Principles Of Digital Communications

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

14.3.2007

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 that they are today. Perhaps one day the input/output interface will separate form the communication/computation hardware. The former will be the only part that we actually carry around 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 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 –hence making keys, credit cards, and wallets obsolete. Getting there will require joint efforts form 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 corrupts it with

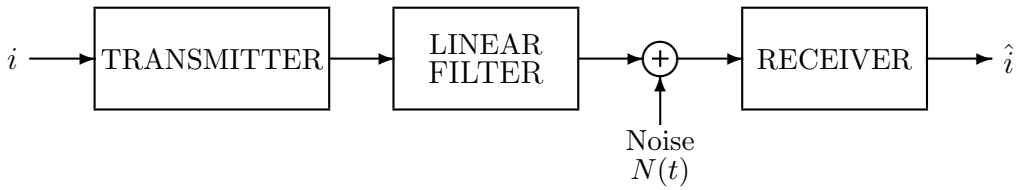


Figure 1.1: Basic point-to-point communication system over a bandlimited Gaussian channel.

noise. The noise is Gaussian since it represents the contribution of various noise sources.[1] The filter in 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 to enforce this constraint is to tell the system designers 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] In some cases the transmit and/or the receive antennas also filter the signal. This is the case for instance when the signal's bandwidth is sufficiently large that the antenna characteristic is not constant over the frequency interval spanned by the signal. The filter in Figure 1.1 accounts for these and possibly other linear time-invariant transformations that acts 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 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. (More on this later.) Without loss of generality we can let the message set consist of the integers $\{0, 1, \ldots, m-1\}$. The elements of the message set are called messages. There is a one-to-one correspondence between messages

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

and elements of the signal set. The "nature" (e.g. discrete vs continuous time) of the signals is important since signals have to be compatible with the channel. The channel is always assumed to be given to the designer who has no control over it. By assumption, the designer can only control 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 $\hat{i}(y)$ of i. Hence the receiver is a map from the space of channel output signals to the message set. Hopefully $i = \hat{i}$ 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 semmed 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. Conceptually, we may label every message with a unique sequence of $\log m$ bits so that communicating the message is equivalent to communicating the corresponding bit sequence. (This is why earlier we said that the nature of the messages is not relevant). Hence 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 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, and know which tradeoffs you have as a system designer.

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 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 me 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 –not always without making errors. At this level of granularity the visible performance figures are the cardinality of the message set, how long we have to wait until we are allowed to choose the next message, and the probability of error. The first two determine how many bits we send per unit of time, i.e., the rate 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)

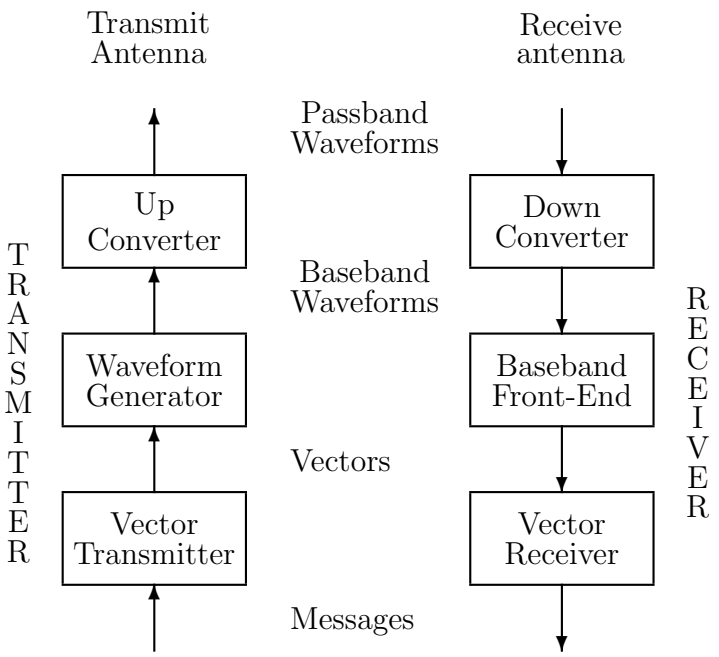


Figure 1.2: Decomposed transmitter and receiver.

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 reach the final stage of the receiver would not be a good idea since it makes little sense to study the transmitter design without having an appreciation of the task and limitations of a receiver. We will instead make many passes over the block diagram of Figure 1.2, each time at a different level and 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 engineer and mathematician of the 20th century, was knows 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.

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* and *Coding Theory*. For the student interested in hands-on experience, EPFL offers *Software-Defined Radio: A Hands On Course*.

Networking is is another branch of communications that has developed almost independently of the material treated in this class. It relies on quite 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, e.g., due to network congestion, hardware failure, queue overflow, etc. 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 paragraph 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." This paragraph will also give you an idea of which fundamental concepts will play a role in this course. Chapter 2 deals with the vector channel case of Figure 1.2. The emphasis will be on the design of an optimal Vector Receiver, assuming that the Vector Transmitter and the Vector Channel are given. This 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 Gaussian Vector Channel. (In Chapter 8 you will realize that the Gaussiann Vector 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 Geneartor is the notion of orthonormal expansion from linear algebra. We will fix an orthonormal basis and we will let the output of the Vector Transmitter be the vector of coefficients that determine the signal produced by the Waveform Transmitter (with respect to the given orthonormal basis). The Baseband Front-End of the receiver reduces the received waveform to a vector (n-tuple) that contains just as much information as needed to decide about the message selected by the sender. To do so the Baseband Front-End projects the received waveform onto each element of mentioned orthonormal basis. The resulting n-tuple is passed to the Vector Receiver. Together the Vector Transmitter and the Waveform Generator form the Waveform Transmitter. Together the Baseband Front-End and the Vector Receiver form the Waveform Receiver. What we do in Chapter

3 holds irrespectively of the specific set of signals that we use to communicate. Chapter 4 deals with general high level implications of a specific signal set. Chapter 5 deals with the problem of choosing a convenient orthonormal basis for the Waveform Generators, namely one that leads to signals that have a desirable power spectral density and that significantly simplifies the complexity of the Baseband Front-End. The main concept here is what is called Nyquist criterion. Chapter 6 deals with the Up/Down Converters. 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 what we have called the Waveform Transmitter and the Waveform Receiver. This will be done using one of the fundamental properties of Fourier transforms. Given our ability to shift the center frequency of the transmitted signal to any desired location, it makes sense to let the Waveform Transmitter and the Waveform Receiver operate in some fixed frequency range if this simplifies their implementation. 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 kind of problems, including the fact that signals 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 let the Waveform Transmitter and Waveform Receiver operate at "baseband", i.e. process signals that have $f = 0$ as their center frequency. As it turns out, the baseband representation of a general signal is complex-valued, even if the signal being represented is real-valued. This means that the Waveform 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 in terms of having to learn how to deal with complex-valued stochastic processes and complex-valued random vectors. Dealing with complex-valued Gaussian processes and vectors is the topic of Chapter 7. Chapter 8 "closes the loop" by showing that the channel "seen" by the Vector Transmitter and the Vector Receiver is indeed the abstract Gaussian Vector Channel that we have assumed in Chapter 2. To emphasize the importance of the Vector Channel we mention that in a typical information theory course (mandatory at the master-level at EPFL) as well as in a typical coding theory course (offered at EPFL in the Ph.D. program), the channel is a Vector Channel (perhaps not called this way) and one takes it for granted that the student knows where it comes from. (The material treated in this class is also assumed as being assimilated in *Advanced Digital Communications* as well as in *Software-Defined Radio: A Hands on Course*, both of which are offered at EPFL at the master level.) Chapter 9 contains is a case study on coding. The communication model is that of Chapter 2 with the Vector Channel being Gaussian. The Vector Transmitter will incorporate a convolutional encoder and the Vector Receiver will be based on the Viterbi algorithm. The performance of the resulting scheme will be analyzed and compared to the uncoded case.

Chapter 2

Optimal Receivers for Vector Channels

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 vector channel. In the next chapter it will become clear why the vector 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 vector receiver when the channel and the transmitter are given. The channel considered in this chapter may be more general than the vector channel of Figure 2.1.

The communication system of interest in this chapter is depicted in Figure 2.2. Its components are:

- The source: It is responsible to produce the message $H \in \mathcal{H} = \{0, 1, \ldots, (m-1)\}$. 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 for a communication system. 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 and/or $m = 2$.

- The transmitter: It is a mapping from $\mathcal{H}$ to $\mathcal{S} = \{\boldsymbol{s_0}, \boldsymbol{s_1}, \ldots, \boldsymbol{s_{m-1}}\}$ where $\boldsymbol{s_i} \in \mathbb{C}^n$ for some n. (We will start with $s_i \in \mathbb{R}^n$ but we will see in the last chapter that it is crucial that we allow for $s_i \in \mathbb{C}^n$).

- The channel: It is described by the probability density of the output for each of the

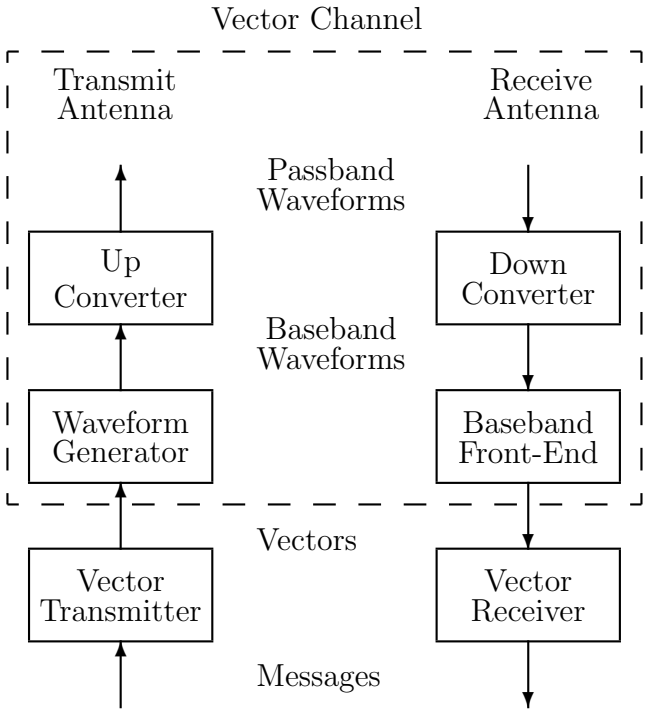


Figure 2.1: Vector channel abstraction.

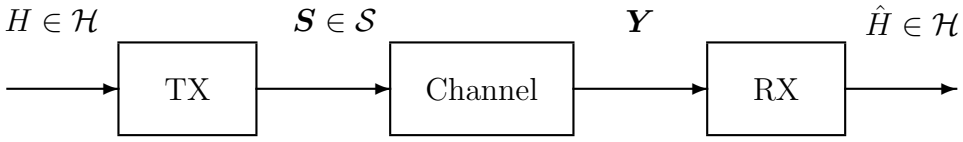


Figure 2.2: Main setup considered Part I.

possible inputs. When the channel input is $\boldsymbol{s}_i$, the probability density of $\boldsymbol{Y}$ will be denoted by $f_{\boldsymbol{Y}|\boldsymbol{S}}(\boldsymbol{y}|\boldsymbol{s}_i)$.

- The 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. This is the so-called *hypothesis testing* problem that comes up in various contexts (not only in communication).

First we give a few examples.

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 additive white Gaussian (AWGN) channel 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 with independent components.*

Specifying the decision rule implemented by the receiver is straightforward once we understand the hypothesis testing problem studied in the next section.

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 on a finite alphabet $\mathcal{H} = \{0, 1, \ldots, m-1\}$, from the outcome of some related random variable Y. The random variable H is called the *Hypothesis* and Y the *observation*.

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. The transmitter sends a signal (typically a distinct signal for each letter of $\mathcal{H}$) and the receiver observes the channel response Y. The receiver decides the value of H based on Y. 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 that makes $P_c = Pr\{\hat{H} = H\}$ as large as possible.[1]

The standard assumption is that we know the *a priori* probability P_H and for each $i \in \mathcal{H}$ we know the conditional probability density function[2] (pdf) $f_{Y|H}(y|i)$ of Y.

EXAMPLE 4. *Here is a good example of a typical hypothesis testing problem. The problem is that of communicating one bit of information (or more generally a sequence of bits) across an optical fiber as shown in the following picture. 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 a LED whose 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* $H = 0$ *or* $H = 1$. *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*

[1] Pr is a short-hand for *probability of the enclosed event.*

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

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:

$$H = 0, \quad Y \sim P_{Y|H}(y|0) = \frac{\lambda_0^y}{y!} e^{-\lambda_0}$$
$$H = 1, \quad Y \sim P_{Y|H}(y|1) = \frac{\lambda_1^y}{y!} e^{-\lambda_1}$$

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

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. □

The relevant quantities may be summarized in the following relationship:

$$\underset{P_H(i)}{H} \longrightarrow \underset{f_{Y|H}(y|i)}{Y} \longrightarrow \hat{H}$$

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 $f_Y(y) = \sum_i P_H(i) f_{Y|H}(y|i)$. $P_{H|Y}(i|y)$ is the *posterior* (also called *a posteriori probability* of H given Y). Once we have observed that $Y = y$, the probability that $H = i$ becomes $P_{H|Y}(i|y)$.

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

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

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.,

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

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*:

$$\hat{H}(y) = \arg\max_i f_{Y|H}(y|i) \qquad \textit{(ML decision rule).} \tag{2.2}$$

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

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

An equivalent rule is

$$\Lambda(y) = \frac{f_{Y|H}(y|1)}{f_{Y|H}(y|0)} \underset{\hat{H}=0}{\overset{\hat{H}=1}{\gtrless}} \frac{P_H(0)}{P_H(1)} = \eta \qquad \textit{(binary MAP rule).} \tag{2.3}$$

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 the region $\{y : \hat{H}(y) = 0\}$ becomes bigger. This is intuitive.

When $P_H(0) = P_H(1) = 1/2$ the threshold becomes unity and the MAP test becomes a ML test that may be written as

$$f_{Y|H}(y|1) \underset{\hat{H}=0}{\overset{\hat{H}=1}{\gtrless}} f_{Y|H}(y|0) \qquad \textit{(binary ML rule).}$$

The decoding region $\mathcal{R}_i$ is the set of y for which the decision is $\hat{H} = i$, $i \in \{0,1\}$.

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 $H = i$,

$$P_e(0) = Pr\{Y \in \mathcal{R}_1 | H = 0\} = \int_{\mathcal{R}_1} f_{Y|H}(y|0)dy \tag{2.4}$$

$$= Pr\{\Lambda(Y) \geq \eta | H = 0\}. \tag{2.5}$$

Whether it is easier to work with the right side of (2.4) or of (2.5) 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 $P_c(1)$. The unconditional error probability is then

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

From (2.3) we see that, for the purpose of performing a MAP test, having $\Lambda(Y)$ is as good as having the observable Y. Any random variable obtained from Y that has this property is called a *sufficient statistic*. More on this later.

2.3 The Q function

The Q function is defined as:

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

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 $Z \sim \mathcal{N}(m, \sigma^2)$, then the probability $Pr\{Z \geq x\}$ can be written using the Q function by noticing that $\{Z \geq x\}$ 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 $Q(x)$.

(a) If $Z \sim \mathcal{N}(0,1)$, $F_Z(z) = Pr\{Z \leq z\} = 1 - Q(z)$. (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 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 omitted. 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

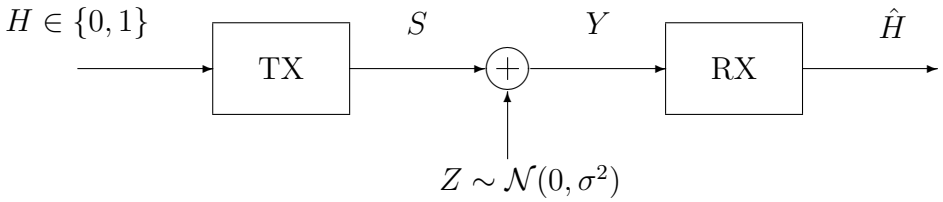
$$\begin{aligned}\frac{Q(\xi)}{2} &= \int_0^{\frac{\pi}{2}} \int_{\frac{\xi}{\sin\theta}}^{\infty} \frac{e^{-\frac{r^2}{2}}}{2\pi} r dr d\theta \\ &= \frac{1}{2\pi} \int_0^{\frac{\pi}{2}} \int_{\frac{\xi^2}{2\sin^2\theta}}^{\infty} e^{-t} dt d\theta \\ &= \frac{1}{2\pi} \int_0^{\frac{\pi}{2}} e^{-\frac{\xi^2}{2\sin^2\theta}}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

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 Binary Hypothesis, Scalar Gaussian Channel

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 say this is

$$f_{Y|H}(y \mid 0) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y-a)^2}{2\sigma^2}}$$

$$f_{Y|H}(y \mid 1) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y-b)^2}{2\sigma^2}}.$$

We compute the likelihood ratio

$$\begin{aligned}\Lambda(y) = \frac{f_{Y|H}(y \mid 1)}{f_{Y|H}(y \mid 0)} &= e^{-\frac{(y-b)^2-(y-a)^2}{2\sigma^2}} \\ &= e^{\frac{b-a}{\sigma^2}(y-\frac{a+b}{2})}\end{aligned}$$

The threshold is $\eta = \frac{P_0}{P_1}$. Now we have all the ingredients for the MAP rule.

Comparing $\Lambda(y)$ to η is the same as comparing $\log \Lambda(y)$ to $\log \eta$. The function $\log \Lambda(y)$ is called *log likelihood ratio*. Hence the MAP decision rule is

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

If $b > a$, 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

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

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

$$\begin{aligned}P_e(0) &= Pr\{Y > \theta \mid H = 0\} \\ &= \int_{\mathcal{R}_1} f_{Y|H}(y \mid 0) dy\end{aligned}$$

This is the probability that a Gaussian random variable with mean a and variance σ^2 exceeds the threshold θ. From our review on the Q function we know immediately that

$$P_e(0) = Q\left(\frac{\theta - a}{\sigma}\right).$$

Similarly,

$$P_c(1) = Q\left(\frac{b - \theta}{\sigma}\right).$$

Finally,

$$P_e = P_H(0) Q\left(\frac{\theta - a}{\sigma}\right) + P_H(1) Q\left(\frac{b - \theta}{\sigma}\right).$$

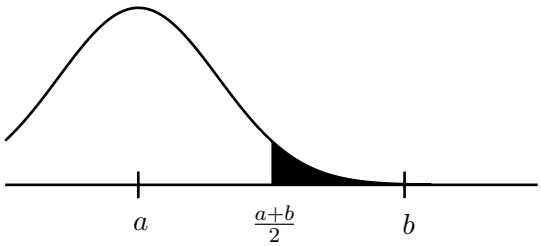


Figure 2.3: The probability of error when $H = 0$ is the black area. It is the probability that the noise makes y exceed the threshold when $H = 0$. The value of the threshold, half way between a and b, is determined assuming $P_H(0) = P_H(1)$.

The most common case is when $P_H(0) = P_H(1) = 1/2$. Then $\frac{\theta-a}{\sigma} = \frac{b-\theta}{\sigma} = \frac{b-a}{2\sigma}$ and

$$P_e = Q\left(\frac{b-a}{2\sigma}\right).$$

Notice that $\frac{b-a}{2}$ is the distance between a (or b) and the threshold. The following picture, which holds for $P_H(0) = P_H(1)$, leads immediately to P_e. Make sure that you understand it. It will be used very frequently.

2.5 Binary Hypothesis, Vector Gaussian Channel

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 over $\mathbb{R}$.

We now assume that the hypothesis i is mapped into the transmitter output $X(i)$ defined by

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

We also assume that $\boldsymbol{Z} \sim \mathcal{N}(0, \sigma^2 I_n)$.

As we did earlier, we start writing down the output statistic for each hypothesis

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

$$H = 1: \quad \boldsymbol{Y} = \boldsymbol{b} + \boldsymbol{Z} \sim \mathcal{N}(\boldsymbol{b}, \sigma^2 I_n).$$

Recall that

$$\begin{aligned} f_Z(\boldsymbol{z}) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{z_i^2}{2\sigma^2}} \\ &= \frac{1}{(2\pi\sigma^2)^{n/2}} e^{-\frac{\sum z_i^2}{2\sigma^2}} \\ &= \frac{1}{(2\pi\sigma^2)^{n/2}} e^{-\frac{\|\boldsymbol{z}\|^2}{2\sigma^2}}. \end{aligned}$$

Similarly,

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

Hence

$$\begin{aligned} \Lambda(\boldsymbol{y}) &= \frac{f_{\boldsymbol{Y}|H}(\boldsymbol{y} \mid 1)}{f_{\boldsymbol{Y}|H}(\boldsymbol{y} \mid 0)} = e^{\frac{\|\boldsymbol{y}-\boldsymbol{a}\|^2 - \|\boldsymbol{y}-\boldsymbol{b}\|^2}{2\sigma^2}}, \\ LLR(\boldsymbol{y}) &= \frac{\| \boldsymbol{y}-\boldsymbol{a} \|^2 - \| \boldsymbol{y}-\boldsymbol{b} \|^2}{2\sigma^2} \\ &= \frac{\| \boldsymbol{a} \|^2 - \| \boldsymbol{b} \|^2}{2\sigma^2} + \frac{1}{\sigma^2}\langle \boldsymbol{y}, \boldsymbol{b}-\boldsymbol{a} \rangle, \end{aligned}$$

and the MAP rule is

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

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

$$\{\boldsymbol{y} \in \mathbb{R}^n : \langle \boldsymbol{y}, \boldsymbol{u} \rangle = \phi\}$$

where $\boldsymbol{u} = \boldsymbol{b} - \boldsymbol{a}$.

When $P_H(0) = P_H(1) = 1/2$, the separating hyperplane separates the points that are closer to $\boldsymbol{a}$ from those that are closer to $\boldsymbol{b}$. We see this by solving for $\boldsymbol{y}$ in

$$LLR(\boldsymbol{y}) = \ln \eta$$

when $\ln \eta = 0$. The $\boldsymbol{y}$ that satisfy this relationship are the one for which

$$\| \boldsymbol{y}-\boldsymbol{a} \|^2 - \| \boldsymbol{y}-\boldsymbol{b} \|^2 = 0.$$

These are the $\boldsymbol{y}$ 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.

We also see that the separating hyperplane moves towards $\boldsymbol{b}$ when ϕ increases, which is the case when $\frac{P_H(0)}{P_H(1)}$ increases. This makes sense: if the prior probability becomes more in favor of $H = 0$ then the decoding region $\mathcal{R}_0$ becomes larger. Moreover, if $\frac{P_H(0)}{P_H(1)}$ exceeds 1, then $\ln \eta$ is positive and ϕ increases with σ^2. This also makes sense: as the observation becomes noisier, we pay more attention to the prior (which favors $H = 0$).

2.6 Multi-Hypothesis Testing

In Section 2.2 we have defined the hypothesis testing problem and derived the maximum a posteriori (MAP) and maximum likelihood (ML) decision rules. This was done for the general case of m hypotheses, that is when $\mathcal{H} = \{0, 1, \ldots, (m-1)\}$. We then turned our attention to binary hypotheses, i.e. $\mathcal{H} = \{0, 1\}$, and deepened our understanding paying particular attention to the special case in which the observation Y is a Gaussian random variable (or random vector $\boldsymbol{Y}$) whose mean depends on H. Now we go back to the m hypothesis testing problem.

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

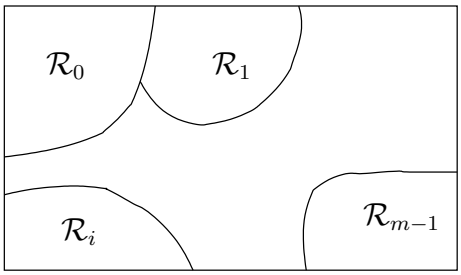
$$\begin{aligned}\hat{H}_{MAP}(\boldsymbol{y}) &= \arg\max_i P_{H|\boldsymbol{Y}}(i|\boldsymbol{y}) \\ &= \arg\max_i \frac{f_{\boldsymbol{Y}|H}(\boldsymbol{y}|i)P_H(i)}{f_{\boldsymbol{Y}}(\boldsymbol{y})} \\ &= \arg\max_i f_{\boldsymbol{Y}|H}(\boldsymbol{y}|i)P_H(i),\end{aligned}$$

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 in 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 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.,

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

In all cases considered here, $f_{\boldsymbol{Y}|H}$ will be known. If the transmitter maps the hypothesis i into the channel input $\boldsymbol{s}_i$, then $f_{\boldsymbol{Y}|H}(\boldsymbol{y}|i) = f_{\boldsymbol{Y}|\boldsymbol{X}}(\boldsymbol{y}|\boldsymbol{s}_i)$, where $f_{\boldsymbol{Y}|\boldsymbol{X}}(\cdot|\boldsymbol{x})$, also denoted



by $f_{\boldsymbol{Y}|\boldsymbol{x}}$, is the probability density function of the channel output when the channel input is $\boldsymbol{x}$.

Note that the decision (or decoding) function $\hat{H}$ assigns an $i \in \mathcal{H}$ to each $\boldsymbol{y} \in \mathbb{R}^n$. It can be equivalently described by the decision (or decoding) 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 of deciding correctly P_c.

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

Now assume the AWGN channel. When $H = i$, $i \in \mathcal{H}$, let $\boldsymbol{S} = \boldsymbol{s}_i$. Assume $P_H(i) = \frac{1}{m}$ (this is a common assumption in communications). The ML decision rule is

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

Hence a ML decision rule for the AWGN channel is a minimum-distance decision rule as shown in Figure 2.4.

Up to ties, $\mathcal{R}_i$ corresponds to the Voronoi region of $\boldsymbol{s}_i$. The Voronoi region of $\boldsymbol{s}_i$ is 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 2.5 shows the signal points and the decoding regions for 6-ary Pulse Amplitude Modulation (why the name makes sense will become clear in the next chapter).*

The error probability for each hypothesis is

$$P_e(0) = P_e(5) = Pr\{Z < -\frac{d}{2}\} = Q(\frac{d}{2\sigma}).$$

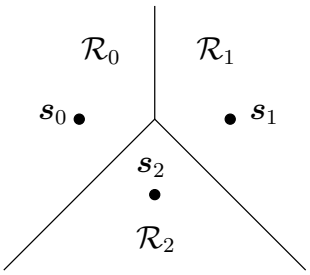


Figure 2.4: Example of Voronoi regions.

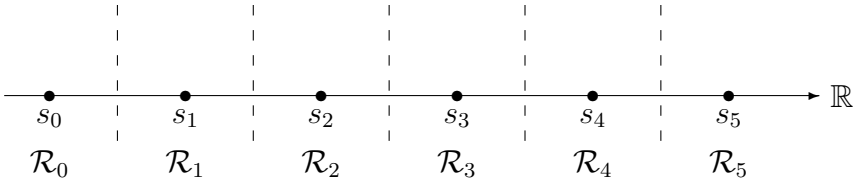


Figure 2.5: PAM signal constellation.

For $i \in \{1,2,3,4\}$,

$$\begin{aligned} P_e(i) &= Pr\{\{Z \geq \frac{d}{2}\} \cup \{Z < -\frac{d}{2}\}\} \\ &= 2Pr\{Z \geq \frac{d}{2}\} = 2Q(\frac{d}{2\sigma}) \end{aligned}$$

Where in the last equality we used the fact that the events under consideration are disjoint. Finally,

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

EXAMPLE 6. *(4-ary QAM) Figure 2.6 shows the signal set for 4-ary Quadrature Amplitude Modulation (QAM).*

We compute the probability of error as follows. First we observe that, due to symmetry,

$$P_e = P_e(0).$$

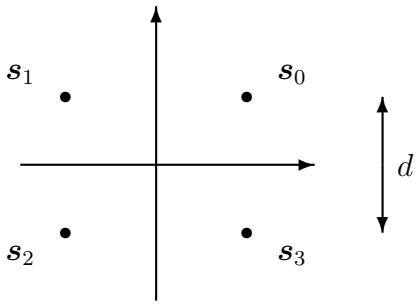


Figure 2.6: QAM signal constellation.

Furthermore,

$$\begin{aligned} P_c(0) &= Pr\left\{\{Z_1 \geq -\frac{d}{2}\} \cap \{Z_2 \geq -\frac{d}{2}\}\right\} \\ &= \left[Pr\{Z_i \geq -\frac{d}{2}\}\right]^2 \\ &= Q^2\left(-\frac{d}{2\sigma}\right) \\ &= \left[1 - Q\left(\frac{d}{2\sigma}\right)\right]^2. \end{aligned}$$

Hence,

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

When decoding regions are rectangular as in this example, one can easily express the error probability by means of the Q function . □

2.7 Union of Events 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 of Events Bound*

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

We now apply the union of events bound to approximate the probability of error in multi-hypothesis testing. Recall that

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

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

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

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

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

with equality if ties are always resolved against i. In fact, the right side contains all the ties (by definition) whereas the left side may or may not contain them.

Now we use the union of events bound:

$$\begin{aligned} P_e(i) &= Pr\{\boldsymbol{Y} \in \mathcal{R}_i^c | H = i\} \\ &\leq Pr\left\{\boldsymbol{Y} \in \bigcup_{j:j\neq i} \mathcal{B}_{i,j} | H = i\right\} \\ &\leq \sum_{j:j\neq i} Pr\{\boldsymbol{Y} \in \mathcal{B}_{i,j} | H = i\} \qquad (2.6) \\ &= \sum_{j:j\neq i} \int_{\mathcal{B}_{i,j}} f_{\boldsymbol{Y}|H}(\boldsymbol{y}|i) d\boldsymbol{y}. \end{aligned}$$

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, for the AWGN channel and ML decision rule

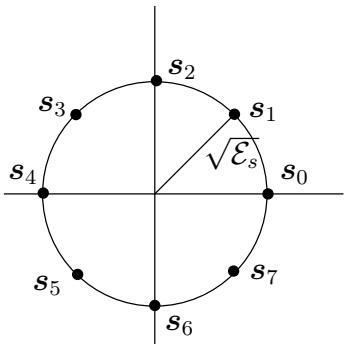
$$\int_{\mathcal{B}_{i,j}} f_{\boldsymbol{Y}|H}(\boldsymbol{y}|i) d\boldsymbol{y} = Q\left(\frac{\| \boldsymbol{s}_j - \boldsymbol{s}_i \|}{2\sigma}\right).$$

Moreover, in the next section we derive an easy-to-compute tight upperbound on

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

Notice that the above integral is the probability of error under $H = i$ when there are only two hypotheses and the other hypothesis is $H = j$.

EXAMPLE 7. *(m-PSK) The figure below 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*

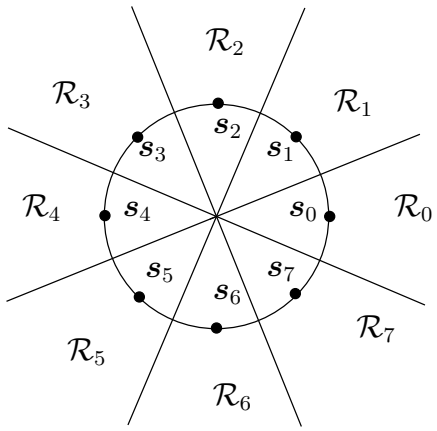
$$\boldsymbol{s}_i = \sqrt{\mathcal{E}_s} \begin{pmatrix} \cos 2\pi i/m \\ \sin 2\pi i/m \end{pmatrix}.$$

The hypothesis testing problem is specified by

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

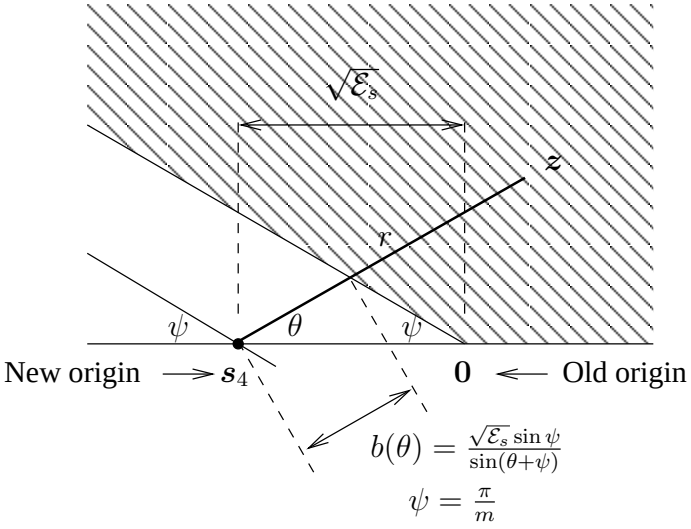
and the prior $P_H(i)$ *is assumed to be uniformly distributed.*

Since we have a uniform prior, the MAP and the ML decision rule are identical. Due to the circular symmetry of the additive noise, the ML decoder is a minimum-distance decoder. The decoding regions (up to ties) are shown in the picture below.



Now we proceed to compute the error probability. By symmetry, $P_e(i)$ *is independent of* i. *Hence* $P_e = P_e(i)$. *To determine* $P_e(i)$, *it is convenient to put the coordinate system at* $\boldsymbol{s}_i$ *as shown in the figure below.*

$$\begin{aligned} P_e(i) &= 2Pr\{\boldsymbol{Z} \in \text{ shaded area}\} \\ &= 2 \int_{\boldsymbol{z} \in \text{shaded area}} \frac{1}{2\pi\sigma^2} \exp\{-\frac{\| \boldsymbol{z} \|^2}{2\sigma^2}\} d\boldsymbol{z}. \end{aligned}$$



Passing to polar coordinates with

$$z_1 = r\cos\theta$$
$$z_2 = r\sin\theta$$

we obtain

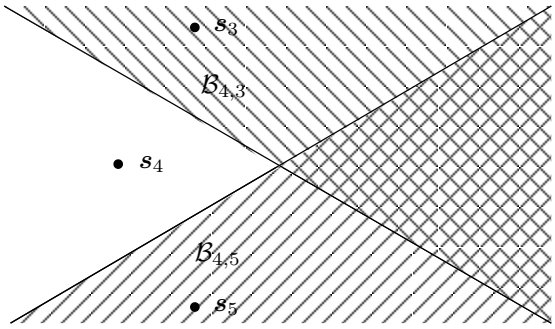
$$\begin{aligned} P_e(i) &= 2\int_0^{\pi-\psi}\int_{b(\theta)}^{\infty}\frac{1}{2\pi\sigma^2}\exp\left\{-\frac{r^2}{2\sigma^2}\right\} r\,dr\,d\theta \\ &= \int_0^{\pi-\psi}\frac{1}{\pi}\exp\left\{-\frac{b^2(\theta)}{2\sigma^2}\right\} d\theta \\ &= \frac{1}{\pi}\int_0^{\pi-\psi}\exp\left\{-\frac{\sin^2\psi}{\sin^2(\theta+\psi)}\frac{\mathcal{E}_s}{2\sigma^2}\right\} d\theta. \qquad \text{(Exact analysis of PSK).} \end{aligned}$$

This is as far as we can go with the exact analysis.

Now we use the union of events bound to determine an upperbound to the error probability. With reference to the figure below we have:

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

Notice that we have been using a version of the union of events bound adapted to the problem: we are getting a tighter bound by using the fact that $\mathcal{R}_i^c = \mathcal{B}_{i,i-1} \cup \mathcal{B}_{i,i+1}$ *rather than* $\mathcal{R}_i^c \subset \cup_{j\neq i}\mathcal{B}_{i,j}$ *.*



How good is the upperbound? Notice that

$$P_e = Pr\{\mathcal{B}_{i,i-1}|H=i\} + Pr\{\mathcal{B}_{i,i+1}|H=i\} - Pr\{\mathcal{B}_{i,i-1} \cap \mathcal{B}_{i,i+1}|H=i\}$$

and we obtained an upper bound by lower-bounding the last term with 0*. We now obtain a lower bound to* P_e *by upperbounding the same term:*

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

Hence,

$$\begin{aligned} P_e &= Pr\{\mathcal{B}_{i,i-1}|H=i\} + Pr\{\mathcal{B}_{i,i+1}|H=i\} - Pr\{\mathcal{B}_{4,3} \cap \mathcal{B}_{i,i+1}|H=i\} \\ &\geq 2Q\left(\sqrt{\frac{\mathcal{E}_s}{\sigma^2}}\sin\psi\right) - \frac{P_e}{m-1}. \end{aligned}$$

Solving for P_e *we obtain the desired lower bound*

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

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 determine even better lower bounds for which this ratio goes to* 1 *as* $\mathcal{E}_s/\sigma^2 \to \infty$*. One such bound is obtained by upperbounding* $Pr\{\mathcal{B}_{i,i-1} \cap \mathcal{B}_{i,i+1}|H=i\}$ *with the probability* $Q\left(\frac{\sqrt{\mathcal{E}_s}}{\sigma}\right)$ *that* Y_1 *is positive given* $H=i$*.* □

2.8 Union Bhattacharyya Bound

Let us summarize. From the union of events 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. What we have gained with the bound is that 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.[3]

Recall that

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

Hence

$$1_{\mathcal{B}_{i,j}}(\boldsymbol{y}) \leq \sqrt{\frac{P_H(j) f_{\boldsymbol{Y}|H}(\boldsymbol{y}|j)}{P_H(i) f_{\boldsymbol{Y}|H}(\boldsymbol{y}|i)}}.$$

With this we obtain the Bhattacharyya bound as follows:

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

What makes the last integral appealing is that we integrate over the entire $\mathbb{R}^n$. For *discrete memoryless channels* the bound further simplifies (see homework).

[3]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 left as an exercise.

As the name indicates, the *Union Bhattacharyya bound* is the union of events bound

$$Pr\{e\} \leq \sum_i \sum_{j:j\neq i} Pr\{\boldsymbol{Y} \in \mathcal{B}_{i,j} | H = i\} P_H(i),$$

with $Pr\{\boldsymbol{Y} \in \mathcal{B}_{i,j} | H = i\}$ upper bounded by the Bhattacharyya bound.

EXAMPLE 8. (Tightness of the Bhattacharyya Bound) *Consider the following scenario consisting of the following transmitter*

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

and where the channel is the binary erasure channel described in the figure:

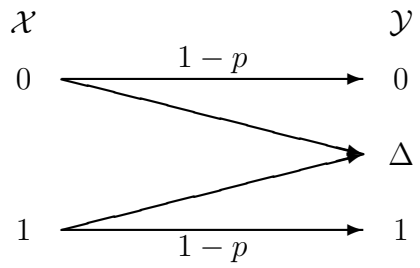


Figure 2.7: Binary erasure channel.

The Bhattacharyya bound is :

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

The same bound applies for $H = 1$*. Hence* $P_e \leq \frac{1}{2}p^n + \frac{1}{2}p^n = p^n$*.*

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*

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

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:*

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

For this channel the Bhattacharyya bound is tight!

2.9 Problems

PROBLEM 1. (Weather Frog.) *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 an 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:*

$$f_{Y|H}(y|0) = \begin{cases} A - \frac{A}{2}y, & 0 \le y \le 1 \\ 0, & \text{otherwise,} \end{cases} \tag{2.7}$$

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

$$f_{Y|H}(y|1) = \begin{cases} B + \frac{B}{3}y, & 0 \le y \le 1 \\ 0, & \text{otherwise.} \end{cases} \tag{2.8}$$

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

(i) *Determine* A *and* B.

(ii) *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.*

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

$$\hat{H}(y) = \begin{cases} 0, & \text{if } y \le \theta \\ 1, & \text{otherwise,} \end{cases} \tag{2.9}$$

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

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

(v) *Determine the probability of error for the decision rule that you have derived in* (iii).

(vi) *Among decision rules that compare the pressure* y *to a threshold like in Eqn. (2.9), is there a decision rule that results in a smaller probability of error than the rule derived in* (iii)*?* Hint: *Express the probability of error as in* (v) *as a function of* θ, *and minimize over all* θ.

PROBLEM 2. (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*

$$f_Z(z) = \frac{1}{2}e^{-|z|}. \tag{2.10}$$

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

(i) *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* $H = 1$.

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

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

PROBLEM 3. (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.*

$$p_{Y|H}(y|0) = \frac{\lambda_0^y}{y!}e^{-\lambda_0}. \tag{2.11}$$

Under hypothesis $H = 1$,

$$p_{Y|H}(y|1) = \frac{\lambda_1^y}{y!}e^{-\lambda_1}. \tag{2.12}$$

(i) *Make up a story around this problem. Clearly identify the meaning of* λ_0 *and* λ_1, *and of the observation* y.

(ii) *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.*

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

(iv) *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.*

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

PROBLEM 4. (iid versus first-order Markov model.) *Consider testing two equally likely hypotheses* $H = 0$ *and* $H = 1$. *The observable*

$$Y = (Y_1, \ldots, Y_k) \tag{2.13}$$

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:

$$Pr(Y_i = y_i | Y_{i-1} = y_{i-1}, \ldots, Y_1 = y_1) \;=\; \begin{cases} 3/4, & \text{if } y_i = y_{i-1} \\ 1/4, & \text{otherwise.} \end{cases} \tag{2.14}$$

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

(ii) *Give a simple sufficient statistic for this decision.*

(iii) *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*

$$y \;=\; 0101010111111\ldots111111010101\ldots. \tag{2.15}$$

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

Problem 5. (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*

$$f_{XY}(x,y) \;=\; \frac{1}{2\pi\sqrt{\det\Sigma}} \exp\left(-\frac{1}{2}\begin{pmatrix} x, & y \end{pmatrix} \Sigma^{-1} \begin{pmatrix} x \\ y \end{pmatrix}\right), \tag{2.16}$$

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

$$\Sigma \;=\; E\left[\begin{pmatrix} X \\ Y \end{pmatrix}(X,Y)\right] = \begin{pmatrix} \sigma_X^2 & \sigma_{XY} \\ \sigma_{XY} & \sigma_Y^2 \end{pmatrix}. \tag{2.17}$$

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

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

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

(iv) *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.*

(v) *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 6. (Correlation and independence.) *Let Z be a random variable with p.d.f.:*

$$f_Z(z) = \begin{cases} 1/2, & -1 \le z \le 1 \\ 0, & \text{otherwise.} \end{cases} \tag{2.18}$$

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

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

(ii) *Are X and Y independent?*

(iii) *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 7. (Transformation of Random Vectors.) *Let R and Φ be independent random variables. R is distributed uniformly over the unit interval, Φ is distributed uniformly over the interval $[0, 2\pi)$.*[4]

(i) *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!*

(ii) *Define the random variables*

$$X = R\cos\Phi \tag{2.19}$$
$$Y = R\sin\Phi. \tag{2.20}$$

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?

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

PROBLEM 8. (Theorem Of Irrelevance and Sufficient Statistics.) *Have you ever tried to drink from a fire hydrant? There are situations in which the observable Y contains too much data. You would like to have a many-to-one function T so that $T(Y)$ contains enough information to make a MAP decision but not too much to be impractical to work with. The Theorem of irrelevance gives a test to check if you have such a function.*

Consider two hypotheses with probabilities $p_H(0) = p_0$ and $p_H(1) = 1 - p_0$. The observable is $Y = (Y_1, \ldots, Y_k)$. Let $f_{Y|H}(y|0)$ and $f_{Y|H}(y|1)$ be given.

[4]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 $[0, 2\pi[$.

(i) (Theorem of irrelevance)*: Suppose it is possible to write*

$$f_{Y|H}(y|0) = g_0(T(y))h(y) \quad (2.21)$$
$$f_{Y|H}(y|1) = g_1(T(y))h(y), \quad (2.22)$$

where $T(\cdot) : \mathbb{R}^k \to \mathbb{R}^d$ *is a function from the observation space* $\mathbb{R}^k$ *to some space of choice* $\mathbb{R}^d$, $g_0(\cdot), g_1(\cdot) : \mathbb{R}^d \to \mathbb{R}^+$ *and* $h(\cdot) : \mathbb{R}^k \to \mathbb{R}^+$.

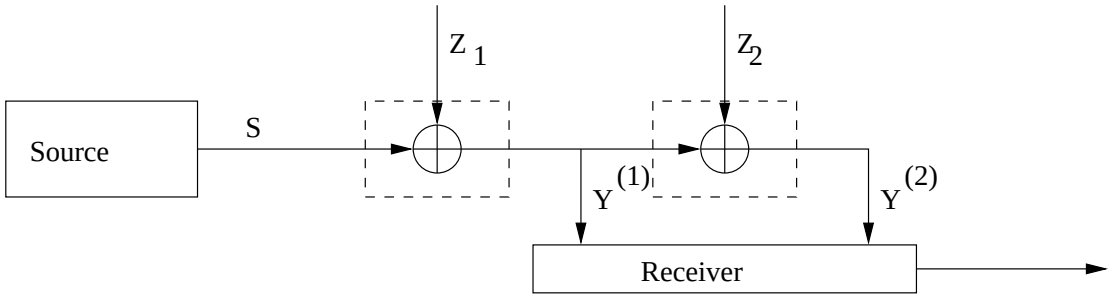
Prove that if you have $T(y)$ *you don't need* y *to make a MAP decision. For this reason* $T(y)$ *is called a sufficient statistic for the hypothesis testing problem.*

(ii) *Sometimes we can partition the observable* $Y \in \mathbb{R}^k$ *into two vectors* $Y' = (Y_1, \ldots, Y_r)$ *and* $Y'' = (Y_{r+1}, \ldots, Y_k)$. *Show that the irrelevance theorem implies the following statement: If* $f_{Y''|H,Y'}(y''|i, y')$ *does not depend on* i, *then* Y' *is a sufficient statistic, i.e.* Y'' *is irrelevant to the decision problem.*

(iii) *Use* (ii) *to answer the following communications problem (see the picture below): Under* $H = 0$, *the source emits* $S = 1$; *under* $H = 1$, *the source emits* $S = -1$. *The receiver has access to two noisy versions of the source output, namely*

$$Y^{(1)} = S + Z_1 \quad (2.23)$$
$$Y^{(2)} = S + Z_1 + Z_2, \quad (2.24)$$

where Z_1 *and* Z_2 *are zero-mean Gaussian random variables of variance* σ^2. *Is* $Y^{(2)}$ *relevant to the hypothesis testing problem? Prove your answer.*



Problem 9. *Consider a binary hypothesis testing problem specified by:*

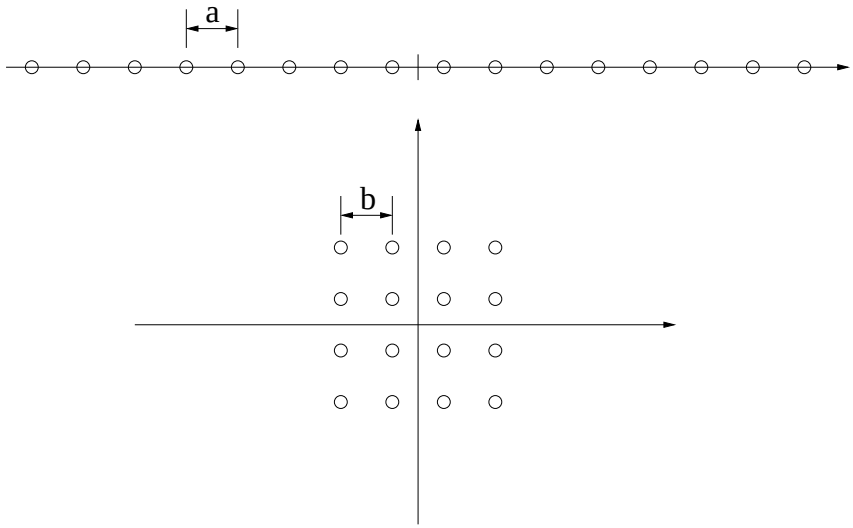
$$H = 0 \ : \ \begin{cases} Y_1 = Z_1 \\ Y_2 = Z_1 Z_2 \end{cases}$$
$$H = 1 \ : \ \begin{cases} Y_1 = -Z_1 \\ Y_2 = -Z_1 Z_2 \end{cases}$$

where Z_1, Z_2 *and* H *are independent random variables.*

(i) *Is* Y_1 *a sufficient statistic? Recall that* Y_1 *is a sufficient statistic if a MAP decoder that observes* (Y_1, Y_2) *makes the same decision (up to ties) as a MAP decoder that observes* Y_1 *alone.*

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

PROBLEM 10. (Comparison of 16-PAM and 16-QAM.) *The following two signal constellations are used to communicate across an additive white Gaussian noise channel. Le 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.*

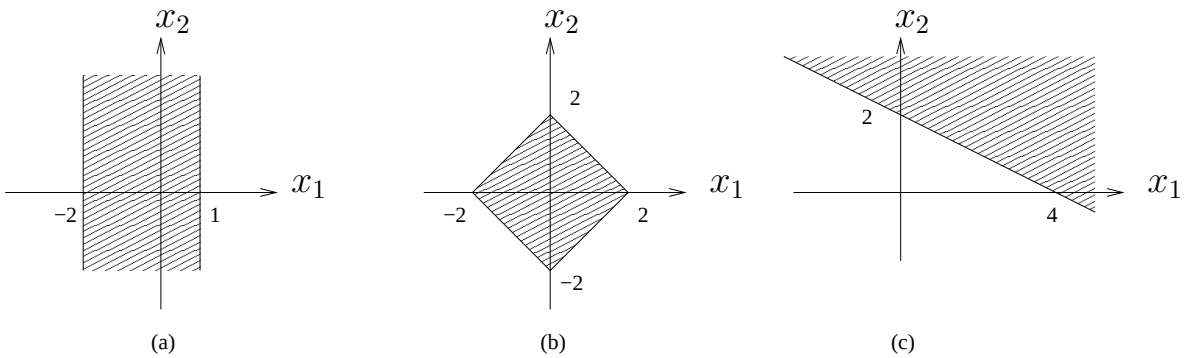
(i) *For each signal constellation, compute the average probability of error,* P_e *, as a function of the parameters* a *and* b*, respectively.*

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

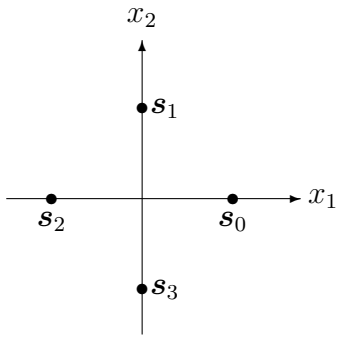
$$E_s = \sum_{i=1}^{16} p_H(i) \parallel \boldsymbol{s}_i \parallel^2 \tag{2.25}$$

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

PROBLEM 11. *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 12. *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 13. (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:

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

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*

$$\sigma_2^2 y_1 + \sigma_1^2 y_2 \;\underset{1}{\overset{0}{\gtrless}}\; 0. \tag{2.26}$$

(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 14. *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 15. *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 (2.8). 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* $P_{0\delta}$.

PROBLEM 16. (Repeat Codes and Bhattacharyya Bound.) *A repeat code is a code that transmits each source output* N *times across the channel. It is clear that the probability of error at the decoder decreases with increasing* N.

Consider two equally likely hypotheses (or, as we could also say, source output values). Under hypothesis $H = 0$, *the signal* $(X_1, \ldots, X_N) = (1, \ldots, 1)$ *is put onto the channel; under hypothesis* $H = 1$, *the signal is* $(X_1, \ldots, X_N) = (-1, \ldots, -1)$. *The transmission channel adds zero-mean independent Gaussian noise of variance* σ^2. *At the receiver, we observe*

$$(Y_1, \ldots, Y_N) \;=\; (X_1 + Z_1, \ldots, X_N + Z_N). \tag{2.27}$$

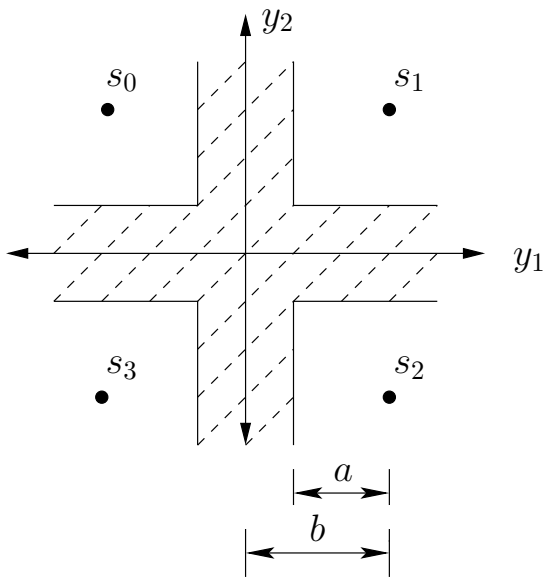


Figure 2.8: Modified QAM demodulator

Based on this observation, we can find the MAP estimator. In fact, it turns out that a sufficient statistic is the sum of the received values, $Y_1+Y_2+\ldots+Y_N$. *The corresponding probability of error was found to be*

$$Pr^{(1)}\{e\} \;=\; Q\left(\frac{\sqrt{N}}{\sigma}\right). \tag{2.28}$$

However, in this case, the receiver has to be able to perform addition of real numbers, and we also have to store them. This is not always possible. Therefore, suppose now that the decoder has access only *to the sign of* Y_i, $1 \le i \le N$. *That is, the observation is*

$$W = (W_1, \ldots, W_N) \;=\; (sgn(Y_1), \ldots, sgn(Y_N)) = (sgn(X_1 + Z_1), \ldots, sgn(X_N + Z_N)) \tag{2.29}$$

where $Z_i \sim \mathcal{N}(0, \sigma^2)$.

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

(ii) *Find the expression for the probability of error* $Pr^{(2)}\{e\}$. *You may assume that* N *is odd.*

(iii) *Your answer to* (ii) *contains a sum that cannot be solved in closed form. Therefore, find the Bhattacharyya bound on* $Pr^{(2)}\{e\}$.

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

PROBLEM 17. (Tighter Union Bhattacharyya Bound: Binary Case) *In this problem we derive a tighter version of the* Union Bhattacharyya Bound *for binary hypotheses. Let*

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

The MAP decision rule is

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

and the resulting probability of error is

$$Pr\{e\} = P_H(0) \int_{\mathcal{R}_1} f_{Y|H}(y \mid 0) dy + P_H(1) \int_{\mathcal{R}_0} f_{Y|H}(y \mid 1) dy. \tag{2.30}$$

(i) *Argue that*

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

(ii) *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,*

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

(iii) *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 18. (Tighter Union Bhattacharyya Bound: M-Ary Case)
In class we have derived the Union Bhattacharyya Bound. *Is this a tight bound or can we do better? To be specific, let us analyze the following M-ary MAP detector:*

$$\begin{aligned} \hat{H}(y) \quad &= \quad \textit{smallest } i \textit{ such that} && (2.31) \\ &\qquad P_H(i) f_{Y/H}(y/i) = \max_j \{P_H(j) f_{Y/H}(y/j)\} && (2.32) \end{aligned}$$

Let

$$\mathcal{B}_{ij} = \begin{cases} y : P_H(j) f_{Y|H}(y|j) \geq P_H(i) f_{Y|H}(y \mid i), & j < i \\ y : P_H(j) f_{Y|H}(y \mid j) > P_H(i) f_{Y|H}(y \mid i), & j > i \end{cases} \tag{2.33}$$

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

Given $H = i$, *the detector will make an error iff:*

$$y \in \bigcup_{j:j\neq i} \mathcal{B}_{ij} \tag{2.34}$$

We calculate the probability of error as:

$$Pr\{e\} = \sum_{i=0}^{M-1} Pr\{e \mid H = i\} P_H(i) \tag{2.35}$$

(ii) *Show that:*

$$\begin{aligned} Pr\{e\} &\leq \sum_{i=0}^{M-1} \sum_{j>i} \left[Pr\{\mathcal{B}_{ij} \mid H = i\} P_H(i) + Pr\{\mathcal{B}_{ji} \mid H = j\} P_H(j) \right] && (2.36) \\ &= \sum_{i=0}^{M-1} \sum_{j>i} \left[\int_{\mathcal{B}_{ij}} f_{Y|H}(y \mid i) P_H(i) dy + \int_{\mathcal{B}_{ij}^c} f_{Y|H}(y \mid j) P_H(j) dy \right] && (2.37) \\ &= \sum_{i=0}^{M-1} \sum_{j>i} \left[\int_y \min\left\{ f_{Y|H}(y \mid i) P_H(i), f_{Y|H}(y \mid j) P_H(j) \right\} dy \right] && (2.38) \end{aligned}$$

*(*Hint: *Apply the Union of Events Bound to equation (2.35) and then group the terms corresponding to* $\mathcal{B}_{ij}$ *and* $\mathcal{B}_{ji}$. *For proving the last part, go back to the definition of* $\mathcal{B}_{ij}$ *.)*

(iii) *Hence show that:*

$$Pr\{e\} \leq \sum_{i=0}^{M-1} \sum_{j>i} \left[\left(\frac{P_H(i) + P_H(j)}{2} \right) \int_y \sqrt{f_{Y|H}(y \mid i) f_{Y|H}(y \mid j)} dy \right] \tag{2.39}$$

*(*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:

$$\begin{aligned} H = 0 &: Y \sim \mathcal{N}(-a, \sigma^2) && (2.40) \\ H = 1 &: Y \sim \mathcal{N}(+a, \sigma^2) && (2.41) \end{aligned}$$

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

$$\begin{aligned} Pr\{e\} &= Pr\{e \mid H = 0\} && (2.42) \\ &\leq \frac{1}{2} \exp\left\{ -\frac{a^2}{2\sigma^2} \right\} && (2.43) \end{aligned}$$

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

$$Q(x) \leq \frac{1}{2} \exp\left\{ -\frac{x^2}{2} \right\}. \tag{2.44}$$

PROBLEM 19. *As an application of the bound derived in problem (10), 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.

(i) *Use the* Tight Bhattacharyya Bound *to derive a bound on* $\Pr\{e\}$.

(ii) *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 20. (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*[5] $P(y|x)$. *When we use this channel to transmit an n-tuple* $\boldsymbol{x} \in \mathcal{X}^n$, *the transition probability is*

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

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.*

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

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

following chain of inequalities.

$$\begin{aligned}
Pr\{e\} &\overset{(a)}{\leq} \frac{1}{2}\sum_{\boldsymbol{y}}\sqrt{P(\boldsymbol{y}|\boldsymbol{s}_0)P(\boldsymbol{y}|\boldsymbol{s}_1)} \\
&\overset{(b)}{\leq} \sum_{\boldsymbol{y}}\sqrt{\prod_{i=1}^{n}P(y_i|s_{0i})P(y_i|s_{1i})} \\
&\overset{(c)}{=} \sum_{y_1,\ldots,y_n}\prod_{i=1}^{n}\sqrt{P(y_i|s_{0i})P(y_i|s_{1i})} \\
&\overset{(d)}{=} \left[\sum_{y_1}\sqrt{P(y_1|s_{01})P(y_1|s_{11})}\right]\cdots\left[\sum_{y_n}\sqrt{P(y_n|s_{0n})P(y_n|s_{1n})}\right] \\
&\overset{(e)}{=} \prod_{i=1}^{n}\sum_{y}\sqrt{P(y|s_{0i})P(y|s_{1i})} \\
&\overset{(f)}{=} \prod_{a\in\mathcal{X},b\in\mathcal{X},a\neq b}\left(\sum_{y}\sqrt{P(y|s_{0i})P(y|s_{1i})}\right)^{n(a,b)}.
\end{aligned}$$

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

(ii) *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*

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

where

$$z=\sum_{y}\sqrt{P(y|a)P(y|b)}.$$

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

(iii) *What is* z *for:*

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

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

*(*Hint: *Use the result from Homework 4, Problem (iii)).*

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

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

Verify your result with that of homework 4 problem 1.

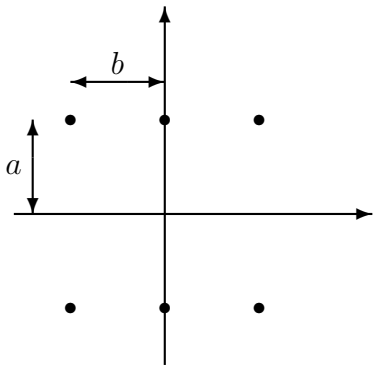
(c) *The Binary Erasure Channel (BEC) with the transition probabilities given by*

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

Verify your result with the one obtained in class.

(iv) Extra question for the curious ones: *Assume that the BSC has been obtained from the binary-input Gaussian channel via a one-bit quantizer applied at the channel output like in homework 4 problem (i). Plot the z of the original and the quantized channel as a function of the input power. By how much do we need to increase the input power of the quantized channel to match the z of the unquantized channel?*

Problem 21. (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 probabilities.

(i) *Draw the boundaries of the decision regions into the above figure.*

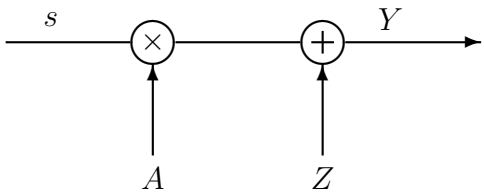
(ii) *Compute the average probability of error, $Pr\{e\}$, for this signal constellation.*

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

Problem 22. (Application of hypothesis testing to 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.



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

(ii) *Calculate the probability of error* $Pr\{e\}$*, based on the above decision rule.*

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 9. *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 [**?**, page 67].

THEOREM 10. *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 11. (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 $\boldsymbol{v}$. Then, consider the matrix

$$V^\dagger AV.$$

Now, 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

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

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,

$$V^\dagger A V = \begin{pmatrix} \lambda & *0 & WSW^\dagger \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & W \end{pmatrix} \begin{pmatrix} \lambda & * \\ 0 & S \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & W^\dagger \end{pmatrix} \tag{2.45}$$

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 12. *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 13. *An 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 11 (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 $R = U^\dagger H U$ we immediately see that R is Hermitian. Hence it is diagonal and 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 [**?**, Section 4.1].

Exercise 14. *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 15. *An Hermitian matrix* $H \in \mathbb{C}^{n\times n}$ *is said to be* positive definite *if*

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

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.*

EXAMPLE 16. *Show that a non-singular covariance matrix is always positive definite.*

THEOREM 17. (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*

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

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.

$$\boldsymbol{u}_i = \frac{1}{\sqrt{\lambda_i}} A\boldsymbol{v}_i, \quad i = 1, 2, \ldots, k. \tag{2.46}$$

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 $\{\boldsymbol{u}_i : i = 1, \ldots, k\}$ form an orthonormal set in $\mathbb{C}^m$. Complete this set to an orthonormal basis for $\mathbb{C}^m$ by choosing $\{\boldsymbol{u}_i : i = k+1, \ldots, m\}$ and let $U = (\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots, \boldsymbol{u}_m)$. Note that (2.46) implies

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

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

$$U \begin{pmatrix} \sqrt{\lambda_1} & & \mathbf{0} \\ & \ddots & \\ \mathbf{0} & & \sqrt{\lambda_n} \\ \cdots & \cdots & \cdots \\ & \mathbf{0}_{m-n} & \end{pmatrix} = AV \tag{2.47}$$

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

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

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 $DD^\dagger = \operatorname{diag}(\lambda_1, \lambda_2, \ldots, \lambda_k, 0, \ldots, 0)$. This shows that λ_i is also an eigenvalues of $AA^\dagger$. We have also shown that $\{\boldsymbol{v}_i : i = k+1, \ldots, n\}$ spans the null space of A and from (2.47) we see that $\{\boldsymbol{u}_i : i = 1, \ldots, k\}$ spans the range of A. □

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

LEMMA 18. *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 $\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_n$ into $\lambda_1\boldsymbol{e}_1, \lambda_2\boldsymbol{e}_2, \ldots, \lambda_n\boldsymbol{e}_n$. Hence, the unit cube is mapped into a rectangle of sides $\lambda_1, \lambda_2, \ldots, \lambda_n$. Its volume is $|\prod \lambda_i| = |\det D| = |\det A|$. □

Appendix 2.B Densities After Linear Transformations

The previous result leads to the following fundamental result.

THEOREM 19. *Let $\boldsymbol{X}$ be an n-rv of given pdf $f_{\boldsymbol{X}}(\boldsymbol{x})$, $A \in \mathbb{R}^{n\times n}$ a non-singular matrix, and $\boldsymbol{Y}$ be defined through the linear transformation $\boldsymbol{Y} = A\boldsymbol{X}$. The pdf of $\boldsymbol{Y}$ is given by*

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

Outline of the proof: The probability that $\boldsymbol{X}$ is inside an infinitesimally small cube $\delta\mathcal{A}$ (in n-dimensions) is $f_{\boldsymbol{X}}(\boldsymbol{x}^*)\delta\mathcal{A}$ (plus terms that become negligible as the volume of $\delta\mathcal{A}$ goes to zero), where $\boldsymbol{x}^*$ is any point inside $\delta\mathcal{A}$. Now $\boldsymbol{x}^*$ maps into $\boldsymbol{y}^* = A\boldsymbol{x}^*$ and $\delta\mathcal{A}$ into some $\delta\mathcal{B}$ of volume $\mathrm{Vol}(\delta\mathcal{B}) = \mathrm{Vol}(\delta\mathcal{A})|\det A|$. Since the probability that $\boldsymbol{Y}$ is inside $\delta\mathcal{B}$ is the same as the probability that $\boldsymbol{X}$ is inside $\delta\mathcal{A}$ we have (in the limit):

$$\mathrm{Vol}(\delta\mathcal{B}) f_{\boldsymbol{Y}}(\boldsymbol{y}^*) = \mathrm{Vol}(\delta\mathcal{A}) f_{\boldsymbol{X}}(\boldsymbol{x}^*).$$

Solving for $f_{\boldsymbol{Y}}(\boldsymbol{y}^*)$ yields the desired result. □

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 $W : \Omega \to \mathbb{R}$ 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 $W \sim \mathcal{N}(0, \sigma^2)$.

An *n-dimensional random vector* (n-rv) $\boldsymbol{X}$ is a mapping $\boldsymbol{X} : \Omega \to \mathbb{R}^n$. 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

$$f_{\boldsymbol{W}}(\boldsymbol{w}) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{w_i^2}{2\sigma^2}\right) \tag{2.48}$$

$$= \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{\boldsymbol{w}^T\boldsymbol{w}}{2\sigma^2}\right). \tag{2.49}$$

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

Definition 20. *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*

$$\boldsymbol{Y} = A\boldsymbol{W} \tag{2.50}$$

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

Note 21. *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 that if $\boldsymbol{Y} = A\boldsymbol{W}$ 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

$$f_{\boldsymbol{Y}}(\boldsymbol{y}) = \frac{1}{\sqrt{(2\pi)^n \det K_{\boldsymbol{Y}}}} \exp\left(-\frac{1}{2}\boldsymbol{y}^T K_{\boldsymbol{Y}}^{-1}\boldsymbol{y}\right) \tag{2.51}$$

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

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

FACT 22. *Let $\boldsymbol{Y}$ be a zero-mean random vector with arbitrary covariance matrix $K_{\boldsymbol{Y}}$ and pdf as in (2.51). Since a covariance matrix is Hermitian, we we can write (see Appendix 2.A)*

$$K_{\boldsymbol{Y}} = U\Lambda U^{\dagger} \tag{2.52}$$

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 (2.51) can always be written in the form $\boldsymbol{Y} = A\boldsymbol{W}$ where $\boldsymbol{W}$ has iid $\sim \mathcal{N}(0,1)$ components.

The contrary is not true in degenerated cases. We have already seen that (2.51) follows from (2.50) when A is a non-singular squared matrix. The derivation extends to any non-rectangular 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 (2.51) is not defined either. An example will show how to handle such degenerated cases.

It should be pointed out that many authors use (2.51) to define a Gaussian random vector. We favor (2.50) 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 23. (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 (2.51) 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 24. *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 (2.50) instead of (2.51), compare this answer with the tedious derivation consisting of integrating over $f_{\boldsymbol{Y}}$ to obtain f_{Y_i} (see Homework 1, Problem 1).

EXERCISE 25. *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 26. (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 27. *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 provide the easy proof using (2.51). 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 28. *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 Inner Product

For $\boldsymbol{a}, \boldsymbol{b} \in \mathbb{C}^n$, we define the inner product

$$\langle \boldsymbol{a}, \boldsymbol{b} \rangle = \sum_{i=1}^{n} a_i b_i^*$$

where b_i^* is the complex conjugate of b_i. The inner product is Hermitian symmetric, meaning that $\langle \boldsymbol{a}, \boldsymbol{b} \rangle = \langle \boldsymbol{b}, \boldsymbol{a} \rangle^*$. It is also Hermitian bilinear in the following sense. If α and β are in $\mathbb{C}$ and a, b, c in $\mathbb{C}^n$, then $\langle \alpha a + \beta b, c \rangle = \alpha \langle a, b \rangle + \beta \langle b, c \rangle$. Using the Hermitian symmetry it immediately follows that $\langle c, \alpha a + \beta b \rangle = \alpha^* \langle c, a \rangle + \beta^* \langle c, b \rangle$. The squared norm of $\boldsymbol{a} \in \mathbb{C}^n$ is

$$\| \boldsymbol{a} \|^2 = \langle \boldsymbol{a}, \boldsymbol{a} \rangle.$$

Using linearity, we immediately obtain

$$\begin{aligned} \| \boldsymbol{a} - \boldsymbol{b} \|^2 &= \langle \boldsymbol{a} - \boldsymbol{b}, \boldsymbol{a} - \boldsymbol{b} \rangle \\ &= \| \boldsymbol{a} \|^2 + \| \boldsymbol{b} \|^2 - 2Re\{\langle \boldsymbol{a}, \boldsymbol{b} \rangle\}. \end{aligned}$$

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

If $\boldsymbol{y}, \boldsymbol{u} \in \mathbb{C}^n$ and $\| \boldsymbol{u} \|^2 = 1$, then we may think of $| \langle \boldsymbol{y}, \boldsymbol{u} \rangle |$ as the length of the vector that we obtain when we project $\boldsymbol{y}$ onto $\boldsymbol{u}$. (Check this out for $u = (1, 0)^T$.)

An *hyperplane* in $\mathbb{C}^n$ is and $n-1$ dimensional subspace of the form

$$\{\boldsymbol{y} \in \mathbb{C}^n : \langle \boldsymbol{y}, \boldsymbol{u} \rangle = 0\}$$

where $\boldsymbol{u} \in \mathbb{C}^n$ is an arbitrary but fixed vector. The hyperplane defined by $\boldsymbol{u}$ is the set of vectors that are orthogonal to $\boldsymbol{u}$. An hyperplane always contains the origin.

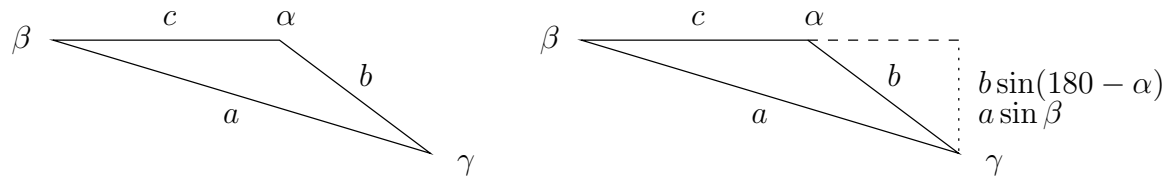
An *affine space* is a translate of an hyperplane. It has the general form

$$\{\boldsymbol{y} \in \mathbb{C}^n : \langle \boldsymbol{y}, \boldsymbol{u} \rangle = c\}$$

where $\boldsymbol{u} \in \mathbb{C}^n$ is an arbitrary but fixed vector and $c \in \mathbb{C}$ is an arbitrary but fixed scalar.

Appendix 2.E A Fact About Triangles

To determine an exact expression of the probability of error, in Example 7 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:

$$\frac{a}{\sin\alpha} = \frac{b}{\sin\beta} = \frac{c}{\sin\gamma}. \tag{2.53}$$

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. □