*

(b) *Sketch a trellis representation of a finite state machine that produces the output sequence* $\underline{S} = (S_1, S_2, \ldots, S_6)$ *from the input sequence* $\underline{U} = (U_1, U_2, \ldots, U_5)$. *Label each trellis transition with the specific value of* $U_i | S_i$.

(c) *Specify a metric* $f(\underline{s}, \underline{y}) = \sum_{i=1}^{6} f(s_i, y_i)$ *whose minimization or maximization with respect to* $\underline{s}$ *leads to a maximum likelihood decision on* $\underline{S}$. *Specify if your metric needs to be minimized or maximized.* Hint: *Think of a vector channel* $\underline{Y} = \underline{S} + \underline{Z}$, *where* $\underline{Z} = (Z_1, \ldots, Z_6)$ *is a sequence of i.i.d. components with* $Z_i \sim \mathcal{N}(0, \sigma^2)$.

(d) *Assume* $\underline{Y} = (Y_1, Y_2, \cdots, Y_5, Y_6) = (2, 0, -1, 1, 0, -1)$. *Find the maximum likelihood estimate of the information sequence* $\underline{U}$. Please: *Do* not *write into the trellis that you have drawn in Part* (b)*; work on a copy of that trellis.*

Problem 6. (Linear Transformations)

(a) (i) *First review the notion of a field. (See e.g. K. Hoffman and R. Kunze,* Linear Algebra, *Prentice Hall or your favorite linear algebra book.)*

Now consider the set $\mathcal{F} = \{0, 1\}$ *with the following addition and multiplication tables:*

+	0	1
0	0	1
1	1	0

×	0	1
0	0	0
1	0	1

Does $\mathcal{F}$, "+", *and* "×" *form a field?*

(ii) *Repeat using* $\mathcal{F} = \{\pm 1\}$ *and the following addition and multiplication tables:*

+	1	−1
1	1	−1
−1	−1	1

×	1	−1
1	1	1
−1	1	−1

(b) (i) *Now first review the notion of a vector space. Let* $\mathcal{F}$, + *and* × *be as defined in* (i)(a). *Let* $\mathcal{V} = \mathcal{F}^\infty$. *(The latter is the set of infinite sequences with components in* $\mathcal{F}$. *Does* $\mathcal{V}$, $\mathcal{F}$, + *and* × *form a vector space?*

(ii) *Repeat using* $\mathcal{F}$, + *and* × *as in* (i)(b).

(c) (i) Review the concept of linear transformation from a vector space $\mathcal{I}$ to a vector space $\mathcal{O}$. Now let $f : \mathcal{I} \to \mathcal{O}$ be the mapping implemented by the encoder described in this chapter. Specifically, for $j = 1, 2, \dots$, let $\boldsymbol{x} = f(\boldsymbol{d})$ be specified by

$$\begin{aligned} x_{2j-1} &= d_{j-1} \oplus d_{j-2} \oplus d_{j-3} \\ x_{2j} &= d_j \oplus d_{j-2}, \end{aligned}$$

where by convention we let $d_0 = d_{-1} = 0$. Show that this $f : \mathcal{I} \to \mathcal{O}$ is linear.

Problem 7. (Rate 1/3 Convolutional Code.) *Consider the following convolutional code, to be used for the transmission of some information sequence $d_i \in \{-1, 1\}$:*

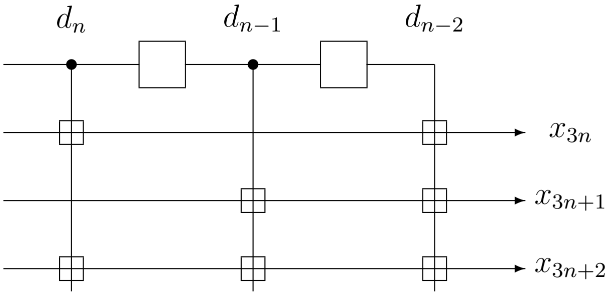


Figure 6.4: Convolutional encoder. $x_{3n} = d_n d_{n-2}$; $x_{3n+1} = d_{n-1} d_{n-2}$; $x_{3n+2} = d_n d_{n-1} d_{n-2}$.

(a) Draw the state diagram for this encoder.

(b) Suppose that this code is decoded using the Viterbi algorithm. Draw the detour flowgraph.

(c) This encoder/decoder is used on an AWGN channel. The energy available per source digit is E_b and the power spectral density of the noise is $N_0/2$. Give an upper bound on the bit error probability P_b as a function of E_b/N_0.

Problem 8. (Convolutional Code.) *The following equations define a convolutional code for a data sequence $d_i \in \{-1, 1\}$:*

$$\begin{aligned} x_{3n} &= d_{2n} \cdot d_{2n-1} \cdot d_{2n-2} && (6.10) \\ x_{3n+1} &= d_{2n+1} \cdot d_{2n-2} && (6.11) \\ x_{3n+2} &= d_{2n+1} \cdot d_{2n} \cdot d_{2n-2} && (6.12) \end{aligned}$$

(a) *Draw an implementation of the encoder of this convolutional code, using only delay elements D and multipliers.* Hint: *Split the data sequence d into two sequences, one containing only the even-indexed samples, the other containing only the odd-indexed samples.*

(b) *What is the rate of this convolutional code?*

(c) *Draw the state diagram for this convolutional encoder.*

(d) *Does the formula for the upper bound on P_b that was derived in class still hold? If not, make the appropriate changes.*

(e) *(optional) Now suppose that the code is used on an AWGN channel. The energy available per source digit is E_b and the power spectral density of the noise is $N_0/2$. Give the detour flowgraph, and derive an upper bound on the bit error probability P_b as a function of E_b/N_0.*

PROBLEM 9. (PSD of a Basic Encoder) *Consider the transmitter shown in Figure 6.5, when $\ldots D_{-i}, D_i, D_{i+1}, \ldots$ is a sequence of independent and uniformly distributed random variables taking value in $\{\pm 1\}$.*

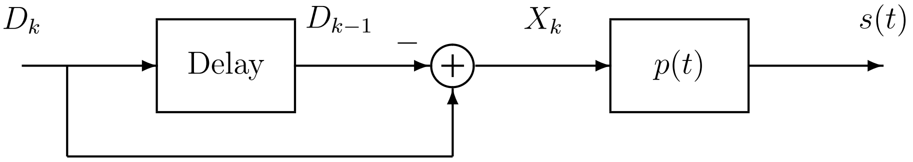


Figure 6.5: Encoder

The transmitted signal is

$$s(t) = \sum_{i=-\infty}^{\infty} X_i p(t - iT - \Theta),$$

where Θ is a random variable, uniformly distributed in $[0, T]$.

$$X_i = D_i - D_{i-1}$$

$$p(t) = 1_{[-\frac{T}{2}, \frac{T}{2}]}(t).$$

(a) *Determine $R_X[k] = E[X_{i+k}X_i]$.*

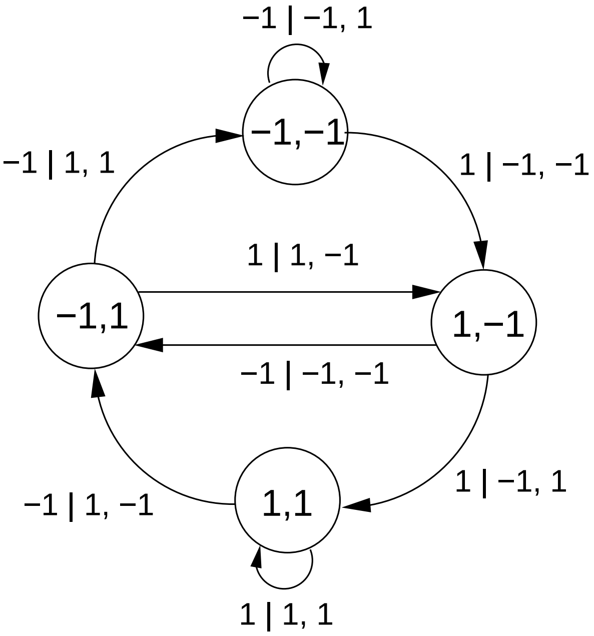
(b) *Determine $R_p(\tau) = \int_{-\infty}^{\infty} p(t+\tau)p(t)dt$.*

(c) *Determine the autocorrelation function $R_s(\tau)$ of the signal $s(t)$.*

(d) *Determine the power spectral density* $S_s(f)$.

PROBLEM 10. (Convolutional Encoder, Decoder and Error Probability) *Consider a channel, where a transmitter wants to send a sequence* $\{D_j\}$ *taking values in* $\{-1,+1\}$, *for* $j = 0, 1, 2, \cdots, k-1$. *This sequence is encoded using a convolutional encoder. The channel adds white Gaussian noise to the transmitted signal. If we let* X_j *denote the transmitted value, then, the received value is:* $Y_j = X_j + Z_j$, *where* $\{Z_j\}$ *is a sequence of i.i.d. zero-mean Gaussian random variables with variance* $\frac{N_0}{2}$. *The receiver has to decide which sequence was transmitted using the optimal decoding rule.*

(a) *Convolutional Encoder Consider the convolutional encoder corresponding to the finite state machine drawn below. The transitions are labeled by* $D_j|X_{2j}, X_{2j+1}$, *and the states by* D_{j-1}, D_{j-2}. *We assume that the initial content of the memory is* $(1,1)$.



(i) *What is the rate of this encoder?*

(ii) *Sketch the filter (composed of shift registers and multipliers) corresponding to this finite state machine. How many shift registers do you need?*

(iii) *Draw a section of the trellis representing this encoder.*

(b) *Viterbi Decoder Let* X_j^i *denote the output of the convolutional encoder at time* j *when we transmit hypothesis* i, $i = 0, \cdots, m-1$, *where* m *is the number of different hypotheses.*

Assume that the received vector is $\bar{Y} = (Y_1, Y_2, Y_3, Y_4, Y_5, Y_6) = (-1, -3, -2, 0, 2, 3)$. *It is the task of the receiver to decide which hypothesis* i *was chosen or, equivalently, which vector* $\bar{X}^i = (X_1^i, X_2^i, X_3^i, X_4^i, X_5^i, X_6^i)$ *was transmitted.*

(i) *Use the Viterbi algorithm to find the most probable transmitted vector* $\bar{X}^i$.

(c) *Performance Analysis*

(i) *Suppose that this code is decoded using the Viterbi algorithm. Draw the detour flow graph, and label the edges by the input weight using the symbol* I, *and the output weight using the symbol* D.

(ii) *Considering the following generating function*

$$T(I,D) = \frac{ID^4}{1-3ID},$$

What is the value of

$$\sum_{i,d} i a(i,d) e^{-\frac{d}{2N_0}},$$

where $a(i,d)$ *is the number of detours with* i *bit errors and* d *channel errors? First compute this expression, then give an interpretation in terms of probability of error of this quantity.*

Hints: Recall that the generating function is defined as $T(I,D) = \sum_{i,d} a(i,d) D^d I^i$. You may also use the formula $\sum_{k=1}^{\infty} k q^{k-1} = \frac{1}{(1-q)^2}$ if $|q| < 1$.

(Viterbi Decoding in Both Directions)

Consider the trellis diagram given in the top figure of Fig. 6.2 in the class notes. Assume the received sequence at the output of the matched filter is

$$(y_1, \ldots, y_{14}) = (-2, -1, 2, 0, 1, -1, 2, -3, -5, -5, 2, -1, 3, 2).$$

(a) Run the Viterbi algorithm from left to right. Show the decoded path on the trellis diagram.

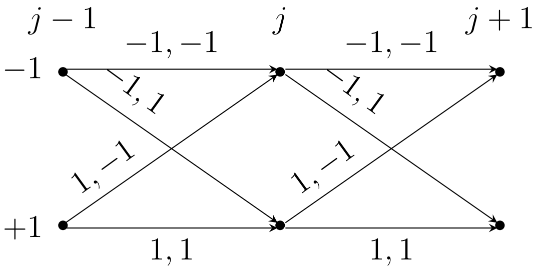
(b) If you run the Viterbi algorithm from right to left instead of left to right, do you think you will get the same answer for the decoded sequence? Why?

(c) Now, run the Viterbi algorithm from left to right only for the observations $(y_1, \ldots, y_6)$. (Do not use the diagram in part-(a), draw a new one.) On the same diagram also run the Viterbi algorithm from right to left for $(y_9, \ldots, y_{14})$. How can you combine the two results to find the maximum likelihood path?

Problem 11. (Trellis with Antipodal Signals) *Assume that the sequence* $X_1, X_2, \dots$ *is sent over an additive white Gaussian noise channel,* i.e.,

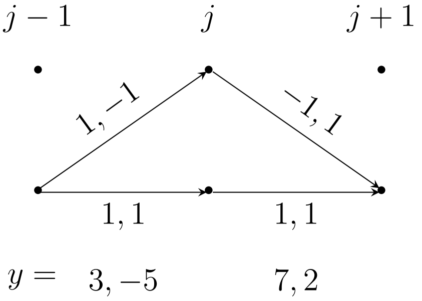
$$Y_i = X_i + Z_i,$$

where the Z_i *are i.i.d. zero-mean Gaussian random variables with variance* σ^2. *The sequence* X_i *is the output of a convolutional encoder described by the following trellis.*

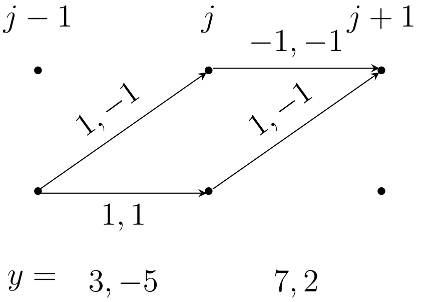


As the figure shows, the trellis has two states labeled with $+1$ *and* -1, *respectively. The probability assigned to each of the two branches leaving any given state is* $1/2$. *The trellis is also labeled with the output produced when a branch is traversed and with the trellis depths* $j-1$, j, $j+1$.

(a) *Consider the two paths in the following picture. Which of the two paths is more likely if the corresponding channel output subsequence* $y_{2j-1}, y_{2j}, y_{2j+1}, y_{2(j+1)}$ *is* $3, -5, 7, 2$?

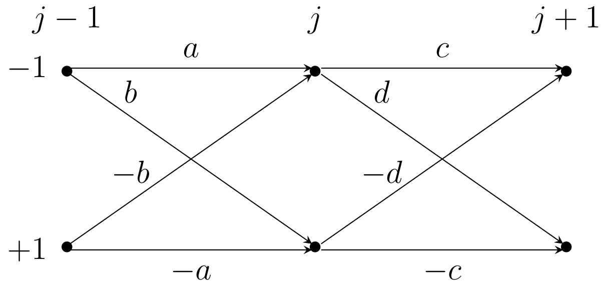


(b) *Now, consider the following two paths with the same channel output as in the previous question. Find again the most likely of the two paths.*



(c) If you have made no mistake in the previous two questions, the state at depth j of the most likely paths is the same in both cases. This is no coincidence as we will now prove.

The first step is to remark that the metric has to be as in the following picture for some value of a, b, c, and d.



(a) Now let us denote by $\sigma_k \in \{\pm 1\}$ the state at depth k, $k = 0, 1, \cdots$, of the maximum likelihood path. Assume that a genie tells you that $\sigma_{j-1} = 1$ and $\sigma_{j+1} = 1$. In terms of a, b, c, d, write down a necessary condition for $\sigma_j = 1$. (The condition is also sufficient up to ties.)

(b) Now assume that $\sigma_{j-1} = 1$ and $\sigma_{j+1} = -1$. What is the condition for choosing $\sigma_j = 1$?

(c) Now assume that $\sigma_{j-1} = -1$ and $\sigma_{j+1} = 1$. What is the condition for $\sigma_j = 1$?

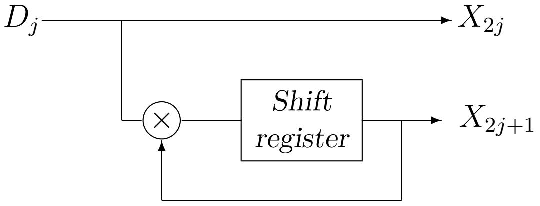
(d) Now assume that $\sigma_{j-1} = -1$ and $\sigma_{j+1} = -1$. What is the condition for $\sigma_j = 1$?

(e) Are the four conditions equivalent? Justify your answer.

(f) Comment on the advantage, if any, implied by your answer to part (v) of question (c).

Problem 12. (Convolutional Code: Complete Analysis)

(a) Convolutional Encoder Consider the following convolutional encoder. The input sequence D_j takes values in $\{-1, +1\}$ for $j = 0, 1, 2, \cdots, k-1$. The output sequence, call it X_j, $j = 0, \cdots, 2k-1$, is the result of passing D_j through the filter shown below, where we assume that the initial content of the memory is 1.



(i) *In the case $k = 3$, how many different hypotheses can the transmitter send using the input sequence (D_0, D_1, D_2), call this number m.*

(ii) *Draw the finite state machine corresponding to this encoder. Label the transitions with the corresponding input and output bits. How many states does this finite state machine have?*

(iii) *Draw a section of the trellis representing this encoder.*

(iv) *What is the rate of this encoder?*
(number of information bits /number of transmitted bits).

(b) *Viterbi Decoder Consider the channel defined by $Y_j = X_j^i + Z_j$. Let X_j^i denote the ouput of the convolutional encoder at time j when we transmit hypothesis i, $i = 0, \cdots, m-1$. Further, assume that Z_j is a zero-mean Gaussian random variable with variance $\sigma^2 = 4$ and let Y_j be the output of the channel.*

Assume that the received vector is $\bar{Y} = (Y_1, Y_2, Y_3, Y_4, Y_5, Y_6) = (1, 2, -2, -1, 0, 3)$. It is the task of the receiver to decide which hypothesis i was chosen or, equivalently, which vector $\bar{X}^i = (X_1^i, X_2^i, X_3^i, X_4^i, X_5^i, X_6^i)$ was transmitted.

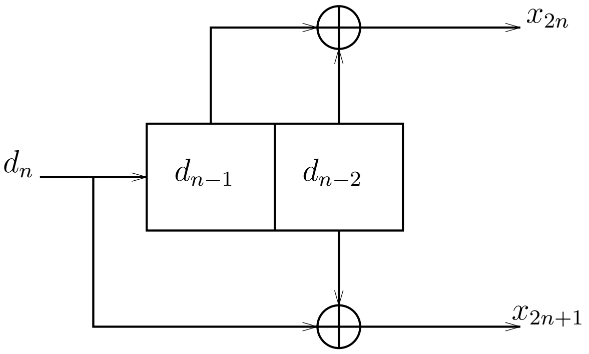
(i) *Without using the Viterbi algorithm, write formally (in terms of $\bar{Y}$ and $\bar{X}^i$) the optimal decision rule. Can you simplify this rule to express it as a function of inner products of vectors? In that case, how many inner products do you have to compute to find the optimal decision?*

(ii) *Use the Viterbi algorithm to find the most probable transmitted vector $\bar{X}^i$.*

(c) *Performance Analysis.*

(i) *Draw the detour flow graph corresponding to this decoder and label the edges by the input weight using the symbol I, the output weight (of both branches) using the symbol D.*

PROBLEM 13. (Viterbi for the Binary Erasure Channel) *Consider the following convolutional encoder. The input sequence belongs to the binary alphabet* $\{0,1\}$. *(This means we are using XOR over* $\{0,1\}$ *instead of multiplication over* $\{\pm 1\}$.*)*



(a) What is the rate of the encoder?

(b) Draw one trellis section for the above encoder.

(c) Consider communication of this sequence through the channel known as Binary Erasure Channel (BEC). The input of the channel belongs to $\{0,1\}$ *and the output belongs to* $\{0,1,?\}$. *The "?" denotes an erasure which means that the output is equally likely to be either* 0 *or* 1. *The transition probabilities of the channel are given by*

$$P_{Y|X}(0|0) = P_{Y|X}(1|1) = 1-\epsilon,$$
$$P_{Y|X}(?|0) = P_{Y|X}(?|1) = \epsilon.$$

Starting from first principles derive the branch metric of the optimal (MAP) decoder. (Hint: Start with $p(x|y)$. *Hopefully you are not scared of* ∞ *?)*

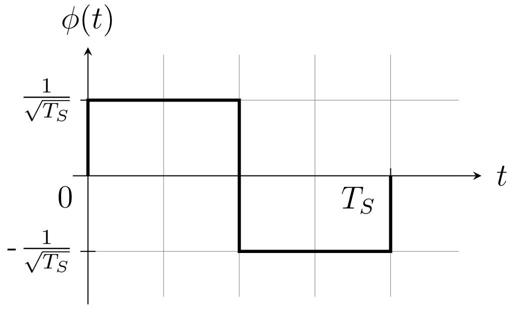
(d) Assuming that the initial state is $(0,0)$, *what is the most likely input corresponding to* $\{0,?,?,1,0,1\}$ *?*

(e) What is the maximum number of erasures the code can correct? (Hint: What is the minimum distance of the code? Just guess from the trellis, don't use the detour graph. :-))

PROBLEM 14. (Power Spectrum: Manchester Pulse) *In this problem you will derive the power spectrum of a signal*

$$X(t) = \sum_{i=-\infty}^{\infty} X_i \phi(t - iT_s - \Theta)$$

where $\{X_i\}_{i=-\infty}^{\infty}$ *is an iid sequence of uniformly distributed random variables taking values in* $\{\pm\sqrt{E_s}\}$, Θ *is uniformly distributed in the interval* $[0, T_s]$, *and* $\phi(t)$ *is the so-called Manchester pulse shown in the following figure*



(a) Let $r(t) = \sqrt{\frac{1}{T_s}}1_{[-\frac{T_s}{4},\frac{T_s}{4}]}(t)$ *be a rectangular pulse. Plot* $r(t)$ *and* $r_{\mathcal{F}}(f)$, *both appropriately labeled, and write down a mathematical expression for* $r_{\mathcal{F}}(f)$.

(b) Derive an expression for $|\phi_{\mathcal{F}}(f)|^2$. *Your expression should be of the form* $A\frac{\sin^m()}{()^n}$ *for some* A, m, *and* n. *Hint: Write* $\phi(t)$ *in terms of* $r(t)$ *and recall that* $\sin x = \frac{e^{jx}-e^{-jx}}{2j}$ *where* $j = \sqrt{-1}$.

(c) Determine $R_X[k] \triangleq E[X_{i+k}X_i]$ *and the power spectrum*

$$S_X(f) = \frac{|\phi_{\mathcal{F}}(f)|^2}{T_S}\sum_{k=-\infty}^{\infty} R_X[k]e^{-j2\pi kfT_s}.$$

Chapter 7

Complex-Valued Random Variables and Processes

7.1 Introduction

In this chapter we define and study complex-valued random variables and complex-valued stochastic processes. We need them to model the noise of the baseband-equivalent channel. Besides being practical in many situations, working with complex-valued random variables and processes turns out to be more elegant than working with the real-valued counterparts.

We will focus on the subclass of complex-valued random variables and processes called *proper* (to be defined). We do so since the class is big enough to contain what we need ad at the same time it is simpler to work with than with the general class of complex-valued random variables and processes.

7.2 Complex-Valued Random Variables

A complex-valued random variable U (hereafter simply called complex random variable) is defined as a random variable of the form

$$U = U_R + jU_I, \quad j = \sqrt{-1},$$

where U_R and U_I are real-valued random variables.

The statistical properties of $U = U_R + jU_I$ are determined by the joint distribution $P_{U_R U_I}(u_R, u_I)$ of U_R and U_I.

A real random variable X is specified by its cumulative distribution function $F_X(x) = \Pr(X \leq x)$. For a complex random variable Z, since there is no natural ordering in the complex plane, the event $Z \leq z$ does not make sense. Instead, we specify a complex random variable by giving the joint distribution of its real and imaginary parts

$F_{\Re\{Z\},\Im\{Z\}}(x,y) = \Pr(\Re\{Z\} \le x,\, \Im\{Z\} \le y)$. Since the pair of real numbers (x,y) can be identified with a complex number $z = x + iy$, we will write the joint distribution $F_{\Re\{Z\},\Im\{Z\}}(x,y)$ as $F_Z(z)$. Just as we do for real valued random variables, if the function $F_{\Re\{Z\},\Im\{Z\}}(x,y)$ is differentiable in x and y, we will call the function

$$p_{\Re\{Z\},\Im\{Z\}}(x,y) = \frac{\partial^2}{\partial x \partial y} F_{\Re\{Z\},\Im\{Z\}}(x,y)$$

the joint density of $(\Re\{Z\},\Im\{Z\})$, and again associating with (x,y) the complex number $z = x + iy$, we will call the function

$$p_Z(z) = p_{\Re\{Z\},\Im\{Z\}}(\Re\{z\},\Im\{z\})$$

the density of the random variable Z.

A complex random vector $\boldsymbol{Z} = (Z_1,\ldots,Z_n)$ is specified by the joint distribution of $(\Re\{Z_1\},\ldots,\Re\{Z_n\},\Im\{Z_1\},\ldots,\Im\{Z_n\})$, and we define the distribution of Z as

$$F_{\boldsymbol{Z}}(\boldsymbol{z}) = \Pr(\Re\{Z_1\} \le \Re\{z_1\},\ldots,\Re\{Z_n\} \le \Re\{z_n\},\Im\{Z_1\} \le \Im\{z_1\},\ldots,\Im\{Z_n\} \le \Im\{z_n\}),$$

and if this function is differentiable in $\Re\{z_1\},\ldots,\Re\{z_n\},\Im\{z_1\},\ldots,\Im\{z_n\}$, then we define the density of $\boldsymbol{Z}$ as

$$p_{\boldsymbol{Z}}(x_1+iy_1,\ldots,x_n+iy_n) = \frac{\partial^{2n}}{\partial x_1 \cdots \partial x_n \partial y_1 \cdots \partial y_n} F_{\boldsymbol{Z}}(x_1+iy_1,\ldots,x_n+iy_n).$$

The expectation of a real random vector $\boldsymbol{X}$ is naturally generalized to the complex case

$$E[\boldsymbol{U}] = E[\boldsymbol{U}_R] + jE[\boldsymbol{U}_I].$$

Recall that the covariance matrix of two real-valued random vectors $\boldsymbol{X}$ and $\boldsymbol{Y}$ is defined as

$$K_{\boldsymbol{XY}} = cov[\boldsymbol{X},\boldsymbol{Y}] \overset{\triangle}{=} E[(\boldsymbol{X}-E[\boldsymbol{X}])(\boldsymbol{Y}-E[\boldsymbol{Y}])^T]. \tag{7.1}$$

To specify the "covariance" of the two complex random vectors $\boldsymbol{U} = \boldsymbol{U}_R + j\boldsymbol{U}_I$ and $\boldsymbol{V} = \boldsymbol{V}_R + j\boldsymbol{V}_I$ the four covariance matrices

$$\begin{aligned} K_{\boldsymbol{U}_R\boldsymbol{V}_R} &= cov[\boldsymbol{U}_R,\boldsymbol{V}_R] \qquad & K_{\boldsymbol{U}_R\boldsymbol{V}_I} &= cov[\boldsymbol{U}_R,\boldsymbol{V}_I] \\ K_{\boldsymbol{U}_I\boldsymbol{V}_R} &= cov[\boldsymbol{U}_I,\boldsymbol{V}_R] \qquad & K_{\boldsymbol{U}_I\boldsymbol{V}_I} &= cov[\boldsymbol{U}_I,\boldsymbol{V}_I] \end{aligned} \tag{7.2}$$

are needed. These four real-valued matrices are equivalent to the following two complex-valued matrices, each of which is a natural generalization of (7.1)

$$\begin{aligned} K_{\boldsymbol{UV}} &\overset{\triangle}{=} E[(\boldsymbol{U}-E[\boldsymbol{U}])(\boldsymbol{V}-E[\boldsymbol{V}])^\dagger] \\ J_{\boldsymbol{UV}} &\overset{\triangle}{=} E[(\boldsymbol{U}-E[\boldsymbol{U}])(\boldsymbol{V}-E[\boldsymbol{V}])^T] \end{aligned} \tag{7.3}$$

The reader is encouraged to verify that the following (straightforward) relationships hold:

$$\begin{aligned} K_{\boldsymbol{UV}} &= K_{\boldsymbol{U}_R\boldsymbol{V}_R} + K_{\boldsymbol{U}_I\boldsymbol{V}_I} + j(K_{\boldsymbol{U}_I\boldsymbol{V}_R} - K_{\boldsymbol{U}_R\boldsymbol{V}_I}) \\ J_{\boldsymbol{UV}} &= K_{\boldsymbol{U}_R\boldsymbol{V}_R} - K_{\boldsymbol{U}_I\boldsymbol{V}_I} + j(K_{\boldsymbol{U}_I\boldsymbol{V}_R} + K_{\boldsymbol{U}_R\boldsymbol{V}_I}). \end{aligned} \tag{7.4}$$

This system may be solved for $K_{\boldsymbol{U}_R\boldsymbol{V}_R}$, $K_{\boldsymbol{U}_I\boldsymbol{V}_I}$, $K_{\boldsymbol{U}_I\boldsymbol{V}_R}$, and $K_{\boldsymbol{U}_R\boldsymbol{V}_I}$ to obtain

$$\begin{aligned} K_{\boldsymbol{U}_R\boldsymbol{V}_R} &= \tfrac{1}{2}\Re\{K_{\boldsymbol{UV}} + J_{\boldsymbol{UV}}\} \\ K_{\boldsymbol{U}_I\boldsymbol{V}_I} &= \tfrac{1}{2}\Re\{K_{\boldsymbol{UV}} - J_{\boldsymbol{UV}}\} \\ K_{\boldsymbol{U}_I\boldsymbol{V}_R} &= \tfrac{1}{2}\Im\{K_{\boldsymbol{UV}} + J_{\boldsymbol{UV}}\} \\ K_{\boldsymbol{U}_R\boldsymbol{V}_I} &= \tfrac{1}{2}\Im\{-K_{\boldsymbol{UV}} + J_{\boldsymbol{UV}}\} \end{aligned} \tag{7.5}$$

proving that indeed the four real-valued covariance matrices in (7.2) are in one-to-one relationship with the two complex-valued covariance matrices in (7.3).

In the literature $K_{\boldsymbol{UV}}$ is widely used and it is called *covariance matrix* (of the complex random vectors $\boldsymbol{U}$ and $\boldsymbol{V}$). Hereafter $J_{\boldsymbol{UV}}$ will be called the *pseudo-covariance matrix* (of $\boldsymbol{U}$ and $\boldsymbol{V}$). For notational convenience we will write $K_{\boldsymbol{U}}$ instead of $K_{\boldsymbol{UU}}$ and $J_{\boldsymbol{U}}$ instead of $J_{\boldsymbol{UU}}$.

Definition 65. *$\boldsymbol{U}$ and $\boldsymbol{V}$ are said to be* uncorrelated *if all four covariances in* (7.2) *vanish.*

From (7.3), we now obtain the following.

Lemma 66. *The complex random vectors $\boldsymbol{U}$ and $\boldsymbol{V}$ are uncorrelated iff $K_{\boldsymbol{UV}} = J_{\boldsymbol{UV}} = 0$.*

Proof. The "if" part follows from (7.5) and the "only if" part from (7.4). □

7.3 Complex-Valued Random Processes

We focus on discrete-time random processes since corresponding results for continuous-time random processes follow in a straightforward fashion.

A discrete-time *complex* random process is defined as a random process of the form

$$U[n] = U_R[n] + jU_I[n]$$

where $U_R[n]$ and $U_I[n]$ are a pair of *real* discrete-time random processes.

Definition 67. *A complex random process is* wide-sense stationary *(w.s.s.) if its real and imaginary parts are jointly w.s.s., i.e., if the real and the imaginary parts are individually w.s.s. and the cross-correlation depends on the time difference.*

Definition 68. *We define*

$$\begin{aligned} r_{\boldsymbol{U}}[m,n] &\overset{\triangle}{=} E\left[U[n+m]U^*[n]\right] \\ s_{\boldsymbol{U}}[m,n] &\overset{\triangle}{=} E\left[U[n+m]U[n]\right] \end{aligned}$$

as the autocorrelation *and* pseudo-autocorrelation *functions of $U[n]$.*

Lemma 69. *A complex random process $U[n]$ is w.s.s. if and only if $E[U[n]]$, $r_{\boldsymbol{U}}[m,n]$, and $s_{\boldsymbol{U}}[m,n]$ are independent of n.*

Proof. The proof is left as an exercise. □

7.4 Proper Complex Random Variables

Proper random variables are of interest to us since they arise in practical applications and since they are mathematically easier to deal with than their non-proper counterparts.[1]

Definition 70. *A complex random vector $\boldsymbol{U}$ is called* proper *if its pseudo-covariance $J_{\boldsymbol{U}}$ vanishes. The complex random vectors $\boldsymbol{U}_1$ and $\boldsymbol{U}_2$ are called* jointly proper *if the composite random vector $\begin{bmatrix} \boldsymbol{U}_1 \\ \boldsymbol{U}_2 \end{bmatrix}$ is proper.*

Lemma 71. *Two jointly proper, complex random vectors $\boldsymbol{U}$ and $\boldsymbol{V}$ are uncorrelated, if and only if their covariance matrix $K_{\boldsymbol{U}\boldsymbol{V}}$ vanishes.*

Proof. The proof easily follows from the definition of joint properness and Lemma 66. □

Note that any subvector of a proper random vector is also proper. By this we mean that if $\left[\begin{smallmatrix} U_1 \\ U_2 \end{smallmatrix}\right]$ is proper, then U_1 and U_2 are proper. However, two individual proper random vectors are not necessarily jointly proper.

Using the fact that (by definition) $K_{\boldsymbol{U}_R\boldsymbol{U}_I} = K^T_{\boldsymbol{U}_I\boldsymbol{U}_R}$, the pseudo-covariance matrix $J_{\boldsymbol{U}}$ may be written as

$$J_{\boldsymbol{U}} = (K_{\boldsymbol{U}_R} - K_{\boldsymbol{U}_I}) + j(K_{\boldsymbol{U}_I\boldsymbol{U}_R} + K^T_{\boldsymbol{U}_I\boldsymbol{U}_R}).$$

Thus:

Lemma 72. *A complex random vector $\boldsymbol{U}$ is proper iff*

$$K_{\boldsymbol{U}_R} = K_{\boldsymbol{U}_I} \quad \text{and} \quad K_{\boldsymbol{U}_I\boldsymbol{U}_R} = -K^T_{\boldsymbol{U}_I\boldsymbol{U}_R},$$

i.e. $J_{\boldsymbol{U}}$ vanishes, iff $\boldsymbol{U}_R$ and $\boldsymbol{U}_I$ have identical auto-covariance matrices and their cross-covariance matrix is skew-symmetric.[2]

Notice that the skew-symmetry of $K_{\boldsymbol{U}_I\boldsymbol{U}_R}$ implies that $K_{\boldsymbol{U}_I\boldsymbol{U}_R}$ has a zero main diagonal, which means that the real and imaginary part of each component U_k of $\boldsymbol{U}$ are uncorrelated. The vanishing of $J_{\boldsymbol{U}}$ does not, however, imply that the real part of U_k and the imaginary part of U_l are uncorrelated for $k \neq l$.

[1] Proper Gaussian random vectors also maximize entropy among all random vectors of a given covariance matrix. Among the many nice properties of Gaussian random vectors, this is arguably the most important one in information theory.

[2] A matrix A is skew-symmetric if $A^T = -A$.

Notice that a *real* random vector is a proper complex random vector, if and only if it is constant (with probability 1), since $K_{U_I} = 0$ and Lemma 72 imply $K_{U_R} = 0$.

Example 73. *One way to satisfy the first condition of Lemma 72 is to let $\boldsymbol{U}_R$ and $\boldsymbol{U}_I$ be identically distributed random vectors. This will guarantee $K_{\boldsymbol{U}_R} = K_{\boldsymbol{U}_I}$. If $\boldsymbol{U}_R$ and $\boldsymbol{U}_I$ are also independent then $K_{U_I U_R} = 0$ and the second condition of lemma 72 is also satisfied. Hence a vector $\boldsymbol{U} = \boldsymbol{U}_R + j\boldsymbol{U}_I$ is proper if $\boldsymbol{U}_R$ and $\boldsymbol{U}_I$ are independent and identically distributed.* □

Example 74. *Let us construct a vector $\boldsymbol{V}$ which is proper in spite of the fact that its real and imaginary parts are not independent. Let $U = X + jY$ with X and Y independent and identically distributed and let*

$$\boldsymbol{V} = (U, aU)^T$$

for some complex-valued number $a = a_R + ja_I$. Clearly $\boldsymbol{V}_R = (X, a_R X - a_I Y)^T$ and $\boldsymbol{V}_I = (Y, a_R Y + a_I X)^T$ are identically distributed. Hence $K_{\boldsymbol{V}_R} = K_{\boldsymbol{V}_I}$. If we let σ^2 be the variance of X and Y we obtain

$$K_{\boldsymbol{V}_I \boldsymbol{V}_R} = \begin{pmatrix} 0 & a_I\sigma^2 \\ -a_I\sigma^2 & (a_R a_I - a_R a_I)\sigma^2 \end{pmatrix} = \begin{pmatrix} 0 & a_I\sigma^2 \\ -a_I\sigma^2 & 0 \end{pmatrix}$$

Hence $\boldsymbol{V}$ is proper in spite of the fact that its real and imaginary parts are correlated. □

The above example was a special case of the following general result.

Lemma 75 (Closure Under Affine Transformations). *Let $\boldsymbol{U}$ be a proper n-dimensional random vector, i.e., $J_{\boldsymbol{U}} = 0$. Then any vector obtained from $\boldsymbol{U}$ by an affine transformation, i.e. any vector $\boldsymbol{V}$ of the form $\boldsymbol{V} = A\boldsymbol{U} + \mathbf{b}$, where $A \in \mathbb{C}^{m\times n}$ and $\mathbf{b} \in \mathbb{C}^m$ are constant, is also proper.*

Proof. From

$$E[\boldsymbol{V}] = AE[\boldsymbol{U}] + \mathbf{b}$$

it follows

$$\boldsymbol{V} - E[\boldsymbol{V}] = A(\boldsymbol{U} - E[\boldsymbol{U}])$$

Hence we have

$$\begin{aligned} J_{\boldsymbol{V}} &= E[(\boldsymbol{V} - E[\boldsymbol{V}])(\boldsymbol{V} - E[\boldsymbol{V}])^T] \\ &= E\{A(\boldsymbol{U} - E[\boldsymbol{U}])(\boldsymbol{U} - E[\boldsymbol{U}])^T A^T\} \\ &= AJ_{\boldsymbol{U}}A^T = 0 \end{aligned}$$

□

Corollary 76. *Let $\boldsymbol{U}$ and $\boldsymbol{V}$ be as in the previous Lemma. Then $\boldsymbol{U}$ and $\boldsymbol{V}$ are jointly proper.*

Proof. The vector having $\boldsymbol{U}$ and $\boldsymbol{V}$ as subvectors is obtained by the affine transformation

$$\begin{bmatrix}\boldsymbol{U}\\ \boldsymbol{V}\end{bmatrix} = \begin{bmatrix}I_n\\ A\end{bmatrix}\boldsymbol{U} + \begin{bmatrix}\mathbf{0}\\ \mathbf{b}\end{bmatrix}.$$

The claim now follows from Lemma 75. □

LEMMA 77. *Let $\boldsymbol{U}$ and $\boldsymbol{V}$ be two independent complex random vectors and let $\boldsymbol{U}$ be proper. Then the linear combination $\boldsymbol{W} = a_1\boldsymbol{U} + a_2\boldsymbol{V}, a_1, a_2 \in \mathbb{C}, a_2 \neq 0$, is proper iff $\boldsymbol{V}$ is also proper.*

Proof. The independence of $\boldsymbol{U}$ and $\boldsymbol{V}$ and the properness of $\boldsymbol{U}$ imply

$$J_{\boldsymbol{W}} = a_1^2 J_{\boldsymbol{U}} + a_2^2 J_{\boldsymbol{V}} = a_2^2 J_{\boldsymbol{V}}.$$

Thus $J_{\boldsymbol{W}}$ vanishes iff $J_{\boldsymbol{V}}$ vanishes. □

7.5 Relationship Between Real and Complex-Valued Operations

Consider now an arbitrary vector $\boldsymbol{u} \in \mathbb{C}^n$ (not necessarily a *random* vector), let $A \in \mathbb{C}^{m\times n}$, and suppose that we would like to implement the operation that maps $\boldsymbol{u}$ to $\boldsymbol{v} = A\boldsymbol{u}$. Suppose also that we implement this operation on a DSP which is programmed at a level at which we can't rely on routines that handle complex-valued operations. A natural question is: how do we implement $\boldsymbol{v} = A\boldsymbol{u}$ using real-valued operations? More generally, what is the relationship between complex-valued variables and operations with respect to their real-valued counterparts? We need this knowledge in the next section to derive the probability density function of proper Gaussian random vectors.

A natural approach is to define the operation that maps a general complex vector $\boldsymbol{u}$ into a real vector $\hat{\boldsymbol{u}}$ according to

$$\hat{\boldsymbol{u}} = \begin{bmatrix}\boldsymbol{u}_R\\ \boldsymbol{u}_I\end{bmatrix} \triangleq \begin{bmatrix}\Re[\boldsymbol{u}]\\ \Im[\boldsymbol{u}]\end{bmatrix} \tag{7.6}$$

and hope for the existence of a real-valued matrix $\hat{A}$ such that

$$\hat{\boldsymbol{v}} = \hat{A}\hat{\boldsymbol{u}}.$$

From $\hat{\boldsymbol{v}}$ we can then immediately obtain $\boldsymbol{v}$. Fortunately such a matrix exists and it is straightforward to verify that

$$\hat{A} = \begin{bmatrix}A_R & -A_I\\ A_I & A_R\end{bmatrix} \triangleq \begin{bmatrix}\Re[A] & -\Im[A]\\ \Im[A] & \Re[A]\end{bmatrix}. \tag{7.7}$$

A set of operations on complex-valued vectors and matrices and the corresponding real-valued operations are described in the following Lemma.

LEMMA 78. *The following properties hold:*

$$\widehat{AB} = \hat{A}\hat{B} \tag{7.8a}$$

$$\widehat{A+B} = \hat{A} + \hat{B} \tag{7.8b}$$

$$\widehat{A^\dagger} = \hat{A}^\dagger \tag{7.8c}$$

$$\widehat{A^{-1}} = \hat{A}^{-1} \tag{7.8d}$$

$$\det(\hat{A}) = |\det(A)|^2 = \det(AA^\dagger) \tag{7.8e}$$

$$\widehat{\boldsymbol{u}+\boldsymbol{v}} = \hat{\boldsymbol{u}} + \hat{\boldsymbol{v}} \tag{7.8f}$$

$$\widehat{A\boldsymbol{u}} = \hat{A}\hat{\boldsymbol{u}} \tag{7.8g}$$

$$\Re(\boldsymbol{u}^\dagger \boldsymbol{v}) = \hat{\boldsymbol{u}}^\dagger \hat{\boldsymbol{v}} \tag{7.8h}$$

Proof. The properties (7.8a), (7.8b) and (7.8c) are immediate. For instance, property (7.8a) is verified as follows:

$$\begin{aligned}\widehat{AB} &= \begin{bmatrix}(AB)_R & -(AB)_I \\ (AB)_I & (AB)_R\end{bmatrix} \\ &= \begin{bmatrix}A_R B_R - A_I B_I & -A_R B_I - A_I B_R \\ A_R B_I + A_I B_R & A_R B_R - A_I B_I\end{bmatrix} \\ &= \begin{bmatrix}A_R & -A_I \\ A_I & A_R\end{bmatrix}\begin{bmatrix}B_R & -B_I \\ B_I & B_R\end{bmatrix} \\ &= \hat{A}\hat{B}\end{aligned}$$

Property (7.8d) follows from (7.8a) and the fact that $\hat{I}_n = I_{2n}$. To prove (7.8e) we use the fact that the determinant of a product is the product of the determinant and the determinant of a block triangular matrix is the product of the determinants of the diagonal blocks. Hence:

$$\det(\hat{A}) = \det\left(\begin{bmatrix}I & jI \\ 0 & I\end{bmatrix}\hat{A}\begin{bmatrix}I & -jI \\ 0 & I\end{bmatrix}\right) = \det\left(\begin{bmatrix}A & 0 \\ \Im(A) & A^*\end{bmatrix}\right) = \det(A)\det(A)^*.$$

Properties (7.8f), (7.8g) and (7.8h) are immediate. □

COROLLARY 79. *If $U \in \mathbb{C}^{n\times n}$ is unitary then $\hat{U} \in \mathbb{R}^{2n\times 2n}$ is orthonormal.*

Proof. $U^\dagger U = I_n \iff (\hat{U})^\dagger \hat{U} = \hat{I}_n = I_{2n}$. □

COROLLARY 80. *If $Q \in \mathbb{C}^{n\times n}$ is non-negative definite, then so is $\hat{Q} \in \mathbb{R}^{2n\times 2n}$. Moreover, $\boldsymbol{u}^\dagger Q \boldsymbol{u} = \hat{\boldsymbol{u}}^\dagger \hat{Q} \hat{\boldsymbol{u}}$.*

Proof. Assume that Q is non-negative definite. Then $\boldsymbol{u}^\dagger Q\boldsymbol{u}$ is a non-negative real-valued number for all $\boldsymbol{u} \in \mathbb{C}^n$. Hence,

$$\begin{aligned}\boldsymbol{u}^\dagger Q\boldsymbol{u} &= \Re\{\boldsymbol{u}^\dagger(Q\boldsymbol{u})\} = \hat{\boldsymbol{u}}^\dagger\widehat{(Q\boldsymbol{u})}\\ &= \hat{\boldsymbol{u}}^\dagger\hat{Q}\hat{\boldsymbol{u}}\end{aligned}$$

where in the last two equalities we used (7.8h) and (7.8g), respectively. □

Exercise 81. *Verify that a random vector $\boldsymbol{U}$ is proper iff $2K_{\hat{\boldsymbol{U}}} = \hat{K}_{\boldsymbol{U}}$.*

7.6 Complex-Valued Gaussian Random Variables

A complex-valued Gaussian random vector $\boldsymbol{U}$ is defined as a vector with jointly Gaussian real and imaginary parts. Following Feller [2, p. 86], we consider Gaussian distributions to include degenerate distributions concentrated on a lower-dimensional manifold, i.e., when the $2n \times 2n$-covariance matrix

$$cov\left(\begin{bmatrix}\boldsymbol{U}_R\\ \boldsymbol{U}_I\end{bmatrix}, \begin{bmatrix}\boldsymbol{U}_R\\ \boldsymbol{U}_I\end{bmatrix}\right) = \begin{bmatrix}K_{\boldsymbol{U}_R} & K_{\boldsymbol{U}_I\boldsymbol{U}_R}\\ K_{\boldsymbol{U}_I\boldsymbol{U}_R} & K_{\boldsymbol{U}_I}\end{bmatrix}$$

is singular and the pdf does not exist unless one admits generalized functions.

Hence, by definition, a complex-valued random vector $\boldsymbol{U} \in \mathbb{C}^n$ with nonsingular covariance matrix $K_{\hat{\boldsymbol{U}}}$ is Gaussian iff

$$f_{\boldsymbol{U}}(\boldsymbol{u}) = f_{\hat{\boldsymbol{U}}}(\hat{\boldsymbol{u}}) = \frac{1}{[\det(2\pi K_{\hat{\boldsymbol{U}}})]^{\frac{1}{2}}} e^{-\frac{1}{2}(\hat{\boldsymbol{u}}-\hat{\boldsymbol{m}})^T K_{\hat{\boldsymbol{U}}}^{-1}(\hat{\boldsymbol{u}}-\hat{\boldsymbol{m}})}. \tag{7.9}$$

Theorem 82. *Let $\boldsymbol{U} \in \mathbb{C}^n$ be a proper Gaussian random vector with mean $\boldsymbol{m}$ and nonsingular covariance matrix $K_{\boldsymbol{U}}$. Then the pdf of $\boldsymbol{U}$ is given by*

$$f_{\boldsymbol{U}}(\boldsymbol{u}) = f_{\hat{\boldsymbol{U}}}(\hat{\boldsymbol{u}}) = \frac{1}{\det(\pi K_{\boldsymbol{U}})} e^{-(\boldsymbol{u}-\boldsymbol{m})^\dagger K_{\boldsymbol{U}}^{-1}(\boldsymbol{u}-\boldsymbol{m})}. \tag{7.10}$$

Conversely, let the pdf of a random $\boldsymbol{U}$ be given by (7.10) *where $K_{\boldsymbol{U}}$ is some Hermitian and positive definite matrix. Then $\boldsymbol{U}$ is proper and Gaussian with covariance matrix $K_{\boldsymbol{U}}$ and mean $\boldsymbol{m}$.*

Proof. If $\boldsymbol{U}$ is proper then by Exercise 81

$$\sqrt{\det 2\pi K_{\hat{\boldsymbol{U}}}} = \sqrt{\det \pi \hat{K}_{\boldsymbol{U}}} = |\det \pi K_{\boldsymbol{U}}| = \det \pi K_{\boldsymbol{U}},$$

where the last equality holds since the determinant of an Hermitian matrix is always real. Moreover, letting $\hat{\boldsymbol{v}} = \hat{\boldsymbol{u}} - \hat{\boldsymbol{m}}$, again by Exercise 81

$$\hat{\boldsymbol{v}}^\dagger(2K_{\hat{\boldsymbol{U}}})^{-1}\hat{\boldsymbol{v}} = \hat{\boldsymbol{v}}^\dagger(\hat{K}_{\boldsymbol{U}})^{-1}\hat{\boldsymbol{v}} = \boldsymbol{v}^\dagger(K_{\boldsymbol{U}})^{-1}\boldsymbol{v}$$

where for last equality we used Corollary 80 and the fact that if a matrix is positive definite, so is its inverse. Using the last two relationships in (7.9) yields the direct part of the theorem. The converse follows similarly. □

Notice that two jointly proper Gaussian random vectors $\boldsymbol{U}$ and $\boldsymbol{V}$ are independent, iff $K_{\boldsymbol{UV}} = 0$, which follows from Lemma 71 and the fact that uncorrelatedness and independence are equivalent for Gaussian random variables.

Appendix 7.A Densities after Linear transformations of complex random variables

We know that $\boldsymbol{X}$ is a real random vector with density $p_{\boldsymbol{X}}$, and if A is a non-singular matrix, then the density of $Y = AX$ is given by

$$p_{\boldsymbol{Y}}(\boldsymbol{y}) = |\det(A)|^{-1} p_{\boldsymbol{X}}(A^{-1}y).$$

If $\boldsymbol{Z}$ is a complex random vector with density $p_{\boldsymbol{Z}}$ and if A is a complex non-singular matrix, then $W = AZ$ is again a complex random vector with

$$\begin{bmatrix} \Re\{W\} \\ \Im\{W\} \end{bmatrix} = \begin{bmatrix} \Re\{A\} & -\Im\{A\} \\ \Im\{A\} & \Re\{A\} \end{bmatrix} \begin{bmatrix} \Re\{Z\} \\ \Im\{Z\} \end{bmatrix}$$

and thus the density of W will be given by

$$p_{\boldsymbol{W}}(\boldsymbol{w}) = \left| \det \left(\begin{bmatrix} \Re\{A\} & -\Im\{A\} \\ \Im\{A\} & \Re\{A\} \end{bmatrix} \right) \right|^{-1} p_{\boldsymbol{Z}}(A^{-1}\boldsymbol{w}).$$

From (7.8e) we know that

$$\det \left(\begin{bmatrix} \Re\{A\} & -\Im\{A\} \\ \Im\{A\} & \Re\{A\} \end{bmatrix} \right) = |\det(A)|^2,$$

and thus the transformation formula becomes

$$p_{\boldsymbol{W}}(\boldsymbol{w}) = |\det(A)|^{-2} p_{\boldsymbol{Z}}(A^{-1}\boldsymbol{w}). \tag{7.11}$$

Appendix 7.B Circular Symmetry

We say that a complex valued random variable Z is *circularly symmetric* if for any $\theta \in [0, 2\pi)$, the distribution of $Ze^{j\theta}$ is the same as the distribution of Z.

Using the linear transformation formula (7.11), we see that the density of Z must satisfy

$$p_Z(z) = p_Z(z \exp(j\theta))$$

for all θ, and thus, p_Z must not depend on the phase of its argument, i.e.,

$$p_Z(z) = p_Z(|z|).$$

We can also conclude that, if Z is circularly symmetric,

$$E[Z] = E[e^{j\theta} Z] = e^{j\theta} E[Z],$$

and taking $\theta = \pi$, we conclude that $E[Z] = 0$. Similarly, $E[Z^2] = 0$.

For (complex) random *vectors*, the definition of circular symmetry is that the distribution of $\boldsymbol{Z}$ should be the same as the distribution of $e^{j\theta}\boldsymbol{Z}$. In particular, by taking $\theta = \pi$, we see that

$$E[\boldsymbol{Z}] = 0,$$

and by taking $\theta = \pi/2$, we see that the pseudo covariance

$$J_{\boldsymbol{Z}} = E[\boldsymbol{Z}\boldsymbol{Z}^T] = 0.$$

We have shown that if $\boldsymbol{Z}$ is circularly symmetric, then it is also zero mean and proper.

If $\boldsymbol{Z}$ is a zero-mean Gaussian random vector, then the converse is also true, i.e., properness implies circular symmetry. To see this let $\boldsymbol{Z}$ be zero-mean proper and Gaussian. Then $e^{-j\theta}\boldsymbol{Z}$ is also zero-mean and Gaussian. Hence $\boldsymbol{Z}$ and $e^{-j\theta}\boldsymbol{Z}$ have the same density iff they have the same covariance and pseudo-covariance matrices. The pseudo-covariance matrices vanish in both cases ($\boldsymbol{Z}$ is proper and $e^{-j\theta}\boldsymbol{Z}$ is also proper since it is the linear transformation of a proper random vector). Using the definition, one immediately sees that $\boldsymbol{Z}$ and $e^{-j\theta}\boldsymbol{Z}$ have the same covariance matrix. Hence they have the same density.

Appendix 7.C On Linear Transformations and Eigenvectors

The material developed in this appendix is not relevant for the subject covered in this textbook. The reason for including it is that it is both instructive and important for topics related to the subject in this class.

The Fourier transform is a useful tool in dealing with linear time-invariant (LTI) systems. This is so since the input/output relationship if a LTI system is easily described in the Fourier domain. In this section we learn that this is just a special case of a more general principle that applies to linear transformations (not necessarily time-invariant). Key ingredients are the eigenvectors.

7.C.1 Linear Transformations, Toepliz, and Circulant Matrices

A linear transformation from $\mathbb{C}^n$ to $\mathbb{C}^n$ can be described by an $n \times n$ matrix H. If the matrix is *Toepliz*, meaning that $H_{ij} = h_{i-j}$, then the transformation which sends $\boldsymbol{u} \in \mathbb{C}^n$ to $\boldsymbol{v} = H\boldsymbol{u}$ can be described by the *convolution* sum

$$v_i = \sum_k h_{i-k} u_k.$$

A Toepliz matrix is a matrix which is constant along its diagonals.

In this section we focus attention on Toepliz matrices of a special kind called circulant. A matrix H is *circulant* if $H_{ij} = h_{[i-j]}$ where here and hereafter the operator $[.]$ applied to an index denotes the index taken modulo n. When H is circulant, the operation that maps $\boldsymbol{u}$ to $\boldsymbol{v} = H\boldsymbol{u}$ may be described by the *circulant convolution*

$$v_i = \sum_k h_{[i-k]} u_k.$$

Example 83.

$$H = \begin{bmatrix} 3 & 1 & 5 \\ 5 & 3 & 1 \\ 1 & 5 & 3 \end{bmatrix} \quad \textit{is a circulant matrix.}$$

A circulant matrix H is completely described by its first column $\boldsymbol{h}$ (or any column or row for that matter). □

7.C.2 The DFT

The discrete Fourier transform of a vector $\boldsymbol{u} \in \mathbb{C}^n$ is the vector $\boldsymbol{U} \in \mathbb{C}^n$ defined by

$$\begin{aligned} \boldsymbol{U} &= F^\dagger \boldsymbol{u} \\ F &= (\boldsymbol{f}_1, \boldsymbol{f}_2, \dots, \boldsymbol{f}_n) \\ \boldsymbol{f}_i &= \frac{1}{\sqrt{n}} \begin{bmatrix} \beta^{i0} \\ \beta^{i1} \\ \vdots \\ \beta^{i(n-1)} \end{bmatrix} \qquad i = 1, 2, \dots, n, \end{aligned} \tag{7.12}$$

where $\beta = e^{j\frac{2\pi}{n}}$ is the primitive n-th root of unity in $\mathbb{C}$. Notice that $\boldsymbol{f}_1, \boldsymbol{f}_2, \dots, \boldsymbol{f}_n$ is an orthonormal basis for $\mathbb{C}^n$.

Usually, the DFT is defined without the $\sqrt{n}$ in (7.12) and with a factor $\frac{1}{n}$ (instead of $1/\sqrt{n}$) in the inverse transform. The resulting transformation is not orthonormal, and a factor n must be inserted in Parseval's identity when it is applied to the DFT. In this class we call $F^\dagger \boldsymbol{u}$ the DFT of $\boldsymbol{u}$.

7.C.3 Eigenvectors of Circulant Matrices

Lemma 84. *Any circulant matrix $H \in \mathbb{C}^{n \times n}$ has exactly n (normalized) eigenvectors which may be taken as $\boldsymbol{f}_1, \dots, \boldsymbol{f}_n$. Moreover, the vector of eigenvalues $(\lambda_1, \dots, \lambda_n)^T$ equals $\sqrt{n}$ times the DFT of the first column of H, namely $\sqrt{n} F^\dagger \boldsymbol{h}$.*

Example 85. *Consider the matrix*

$$H = \begin{bmatrix} h_0 & h_1 \\ h_1 & h_0 \end{bmatrix} \in \mathbb{C}^{2 \times 2}.$$

This is a circulant matrix. Hence

$$f_1 = \frac{1}{\sqrt{2}}\begin{bmatrix}1\\-1\end{bmatrix} \quad \textit{and} \quad f_2 = \frac{1}{\sqrt{2}}\begin{bmatrix}1\\1\end{bmatrix}$$

are eigenvectors and the eigenvalues are

$$\begin{bmatrix}\lambda_1\\\lambda_2\end{bmatrix} = \sqrt{2}F^\dagger \boldsymbol{h} = \begin{bmatrix}1 & -1\\1 & 1\end{bmatrix}\begin{bmatrix}h_0\\h_1\end{bmatrix} = \begin{bmatrix}h_0 - h_1\\h_0 + h_1\end{bmatrix}$$

indeed

$$H\boldsymbol{f}_1 = \frac{1}{\sqrt{2}}\begin{bmatrix}h_0 - h_1\\h_1 - h_0\end{bmatrix} = \frac{h_0 - h_1}{\sqrt{2}}\begin{bmatrix}1\\-1\end{bmatrix} = \lambda_1 \boldsymbol{f}_1$$

and

$$H\boldsymbol{f}_2 = \frac{1}{\sqrt{2}}\begin{bmatrix}h_0 + h_1\\h_1 + h_0\end{bmatrix} = \frac{h_0 + h_1}{\sqrt{2}}\begin{bmatrix}1\\1\end{bmatrix} = \lambda_2 \boldsymbol{f}_2$$

Proof.

$$\begin{aligned}(H\boldsymbol{f}_i)_k &= \frac{1}{\sqrt{n}}\sum_{e=0}^{n-1} h_{k-e}\beta^{ie}\\ &= \left(\sum_{m=0}^{n-1} h_m \beta^{-im}\right)\frac{1}{\sqrt{n}}\beta^{ik}\\ &= \sqrt{n}\boldsymbol{f}_i^\dagger \boldsymbol{h}\frac{1}{\sqrt{n}}\beta^{ik} = \lambda_i \frac{1}{\sqrt{n}}\beta^{ik},\end{aligned}$$

where $\lambda_i = \sqrt{n}\boldsymbol{f}_i^\dagger \boldsymbol{h}$. Going to vector notation we obtain $H\boldsymbol{f}_i = \lambda_i \boldsymbol{f}_i$. □

7.C.4 Eigenvectors to Describe Linear Transformations

When the eigenvectors of a transformation $H \in \mathbb{C}^{n\times n}$ (not necessarily Toepliz) span $\mathbb{C}^n$, both the vectors and the transformation can be represented with respect to a basis of eigenvectors. In that new basis the transformation takes the form $H' = \text{diag}(\lambda_1, \dots, \lambda_n)$, where diag( ) denotes a matrix with the arguments on the main diagonal and 0s elsewhere, and λ_i is the eigenvalue of the i-th eigenvector. In the new basis the input/output relationship is

$$\boldsymbol{v}' = H'\boldsymbol{u}'$$

or equivalently, $v_i' = \lambda_i u_i'$, $i = 1, 2, \dots, n$. To see this, let $\boldsymbol{\varphi}_i, i = 1 \dots n$, be n eigenvectors of H spanning $\mathbb{C}^n$. Letting $\boldsymbol{u} = \sum_i \boldsymbol{\varphi}_i u_i'$ and $\boldsymbol{v} = \sum_i \boldsymbol{\varphi}_i v_i'$ and plugging into $H\boldsymbol{u}$ we obtain

$$H\boldsymbol{u} = H\Big(\sum_i \boldsymbol{\varphi}_i u_i'\Big) = \sum_i H\boldsymbol{\varphi}_i u_i' = \sum \boldsymbol{\varphi}_i \lambda_i u_i'$$

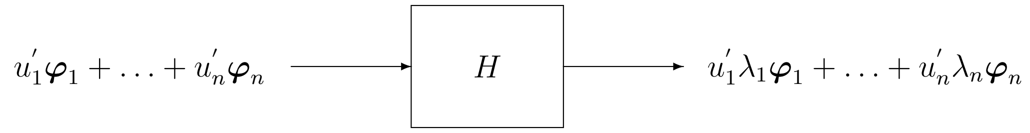

$u_1^{'}\boldsymbol{\varphi}_1 + \ldots + u_n^{'}\boldsymbol{\varphi}_n$
H
$u_1^{'}\lambda_1\boldsymbol{\varphi}_1 + \ldots + u_n^{'}\lambda_n\boldsymbol{\varphi}_n$

Figure 7.1: Input/output representation via eigenvectors.

showing that $v_i' = \lambda_i u_i'$.

Notice that the key aspects in the proof are the linearity of the transformation and the fact that $\boldsymbol{\varphi}_i u_i'$ is sent to $\boldsymbol{\varphi}_i \lambda_i u_i'$, as shown in Figure 7.1.

It is often convenient to use matrix notation. To see how the proof goes with matrix notation we define $\Phi = (\boldsymbol{\varphi}_1, \ldots, \boldsymbol{\varphi}_n)$ as the matrix whose columns span $\mathbb{C}^n$. Then $\boldsymbol{u} = \Phi\boldsymbol{u}'$ and the above proof in matrix notation is

$$\boldsymbol{v} = H\boldsymbol{u} = H\Phi\boldsymbol{u}' = \Phi H'\boldsymbol{u}',$$

showing that $\boldsymbol{v}' = H'\boldsymbol{u}'$.

For the case where H is circulant, $\boldsymbol{u} = F\boldsymbol{u}'$ and $\boldsymbol{v} = F\boldsymbol{v}'$. Hence $\boldsymbol{u}' = F^{\dagger}\boldsymbol{u}$ and $\boldsymbol{v}' = F^{\dagger}\boldsymbol{v}$ are the DFT of $\boldsymbol{u}$ and $\boldsymbol{v}$, respectively. Similarly, the diagonal elements of H' are $\sqrt{n}$ times the DFT of the first column of H. Hence the above representation via the new basis says (the well-know result) that a circular convolution corresponds to a multiplication in the DFT domain.

7.C.5 Karhunen-Loève Expansion

In Appendix 7.C, we have seen that the eigenvectors of a linear transformation H can be used as a basis and in the new basis the linear transformation of interest becomes a componentwise multiplication.

A similar idea can be used to describe a random vector $\boldsymbol{u}$ as a linear combination of deterministic vectors with orthogonal random coefficient. Now the eigenvectors are those of the correlation matrix $r_{\boldsymbol{u}}$. The procedure, that we now describe, is the Karhunen-Loève expansion.

Let $\boldsymbol{\varphi}_1, \ldots, \boldsymbol{\varphi}_n$ be a set of eigenvectors of $r_{\boldsymbol{u}}$ that form an orthonormal basis of $\mathbb{C}^n$. Such a set exists since $r_{\boldsymbol{u}}$ is Hermitian. Hence

$$\lambda_i\boldsymbol{\varphi}_i = r_{\boldsymbol{u}}\boldsymbol{\varphi}_i, i = 1, 2, \ldots, n$$

or, in matrix notation,

$$\Phi\Lambda = r_{\boldsymbol{u}}\Phi$$

where $\Lambda = diag(\lambda_1, \ldots, \lambda_n)$ and $\Phi = [\boldsymbol{\varphi}_1, \ldots, \boldsymbol{\varphi}_n]$ is the matrix whose columns are the eigenvectors. Since the eigenvectors are orthonormal, Φ is unitary (i.e. $\Phi^{\dagger}\Phi = I$).

Solving for Λ we obtain

$$\Lambda = \Phi^\dagger r_{\boldsymbol{u}} \Phi.$$

Notice that if we solve for $r_{\boldsymbol{u}}$ we obtain $r_{\boldsymbol{u}} = \Phi\Lambda\Phi^\dagger$ which is the well known result that an Hermitian matrix can be diagonalized.

Since Φ forms a basis of $\mathbb{C}^n$ we can write

$$\boldsymbol{u} = \Phi \boldsymbol{u}' \tag{7.13}$$

for some vector of coefficient $\boldsymbol{u}'$ with correlation matrix

$$\begin{aligned} r_{\boldsymbol{u}'} &= E[\boldsymbol{u}'(\boldsymbol{u}')^\dagger] = \Phi^\dagger E[\boldsymbol{u}\boldsymbol{u}^\dagger]\Phi = \Phi^\dagger r_{\boldsymbol{u}}\Phi \\ &= \Lambda \end{aligned}$$

Hence (7.13) expresses $\boldsymbol{u}$ as a linear combination of deterministic vectors $\boldsymbol{\varphi}_1, \dots, \boldsymbol{\varphi}_n$ with orthogonal random coefficients $u'_1, \dots, u'_n$. This is the Karhunen-Loève expansion of $\boldsymbol{u}$.

If $r_{\boldsymbol{u}}$ is circulant, then $\Phi = F$ and $\boldsymbol{u}' = \Phi^\dagger \boldsymbol{u}$ is the DFT of $\boldsymbol{u}$.

Remark 86. $\|\boldsymbol{u}\|^2 = \|\boldsymbol{u}'\|^2 = \sum |u'_i|^2$. *Also* $E\|\boldsymbol{u}\|^2 = \sum \lambda_i$.

7.C.6 Circularly Wide-Sense Stationary Random Vectors

We consider random vectors in $\mathbb{C}^n$. We will continue using the notation that $\boldsymbol{u}$ and $\boldsymbol{U}$ denotes DFT pairs. Observe that if $\boldsymbol{U}$ is random then $\boldsymbol{u}$ is also random. This forces us to abandon the convention that we use capital letters for random variables.

The following definitions are natural.

Definition 87. *A random vector* $\boldsymbol{u} \in \mathbb{C}^n$ *is* circularly wide sense stationary *(c.w.s.s.) if*

$$\begin{aligned} m_{\boldsymbol{u}} &\overset{\triangle}{=} E[\boldsymbol{u}] \text{ is a constant vector} \\ r_{\boldsymbol{u}} &\overset{\triangle}{=} E[\boldsymbol{u}\boldsymbol{u}^\dagger] \text{ is a circulant matrix} \\ s_{\boldsymbol{u}} &\overset{\triangle}{=} E[\boldsymbol{u}\boldsymbol{u}^T] \text{ is a circulant matrix} \end{aligned}$$

Definition 88. *A random vector* $\boldsymbol{u}$ *is uncorrelated if* $K_{\boldsymbol{u}}$ *and* $J_{\boldsymbol{u}}$ *are diagonal.*

We will call $r_{\boldsymbol{u}}$ *and* $s_{\boldsymbol{u}}$ *the* circular correlation matrix *and* circular pseudo-correlation matrix, *respectively.*

Theorem 89. *Let* $\boldsymbol{u} \in \mathbb{C}^n$ *be a zero-mean proper random vector and* $\boldsymbol{U} = F^\dagger \boldsymbol{u}$ *be its DFT. Then* $\boldsymbol{u}$ *is c.w.s.s. iff* $\boldsymbol{U}$ *is uncorrelated. Moreover,*

$$r_{\boldsymbol{u}} = \operatorname{circ}(\boldsymbol{a}) \tag{7.14}$$

if and only if

$$r_{\boldsymbol{U}} = \sqrt{n}\ \mathrm{diag}(\boldsymbol{A}) \tag{7.15}$$

for some $\boldsymbol{a}$ *and its DFT* $\boldsymbol{A}$.

Proof. Let $\boldsymbol{u}$ be a zero-mean proper random vector. If $\boldsymbol{u}$ is c.w.s.s. then we can write $r_{\boldsymbol{u}} = \mathrm{circ}(\boldsymbol{a})$ for some vector $\boldsymbol{a}$. Then, using Lemma 84,

$$\begin{aligned} r_{\boldsymbol{U}} &\overset{\triangle}{=} E[F^\dagger \boldsymbol{u}\boldsymbol{u}^\dagger F] = F^\dagger r_{\boldsymbol{u}} F \\ &= F^\dagger \sqrt{n} F \mathrm{diag}(F^\dagger \boldsymbol{a}) \\ &= \sqrt{n}\, \mathrm{diag}(\boldsymbol{A}), \end{aligned}$$

proving (7.15). Moreover, $m_{\boldsymbol{U}} = 0$ since $m_{\boldsymbol{u}} = 0$ and therefore $s_{\boldsymbol{U}} = J_{\boldsymbol{U}}$. But $J_{\boldsymbol{U}} = 0$ since the properness of $\boldsymbol{u}$ and Lemma 75 imply the properness of $\boldsymbol{U}$. Conversely, let $r_{\boldsymbol{U}} = \mathrm{diag}(\boldsymbol{A})$. Then

$$r_{\boldsymbol{u}} = E[\boldsymbol{u}\boldsymbol{u}^\dagger] = F r_{\boldsymbol{U}} F^\dagger.$$

Due to the diagonality of $r_{\boldsymbol{U}}$, the element (k,l) of $r_{\boldsymbol{u}}$ is

$$\begin{aligned} \sqrt{n}\sum_m F_{k,m} A_m (F^\dagger)_{m,l} &= \sum_m F_{k,m} F^*_{l,m} A_m \sqrt{n} \\ &= \frac{1}{\sqrt{n}} \sum_m A_m e^{j\frac{2\pi}{n}m(k-l)} \\ &= a_{k-l} \end{aligned}$$

□

Appendix 7.D Problems

To be filled in

Chapter 8

Passband Communication via Up/Down Conversion

In Chapter 5 we have learned how a wide-sense-stationary symbol sequence $\{X_j : j \in \mathbb{N}\}$ and a finite-energy pulse $\psi(t)$ determine the power spectral density of the random process

$$X(t) \;=\; \sum_{i=-\infty}^{\infty} X_i \psi(t - iT - \Theta), \tag{8.1}$$

where T is an arbitrary positive number and Θ is uniformly distributed in an arbitrary interval of length T. An important special case is when the wide-sense-stationary sequence $\{X_j : j \in \mathbb{N}\}$ is also uncorrelated. Then, up to a scaling factor, the power spectral density of $X(t)$ *is* $\|\psi_{\mathcal{F}}(f)\|^2$. This is a particularly convenient case since Nyquist criterion to check whether or not $\psi(t)$ is orthogonal to its shifts by multiples of T also depends solely on $\|\psi_{\mathcal{F}}(f)\|^2$.

From a practical point of view the above is saying that we should choose $\psi(t)$ by starting from $\|\psi_{\mathcal{F}}(f)\|^2$. For a baseband pulse which is sufficiently narrow in the frequency domain, meaning that that the width of the support set does not exceed $2/T$, the condition for $\|\psi_{\mathcal{F}}(f)\|^2$ to fulfill Nyquist criterion is particularly straightforward (see item (a) of the discussion following Theorem 62). Of course we are particularly interested in bandpass communication. In this chapter we learn how to transform a real or complex-valued baseband process that has power spectral density $\mathcal{S}(f)$ into a real-valued passband process that has power spectral density $[\mathcal{S}(f - f_0) + \mathcal{S}(-f + f_0)]/2$ for an arbitrary[1] *center frequency* f_0. The transformation and its inverse are handled by the top layer of Figure 2.1. As a "sanity check" notice that $[\mathcal{S}(f - f_0) + \mathcal{S}(-f + f_0)]/2$ is an even function of f, which must be the case or else it can't be the power spectral density of a real-valued process.

[1] The expression $[\mathcal{S}(f - f_0) + \mathcal{S}(-f + f_0)]/2$ is the correct power spectral density provided that the center frequency f_0 is sufficiently large, i.e., provided that the support of $\mathcal{S}(f - f_0)$ and that of $\mathcal{S}(f - f_0)$ do not overlap. In all typical scenarios this is the case.

The up/down-conversion technique discussed in this chapter is an elegant way to separate the choice of the center frequency from anything else. Being able to do so is quite convenient since in a typical wireless communication scenario the sender and the receiver need to have the agility to vary the center frequency f_0 so as to minimize the interference with other signals. The up/down-conversion technique has also other desired properties including the fact that it fits well with the modular approach that we have pursued so far (see once again Figure 2.1) and has implementation advantages to be discussed later. In this chapter we also develop the equivalent channel model seen from the up-converter input to the down-converter output. Having such a model makes it possible to design the core of the sender and that of the receiver pretending that the channel is baseband.

8.1 Baseband-Equivalent of a Passband Signal

In this section we learn how to go back and forth between a passband signal $x(t)$ and its *baseband-equivalent* $x_E(t)$, passing through the *analytic-equivalent* $\hat{x}(t)$. These are precisely what we will need in the next section to describe up/down conversion.

We start by recalling a few basic facts from Fourier analysis. If $x(t)$ is a *real-valued* signal, then its Fourier transform $x_{\mathcal{F}}(f)$ satisfies the *symmetry property*

$$x_{\mathcal{F}}(f) = x_{\mathcal{F}}^*(-f)$$

where $x_{\mathcal{F}}^*$ denotes the complex conjugate of $x_{\mathcal{F}}$. If $x(t)$ is a *purely imaginary* signal, then its Fourier transform satisfies the *anti-symmetry property*

$$x_{\mathcal{F}}(f) = -x_{\mathcal{F}}^*(-f)$$

The symmetry and the anti-symmetry properties can easily be verified from the definition of the Fourier transform using the fact that the complex conjugate operator commutes with the integral, i.e., $[\int x(t)dt]^* = \int [x(t)]^* dt$.

The symmetry property implies that the Fourier transform $x_{\mathcal{F}}(f)$ of a *real-valued* signal $x(t)$ has redundant information: if we know $x_{\mathcal{F}}(f)$ for $f \geq 0$ then we can infer $x_{\mathcal{F}}(f)$ for $f \leq 0$. This implies that the set of real-valued signals in $\mathcal{L}_2$ is in one-to-one correspondence with the set of complex-valued signals in $\mathcal{L}_2$ that have vanishing negative frequencies. The correspondence map associates a real-valued signal $x(t)$ to the signal obtained by setting to zero the negative frequency components of $x(t)$. The latter, scaled appropriately so as to have the same $\mathcal{L}_2$ norm as $x(t)$, will be referred to as the *analytic equivalent* of $x(t)$ and will be denoted by $\hat{x}(t)$.

To remove the negative frequencies of $x(t)$ we use the filter of impulse response $h_>(t)$ that has Fourier transform

$$h_{>,\mathcal{F}}(f) = \begin{cases} 1 & \text{for } f > 0 \\ 1/2 & \text{for } f = 0 \\ 0 & \text{for } f < 0. \end{cases} \tag{8.2}$$

Hence an arbitrary real-valued signal $x(t)$ and its analytic-equivalent may be described in the Fourier domain by

$$\hat{x}_{\mathcal{F}}(f) = \sqrt{2}x_{\mathcal{F}}(f)h_{>,\mathcal{F}}(f)$$

where the factor $\sqrt{2}$ ensures that the original and the analytic-equivalent have the same norm. (The part removed by filtering contains half of the signal's energy.)

How to go back from $\hat{x}(t)$ to $x(t)$ may seem less obvious at first but it turns out to be even simpler. We claim that

$$x(t) = \sqrt{2}\Re\left\{\hat{x}(t)\right\}. \tag{8.3}$$

One way to see this is to use the relationship

$$h_{>,\mathcal{F}}(f) = \frac{1}{2} + \frac{1}{2}\operatorname{sign}(f)$$

to obtain

$$\begin{aligned}\hat{x}_{\mathcal{F}}(f) &= \sqrt{2}x_{\mathcal{F}}(f)h_{>,\mathcal{F}}(f)\\ &= \sqrt{2}x_{\mathcal{F}}(f)[\frac{1}{2} + \frac{1}{2}\operatorname{sign}(f)]\\ &= \frac{x_{\mathcal{F}}(f)}{\sqrt{2}} + \frac{x_{\mathcal{F}}(f)}{\sqrt{2}}\operatorname{sign}(f).\end{aligned}$$

The first term of last line satisfies the symmetry property (by assumption) and therefore the second term satisfies the anti-symmetry property. Hence, taking the inverse Fourier transform, $\hat{x}(t)$ equals $\frac{x(t)}{\sqrt{2}}$ plus an imaginary term, implying (8.3). Another way to prove the same is to write

$$\sqrt{2}\Re\left\{\hat{x}(t)\right\} = \frac{1}{\sqrt{2}}\left(\hat{x}(t) + \hat{x}^*(t)\right)$$

and take the Fourier transform on the right side. The result is

$$x_{\mathcal{F}}(f)h_{>,\mathcal{F}}(f) + x^*_{\mathcal{F}}(-f)h^*_{>,\mathcal{F}}(-f).$$

For positive frequencies the first term equals $x_{\mathcal{F}}(f)$ and the second term vanishes. Hence $\sqrt{2}\Re\left\{\hat{x}(t)\right\}$ and $x(t)$ agree for positive frequencies. Since they are real-valued they must agree everywhere.

To go from $\hat{x}(t)$ to the baseband-quivalent $x_E(t)$ we use the *frequency shift* property of the Fourier transform that we rewrite for reference:

$$x(t)\exp\{j2\pi f_0 t\} \longleftrightarrow x_{\mathcal{F}}(f - f_0).$$

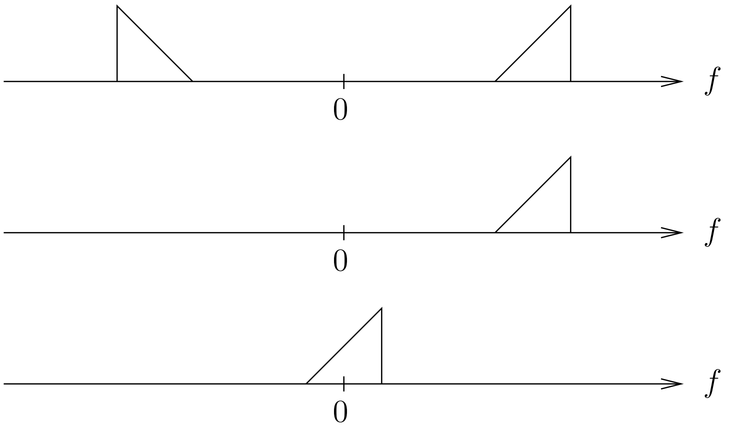
The *baseband-equivalent of* $x(t)$ is the signal

$$x_E(t) = \hat{x}(t)\exp\{-j2\pi f_0 t\}$$

and its Fourier transform is

$$x_{E,\mathcal{F}}(f) = \hat{x}_{\mathcal{F}}(f + f_0).$$

The transformation from $|x_{\mathcal{F}}(f)|$ to $|x_{E,\mathcal{F}}(f)|$ is depicted in the following frequency-domain representation (factor $\sqrt{2}$ omitted).



Going the other way is straightforward:

$$x(t) = \sqrt{2}\Re\big\{x_E(t)\exp\{j2\pi f_0 t\}\big\}.$$

In this subsection the signal $x(t)$ was real-valued but otherwise arbitrary.

When $x(t)$ is the transmitted signal $s(t)$, the concepts developed in this section lead to the relationship between $s(t)$ and its baseband-equivalent $s_E(t)$. In many implementations the sender first forms $s_E(t)$ and then it converts it to $s(t)$ via a stage that we refer to as the up-converter. The block diagram of Figure 1.2 reflects this approach. At the receiver the down-converter implements the reverse processing. Doing so is advantageous since the up/down-converters are hen the only transmitter/receiver stages that explicitly depend on the carrier frequency. As we will discuss, there are other implementation advantages to this approach. Up/down-conversion is discussed in the next section.

When the generic signal $x(t)$ is the channel impulse response $h(t)$ then, up to a scaling factor introduced for a valid reason, $h_E(t)$ becomes the baseband-equivalent impulse response. This idea, developed in the section following next, is useful to relate the baseband-equivalent received signal to the baseband-equivalent transmitted signal via a baseband-equivalnt channel model. The baseband-equivalent channel model is an abstraction that allows us to hide the passband issues in the channel model.

8.2 Up/Down Conversion

In this section we describe the top layer of Figure 1.2. To generate a passband signal, the transmitter first generates a complex-valued baseband signal $s_E(t)$. The signal is then converted to the signal $s(t)$ that has the desired center frequency f_0. This is done by means of the operation

$$s(t) = \sqrt{2}\Re\big\{s_E(t)\exp\{j2\pi f_0 t\}\big\}.$$

With the frequency domain in mind, the process is called *up-convertion*. We see that the signal $s_E(t)$ is the baseband-equivalent of the transmitted signal $s(t)$.

The up-converter block diagram is depicted in the top part of Figure 8.1. The rest of the figure shows the channel and the down-converter at the receiver leading to

$$R_E(t) = \sqrt{2}(R * h_>)(t)\exp\{-j2\pi f_0 t\}.$$

The signal $R_E(t)$ at the down-converter output is a sufficient statistic since the two operations performed by the donw-converter are reversible.

In the next section we model the channel between the up-converter input and the dwon-converter output.

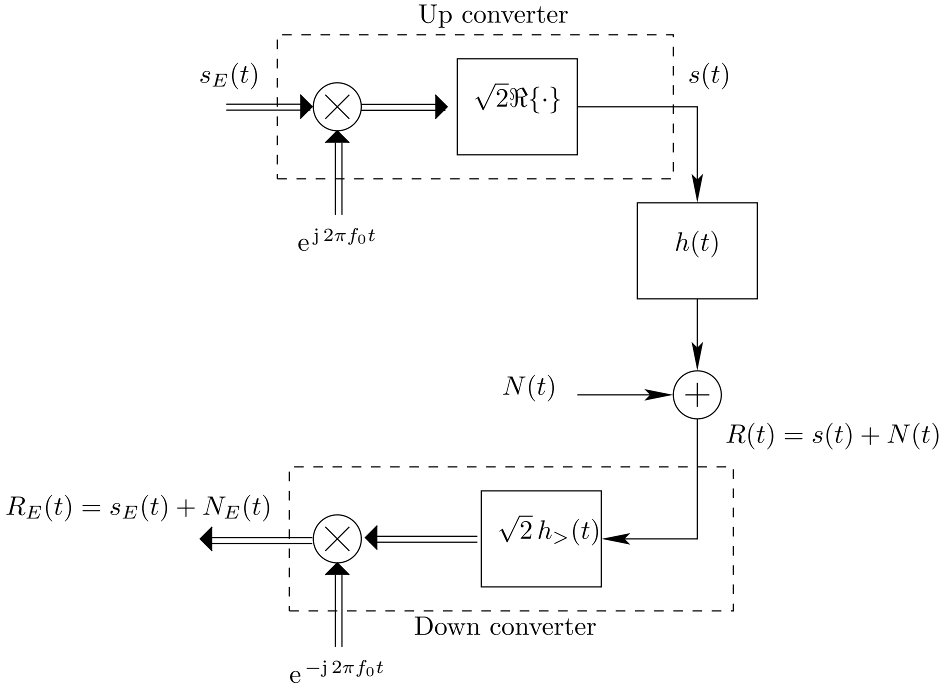


Figure 8.1: Up/down conversion. Double lines denote complex signals.

8.3 Baseband-Equivalent Channel Model

In this section we show that the channel seen from the up-converter input to the down-converter output is a baseband additive Gaussian noise channel as modeled in Figure 8.2. We start by describing the baseband-equivalent impulse response, denoted $\frac{h_E(t)}{\sqrt{2}}$ in the figure, and then turn our attention to the baseband-equivalent noise $N_E(t)$. Notice that we are allowed to study the signal and the noise separately since the system is linear, albeit time-varying.

Assume that the bandpass channel has an arbitrary real-valued impulse response $h(t)$. Without noise the input/output relationship is

$$r(t) = (h \star s)(t).$$

Taking the Fourier transform on both sides we get the first of the equations below. The other follow via straightforward manipulations of the first equality and the notions developed in the previous subsection.

$$\begin{aligned}
r_{\mathcal{F}}(f) &= h_{\mathcal{F}}(f) s_{\mathcal{F}}(f) \\
r_{\mathcal{F}}(f) h_{>,\mathcal{F}}(f)\sqrt{2} &= h_{\mathcal{F}}(f) h_{>,\mathcal{F}}(f) s_{\mathcal{F}}(f) h_{>,\mathcal{F}}(f)\sqrt{2} \\
\hat{r}_{\mathcal{F}}(f) &= \frac{\hat{h}_{\mathcal{F}}(f)}{\sqrt{2}} \hat{s}_{\mathcal{F}}(f) \\
r_{E,\mathcal{F}(f)} &= \frac{h_{E,\mathcal{F}}(f)}{\sqrt{2}} s_{E,\mathcal{F}}(f). \qquad (8.4)
\end{aligned}$$

Hence, when we send a signal $s(t)$ through a channel of impulse response $h(t)$ it is as sending the baseband equivalent signal $s_E(t)$ through a channel of baseband equivalent impulse response $\frac{h_E(t)}{\sqrt{2}}$. This is the baseband-equivalent channel impulse response.

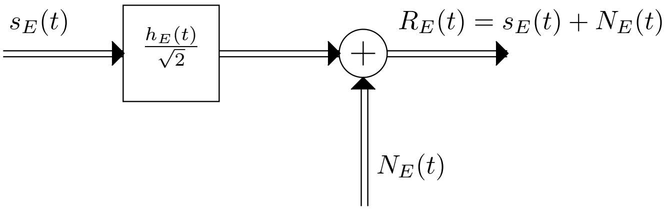


Figure 8.2: Baseband-equivalent channel model.

Let us now focus on the noise. With reference to Figure 8.1, observe that $N_E(t)$ is a zero-mean (complex-valued) Gaussian random process. Indeed it is obtained from linear (complex-valued) operations on Gaussian noise. Furthermore:

(a) The analytic equivalent $\hat{N}(t)$ of $N(t)$ is a Gaussian process since obtained by filtering a Gaussian noise. Its power spectral density is

$$\mathcal{S}_{\hat{N}}(f) = \mathcal{S}_N(f) \left|\sqrt{2}\, h_{>,\mathcal{F}}(f)\right|^2 = \begin{cases} 2\mathcal{S}_N(f), & f > 0 \\ \frac{1}{2}\mathcal{S}_N(f), & f = 0 \\ 0, & f < 0. \end{cases} \qquad (8.5)$$

(b) Let $N_E(t) = \hat{N}(t)\,\mathrm{e}^{-\mathrm{j}\,2\pi f_0 t}$ be the baseband-equivalent noise. The autocorrelation of $N_E(t)$ is given by:

$$\begin{aligned}
\mathcal{R}_{N_E}(\tau) &= E\left[\hat{N}(t+\tau)\,\mathrm{e}^{-\mathrm{j}\,2\pi f_0 (t+\tau)} \hat{N}^*(t)\,\mathrm{e}^{\mathrm{j}\,2\pi f_0 t}\right] \\
&= R_{\hat{N}}(\tau)\mathrm{e}^{-\mathrm{j}\,2\pi f_0 \tau} \qquad (8.6)
\end{aligned}$$

where we have used the fact that $\hat{N}(t)$ is WSS (since it is obtained from filtering a WSS process). We see that $N_E(t)$ is itself WSS. Its power spectral density is given by:

$$\mathcal{S}_{N_E}(f) = \mathcal{S}_{\hat{N}}(f+f_0) = \begin{cases} 2\mathcal{S}_N(f+f_0), & f > -f_0 \\ \frac{1}{2}\mathcal{S}_N(f+f_0), & f = -f_0 \\ 0, & f < -f_0. \end{cases} \tag{8.7}$$

(c) We now show that $\hat{N}(t)$ is proper.

$$\begin{aligned} E[\hat{N}(t)\hat{N}(s)] &= E\left[\int_{-\infty}^{+\infty} \sqrt{2}h_>(\alpha)N(t-\alpha)\,d\alpha \int_{-\infty}^{+\infty} \sqrt{2}h_>(\beta)N(s-\beta)\,d\beta\right] \\ &= 2\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} h_>(\alpha)h_>(\beta)\mathcal{R}_N(t-\alpha-s+\beta)\,d\alpha\,d\beta \\ &= 2\int_\alpha\int_\beta h_>(\alpha)h_>(\beta)\,d\alpha\,d\beta \int_{-\infty}^{+\infty} \mathcal{S}_N(f)\,\mathrm{e}^{\mathrm{j}\,2\pi f(t-\alpha-s+\beta)}\,df \\ &= 2\int_f \mathcal{S}_N(f)\,\mathrm{e}^{\mathrm{j}\,2\pi f(t-s)}h_{>,\mathcal{F}}(f)h_{>,\mathcal{F}}(-f)\,df \\ &= 0 \end{aligned} \tag{8.8}$$

since $h_{>,\mathcal{F}}(f)h_{>,\mathcal{F}}(f) = 0$ for all frequencies except for $f = 0$. Hence the integral vanishes. Thus $\hat{N}(t)$ is proper.

(d) $N_E(t)$ is also proper since

$$\begin{aligned} E[N_E(t)N_E(s)] &= E\left[\hat{N}(t)\,\mathrm{e}^{-\mathrm{j}\,2\pi f_0 t}\hat{N}(s)\,\mathrm{e}^{-\mathrm{j}\,2\pi f_0 s}\right] \\ &= \mathrm{e}^{-\mathrm{j}\,2\pi f_0(t+s)}E\left[\hat{N}(t)\hat{N}(s)\right] \\ &= 0 \end{aligned} \tag{8.9}$$

(We could have simply argued that $N_E(t)$ is proper since it is obtained from the proper process $\hat{N}(t)$ via a linear transformation.)

(e) The real and imaginary components of $N_E(t)$ have the same autocorrelation function. Indeed,

$$\begin{aligned} 0 = E[N_E(t)N_E(s)] &= E\left[(\Re\{N_E(t)\}\Re\{N_E(s)\} - \Im\{N_E(t)\}\Im\{N_E(s)\})\right. \\ &\quad \left.+\,\mathrm{j}\,(\Re\{N_E(t)\}\Im\{N_E(s)\} + \Im\{N_E(t)\}\Re\{N_E(s)\})\right] \end{aligned} \tag{8.10}$$

implies

$$E\left[(\Re\{N_E(t)\}\Re\{N_E(s)\}\right] = E\left[\Im\{N_E(t)\}\Im\{N_E(s)\}\right]$$

As claimed.

(f) Furthermore, if $\mathcal{S}_N(f_0 - f) = \mathcal{S}_N(f_0 + f)$) then the real and imaginary parts of $N_E(t)$ are uncorrelated, hence they are independent. To see this we expand as follows

$$\begin{aligned} E[N_E(t)N_E^*(s)] &= E\left[(\Re\{N_E(t)\}\Re\{N_E(s)\} + \Im\{N_E(t)\}\Im\{N_E(s)\})\right. \\ &\quad \left. - \mathrm{j}\,(\Re\{N_E(t)\}\Im\{N_E(s)\} - \Im\{N_E(t)\}\Re\{N_E(s)\})\right]. \end{aligned} \tag{8.11}$$

and observe that if the power spectral density of $N_E(t)$ is an even function then the autocorrelation of $N_E(t)$ is real-valued. Thus

$$E\left[\Re\{N_E(t)\}\Im\{N_E(s)\} - \Im\{N_E(t)\}\Re\{N_E(s)\}\right] = 0.$$

On the other hand, from (8.10) we have

$$E\left[\Re\{N_E(t)\}\Im\{N_E(s)\} + \Im\{N_E(t)\}\Re\{N_E(s)\}\right] = 0.$$

The last two expressions imply

$$E\left[\Re\{N_E(t)\}\Im\{N_E(s)\}\right] = E\left[\Im\{N_E(t)\}\Re\{N_E(s)\}\right] = 0,$$

which is what we have claimed.

We summarize what concerns the noise. $N_E(t)$ is a proper zero-mean Gaussian random process. Furthermore, from (8.7) we see that for the interval $f \geq -f_0$, the power spectral density of of $N_E(t)$ equals that of $N(t)$ translated towards baseband by f_0 and scaled by a factor 2. The fact that this relationship holds only for $f \geq -f_0$ is immaterial for all practical cases. Indeed, in practice, the center frequency f_0 is much larger than the signal bandwidth and any noise component which is outside the signal bandwidth will be eliminated by the front end receiver that projects the baseband equivalent of the received signal onto the baseband equivalent signal space. Even in suboptimal receivers that do not project the signal onto the signal space there is a front-end filter that eliminates the out of band noise. For these reasons we may simplify our expressions and assume that, for all frequencies, the the power spectral density of of $N_E(t)$ equals that of $N(t)$ translated towards baseband by f_0 and scaled by a factor 2.

To remember where the factor 2 goes, it suffices to keep in mind that the variance of the noise within the band of interest is the same for both processes. To find the variance of $N(t)$ in the band of interest we have to integrate its power spectral density over $2B$ Hz. For that of $N_E(t)$ we have to integrate over B Hz. Hence the power spectral density of $N_E(t)$ must be twice that of $N(t)$.

The real and imaginary parts of $N_E(t)$ have the same autocorrelation functions hence the same power spectral densities. If $\mathcal{S}(f)$ is symmetric with respect to f_0, then the real and imaginary parts of $N_E(t)$ are uncorrelated, and since they are Gaussian they are independent. In this case their power spectral density must be half that of $N_E(t)$.

8.4 Implications

What we have learned in this chapter has several implications. Let us take a look at what they are.

Signal Design: By signal design we mean the choice of parameters that specify $s(t)$. The signal design ideas and techniques we have learned in previous chapters apply generally to any additive white Gaussian noise channel, regardless whether it is baseband or passband. However, applying Nyquist criterion to a passband pulses $\psi(t)$ is a bit more tricky. (Point to a problem). In particular, for a desirable baseband power spectral density $\mathcal{S}(f)$ that "fulfills Nyqist criterion" in the sense that $|\psi_{\mathcal{F}}(f)|^2 = \frac{TS(f)}{\mathcal{E}}$ fulfills Nyquist criterion form some T and$\mathcal{E}$, it is possible to find a pulse $\tilde{\psi}(t)$ that fulfills Nyquist criterion and leads to the spectrum $\mathcal{S}(t-f_0)$ for some f_0 and not for others. This is annoying enough. What is also annoying is that if we rely on the pulse $\tilde{\psi}(t)$ to determine not only the "shape" but also the center frequency of the power spectral density then the pulse will depend on the center frequency. Fortunately the technique developed in the this chapter allows us to do the signal design assuming a baseband-equivalent signal which is independent of f_0.

Performance Analysis Once we have a tentative baseband-equivalent signal, the next step is to assess the resulting error probability. For this we need to have a channel model. The technique developed in this chapter allows us to determine the channel model seen by the baseband-equivalent signal. It is often the case that over the passband interval of interest the channel impulse response has a flat magnitude and linear phase. Then the baseband-equivalent impulse response has also a flat magnitude and linear phase around the origin. Since the transmitted signal is not affected by the channels frequency response outside the baseband interval of interest, we may as well assume that the magnitude is flat and the phase is linear over the entire frequency range. Hence we may assume that $\frac{h_E(t)}{\sqrt{2}} = a\delta(t-\tau)$ for some constants a and τ. The effect of the baseband equivalent channel is to scale the symbol alphabet by a and delay by τ. As we see in this example, the impulse response of the baseband equivalent channel need not be more complicated than that of the actual (passband) channel.

Implementation Decomposing the transmitter into a baseband transmitter and an up-converter is a good thing to do also for practical reasons. We summarize a few facts that justify this claim.

(a) *Task Decomposition* Senders and a a receivers are signal processing devices. In many situations, notably in wireless communication, the sender and the receiver exchange radio-frequency signals. Very few people are expert in implementing signal processing as well as in designing radio-frequency devices. Decomposing the design into baseband and a radio-frequency part makes it possible to partition the design task into subtasks that can be accomplished independently by people that have the appropriate skills.

(b) *Complexity* At various stages of a sender and a receiver there are filters and amplifiers. In the baseband transmitter those filters do not depend on f_0 and the amplifiers have to fulfill the design specifications only over the frequency range occupied by the baseband-equivalent signal. Such filters and amplifiers would be more expensive if their characteristic depended on f_0.

(c) *Oscillation* The "ringing" produced by a sound system when the microphone is placed too close to the speaker is a well-known effect that occurs when the signal at the output of an amplifier manages to feed back to the amplifier input. When this happens the amplifiers turns into an oscillator. The problem is particularly challenging in dealing with radio-frequency amplifiers. In fact a wire of appropriate length can act as a transmit or as a receive antenna of a radio-frequency signal. The signal produced by the up-converter output is sufficiently strong to travel long-distance over the air, which means that in principle it can easily feed back to the up-converter input. This is not a problem though since the input has a filter that passes "only" baseband signals.

8.5 Problems

In the problem session we should have a problem that shows how a delay causes the baseband equivalent signal to rotate.

Problem 1. (Fourier Transform)

(a) *Prove that if $x(t)$ is a real-valued signal, then its Fourier transform $X(f)$ satisfies the symmetric property*

$$X(f) = X^*(-f) \quad \text{(Symmetry Property)}$$

where X^* is the complex conjugate of X.

(b) *Prove that if $x(t)$ is a purely imaginary-valued signal, then its Fourier transform $X(f)$ satisfies the anti-symmetry property*

$$X(f) = -X^*(-f) \quad \text{(Anti-Symmetry Property)}$$

Problem 2. (Baseband Equivalent Relationship) *In this problem we neglect noise and consider the situation in which we transmit a signal $X(t)$ and receive*

$$R(t) = \sum_i \alpha_i X(t - \tau_i).$$

Show that the baseband equivalent relationship is

$$R_E(t) = \sum_i \beta_i \, X_E(t - \tau_i).$$

Express β_i explicitly.

Problem 3. (Equivalent Representations) *A bandpass signal $x(t)$ may be written as $x(t) = \sqrt{2}\Re\{x_E(t)e^{j2\pi f_0 t}\}$, where $x_E(t)$ is the baseband equivalent of $x(t)$.*

(a) *Show that a signal $x(t)$ can also be written as $x(t) = a(t)\cos[2\pi f_0 t + \theta(t)]$ and describe $a(t)$ and $\theta(t)$ in terms of $x_E(t)$. Interpret this result.*

(b) *Show that the signal $x(t)$ can also be written as $x(t) = x_{EI}(t)\cos 2\pi f_0 t - x_{EQ}(t)\ \sin(2\pi f_0 t)$, and describe $x_{EI}(t)$ and $x_{EQ}(t)$ in terms of $x_E(t)$. (This shows how you can obtain $x(t)$ without doing complex-valued operations.)*

(c) *Find the baseband equivalent of the signal $x(t) = A(t)\cos(2\pi f_0 t + \varphi)$, where $A(t)$ is a real-valued lowpass signal. Hint: You may find it easier to guess an answer and verify that it is correct.*

Problem 4. (Equivalent Baseband Signal)

(a) *You are given a "passband" signal $\psi(t)$ whose spectrum is centered around f_0. Write down in a generic form the different steps needed to find the baseband equivalent signal.*

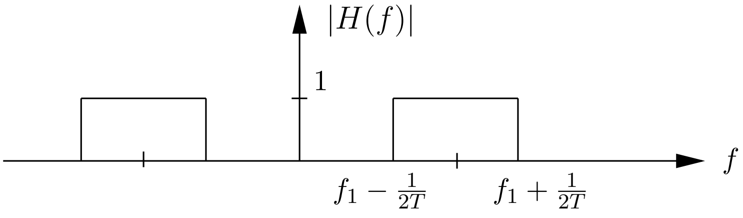
(b) *Consider the waveform*

$$\psi(t) = \text{sinc}\left(\frac{t}{T}\right)\cos(2\pi f_0 t).$$

What is the equivalent baseband signal of this waveform.

(c) *Assume that the signal $\psi(t)$ is passed through the filter with impluse response $h(t)$ where $h(t)$ is specified by its baseband equivalent impulse response $h_E(t) = \frac{1}{T\sqrt{2}}\text{sinc}^2\left(\frac{t}{2T}\right)$. What is the output signal, both in passband as well as in baseband?* Hint: The Fourier transform of $\cos(2\pi f_0 t)$ is $\frac{1}{2}\delta(f - f_0) + \frac{1}{2}\delta(f + f_0)$. The Fourier transform of $\frac{1}{T}\text{sinc}(\frac{t}{T})$ is equal to $\mathbf{1}_{[-\frac{1}{2T},\frac{1}{2T}]}(f)$ with $\mathbf{1}_{[-\frac{1}{2T},\frac{1}{2T}]}(f) = 1$ if $f \in [-\frac{1}{2T}, \frac{1}{2T}]$ and 0 otherwise.

Problem 5. (Up-Down Conversion) *We want to send a "passband" signal $\psi(t)$ whose spectrum is centered around f_0, through a waveform channel defined by its impulse response $h(t)$. The Fourier transform $H(f)$ of the impulse response is given by*



where $f_1 \neq f_0$.

(a) *Write down in a generic form the different steps needed to send $\psi(t)$ at the correct frequency f_1.*

(b) *Consider the waveform*

$$\psi(t) = sinc\left(\frac{t}{T}\right)\cos(2\pi f_0 t).$$

What is the output signal, in passband (at center frequency f_1) as well as in baseband?

(c) *Assume that $f_0 = f_1 + \epsilon$, with $\epsilon \ll \frac{1}{2T}$, and that the signal $\psi(t)$ is directly transmitted without any frequency shift. What will be the central frequency of the output signal?* Hint: The Fourier transform of $\cos(2\pi f_0 t)$ is $\frac{1}{2}\delta(f - f_0) + \frac{1}{2}\delta(f + f_0)$. The Fourier transform of $\frac{1}{T}\text{sinc}(\frac{t}{T})$ is equal to $\mathbf{1}_{[-\frac{1}{2T},\frac{1}{2T}]}(f)$ with $\mathbf{1}_{[-\frac{1}{2T},\frac{1}{2T}]}(f) = 1$ if $f \in [-\frac{1}{2T}, \frac{1}{2T}]$ and 0 otherwise.

Problem 6. (Smoothness of Bandlimited Signals) *In communications one often finds the statement that if $s(t)$ is a signal of bandwidth W, then it can't vary too much in a small interval $\tau << 1/W$. Based on this, people sometimes substitute $s(t)$ for $s(t+\tau)$. In this problem we will derive an upper bound for $|s(t+\tau) - s(t)|$. It is assumed that $s(t)$ is a finite energy signal with Fourier transform satisfying $S(f) = 0$, $|f| > W$.*

(a) *Let $H(f)$ be the frequency response of the ideal lowpass-filter defined as 1 for $|f| \leq W$ and 0 otherwise. Show that*

$$s(t+\tau) - s(t) = \int s(\xi)[h(t+\tau-\xi) - h(t-\xi)]d\xi. \tag{8.12}$$

(b) *Use Schwarz inequality to prove that*

$$|s(t+\tau) - s(t)|^2 \leq 2E_s[E_h - R_h(\tau)], \tag{8.13}$$

where E_s is the energy of $s(t)$,

$$R_h(\tau) = \int h(\xi+\tau)h(\xi)d\xi$$

is the (time) autocorrelation function of $h(t)$, and $E_h = R_h(0)$.

(c) *Show that $R_h(\tau) = h * h(\tau)$, i.e., for h the convolution with itself equals its autocorrelation function. What makes h have this property?*

(d) *Show that $R_h(\tau) = h(\tau)$.*

(e) *Put things together to derive the upperbound*

$$|s(t+\tau)-s(t)| \leq \sqrt{2E_s[E_h - h(\tau)]} = \sqrt{4WE_s\left(1-\frac{\sin(2\pi W\tau)}{2\pi W\tau}\right)}. \tag{8.14}$$

[Can you determine the impulse response $h(t)$ without looking it up an without solving integrals? Remember the "mnemonics" given in class?] Verify that for $\tau = 0$ the bound is tight.

(f) *Let E_D be the energy in the difference signal $s(t+\tau)-s(t)$. Assume that the duration of $s(t)$ is T and determine an upperbound on E_D.*

(g) *Consider a signal $s(t)$ with parameters $2W = 5$ Mhz and $T = 5/2W$. Find a numerical value T_m for the time difference τ so that $E_D(\tau) \leq 10^{-2}E_s$ for $|\tau| \leq T_m$.*

Bibliography

[1] R. A. Horn and C. R. Johnson, *Matrix analysis.* Cambridge: Cambridge University Press, 1999.

[2] W. Feller, *An Introduction to Probability Theory and its Applications*, vol. II. New York: Wiley, 1966.