Syracuse University

SURFACE

Dissertations - ALL

SURFACE

July 2016

Optimal Inference for Distributed Detection

Earnest Akofor
Syracuse University

Follow this and additional works at: https://surface.syr.edu/etd

Part of the Engineering Commons

Recommended Citation

Abstract

In distributed detection, there does not exist an automatic way of generating optimal decision strategies for non-affine decision functions. Consequently, in a detection problem based on a non-affine decision function, establishing optimality of a given decision strategy, such as a generalized likelihood ratio test, is often difficult or even impossible.

In this thesis we develop a novel detection network optimization technique that can be used to determine necessary and sufficient conditions for optimality in distributed detection for which the underlying objective function is monotonic and convex in probabilistic decision strategies. Our developed approach leverages on basic concepts of optimization and statistical inference which are provided in appendices in sufficient detail. These basic concepts are combined to form the basis of an optimal inference technique for signal detection.

We prove a central theorem that characterizes optimality in a variety of distributed detection architectures. We discuss three applications of this result in distributed signal detection. These applications include interactive distributed detection, optimal tandem fusion architecture, and distributed detection by acyclic graph networks. In the conclusion we indicate several future research directions, which include possible generalizations of our optimization method and new research problems arising from each of the three applications considered.

Optimal Inference for Distributed Detection

by

Earnest Akofor

Ph.D. Physics, Syracuse University, 2010

B.Sc. Physics, University of Buea, 2001

Dissertation

Submitted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in Electrical & Computer Engineering.

Syracuse University

July 2016

Acknowledgments

Contents

III Concepts   99

Bibliography   145

List of Figures

Chapter 1

Introduction

1.1 Problem description and relevance

The problem

In complex statistical decision problems such as in distributed, sequential, or dynamic settings, the decisions from earlier stages serve as part of the data for decisions in the later stages. Therefore, even if the decision function for the decision at the first stage is an affine function of the initial decision probabilities, the decision functions at later stages are in general nonlinear in the probabilities of earlier decisions.

For distributed detection in particular, various types of decision functions appear in the literature, along with a variety of numerical algorithms for optimizing seemingly different classes of decision functions. However, there does not seem to exist any attempt to provide an efficient optimization procedure capable of stating explicit model-independent decision rules applicable to all *monotonic convex* decision functions (i.e., decision functions which are monotonic and convex in de-

cision probabilities) without resorting to suboptimal techniques (e.g., numerical programming and simulation) even for the simplest types of problems.

We intend to provide such a decision optimization framework and, hopefully, generalize the discussion to include monotonic subharmonic decision functions. We will show, in particular, that given any convex decision function to be optimized, it is always possible to decrease the space of optimization variables (no matter how large) to a set whose cardinality is no larger than the product of the cardinalities of the sets of decisions, hypotheses, and network components such as sensors. This reduction is completely independent of any network model of distributed detection.

The key observation that makes the reduction noted above possible is the fact that every extremum, i.e., maximum or minimum, of a differentiable convex function is either a boundary point of its domain or a point where its derivative equals zero.

Importance

It is not too difficult to observe that the optimization of two different decision functions F_1 and F_2 can yield two decision rules R_1 and R_2 that are identical or equivalent in the sense that they have decision regions of the same analytical form and there is a one-to-one correspondence between the set of threshold parameters T_1 that determines R_1 and the set of threshold parameters T_2 that determines R_2. Therefore it is clearly inefficient to directly compute R_2 when R_1 has already been computed.

We aim to show that there is only one type of decision rule or strategy (up to the equivalence stated above) that optimizes every monotonic convex decision function, even in the distributed setting. This should significantly reduce the ef-

fort involved in computing decision rules for decision functions in the monotonic convex class. Moreover, this analysis reveals that if sensor observations are conditionally independent and follow certain simple distributions (e.g., exponential family), then the decision problem becomes analytically tractable even for certain complex situations, such as that of distributed detection over acyclic graphs, as long as the decision functions are monotonic and convex.

In distributed detection literature, apparently different algorithms exist for computing decision rules for objective functions in the monotonic convex class. However, with our analysis, only one such algorithm may be necessary.

1.2 Related work and contributions

Almost every research paper on distributed detection first specifies a decision function, and then proceeds to obtain decision rules serving as necessary (and sometimes sufficient) conditions for optimality. To provide these rules, the authors tend to rely on the following.

(a) Susceptibility of the optimization problem to person-by-person optimization (PBPO) methods, especially when the underlying objective function is affine in decision probabilities. Each local sensor rule is derived under the assumption that optimal rules of all other sensors are given. For example, PBPO methods have been employed in [1, 2, 3, 4, 5, 6, 7, 8, 9].

(b) Suboptimal methods (e.g., generalized likelihood ratio tests) based on well known optimal solutions of simpler problems. At least one of the basic hypotheses is composite, and detection of *a given composite hypothesis* involves optimization over *its components*. Generalized likelihood ratio tests

have been used for example in [10, 11, 12, 13].

(c) Susceptibility of the optimization problem to dynamic programming techniques, especially in the context of sequential distributed detection. Optimization is performed repeatedly in several consecutive steps, where optimization at any given step utilizes suboptimal input from previous steps. For example, dynamic programming methods are found in [14, 15, 16, 9, 17].

Any success with the first two (and possibly the third) methods above is mostly a consequence of the monotonic and convex nature of the underlying decision function. The third method, i.e., dynamic programming, attempts to avoid the problem of a large space of optimization variables by sequentially incrementing the number of active optimization variables until a desired level of accuracy is reached.

All of these methods fail to recognize, and to properly utilize, the automatic reduction in the space of optimization variables associated with convex decision functions in general, as well as automatic optimality conditions which hold for monotonic convex decision functions in particular. Consequently, much greater effort than necessary is often required in establishing sufficiency (and hence optimality) of necessary conditions given in the form of local sensor decision strategies. This is a problem we intend to address in some detail.

The main contributions of this thesis are the following.

1. *Optimal hypothesis testing* (Chapter 3): We extend work on optimal detection initiated in [18, 19]. Specifically, we prove that every monotonic convex decision function has a unique optimum. We derive the general structure of optimal decision rules that represent the necessary and sufficient conditions for this optimum.

2. *Interactive distributed detection* (Chapter 4): Based on the optimality criterion obtained in Chapter 3, we present work done in [19] on interactive distributed detection, which is related work done in [18, 20]. We consider a decision fusion setup in which two sensors in tandem interact once in a memoryless fashion, by exchanging 1-bit decisions in a two-way communication process. It is shown that this interactive fusion can improve fixed sample performance of the Neyman-Pearson (NP) test but not large sample asymptotic performance of the test. This result is then extended to more realistic situations involving multiple rounds of memoryless interaction, multiple peripheral sensors, and the exchange of multibit decisions.

3. *Optimal fusion architecture* (Chapter 5): Again, based on the optimality criterion in Chapter 3, we present work done in [21] on the problem of determining the preferred two-sensor tandem fusion architecture in distributed detection of a deterministic, or Gaussian-distributed random, signal in Gaussian noises. Using an optimal version of a suboptimal decision strategy employed in [12, 13], as well as some techniques used therein, we determine that for low signal-to-noise ratio (SNR), the better sensor, i.e., the one with a larger SNR, should serve as the fusion center.

4. *Detection over acyclic graphs* (Chapter 6): We present some preliminary work on Bayesian distributed detection with sensor networks in the form of acyclic directed graphs. Specifically, we prove that if the communicated messages among sensors are such that each sensor passes the same message to every sensor receiving from it, then the optimal local decision rules for such a network are not more complicated than those of the simple tandem and parallel networks. Similar work was done in [3, 4] under assumption of binary

hypotheses, binary decisions, and at most a single connecting path between any two sensors. Our conclusions above do not require these assumptions.

We would like to remark that the results of Chapter 6 in particular may, or may not, be known. However, what is important for us in that chapter is not novelty but the relative ease with which the results therein can be obtained with the help of Proposition 3.1. In other words, Chapter 6 is mainly illustrating applicability of optimal hypothesis testing as described in Section 3.3.

1.3 Organization and prerequisite

The material in this thesis can be subdivided into three parts as follows.

For convenience, we have presented a review of essential preliminary material (Part III) as appendices. This part contains a brief review of *basic concepts* of optimal inference. These concepts include those of optimization of convex functions (Appendix A) and of statistical information inference (Appendix B). The latter includes a discussion of probability, statistics, point estimation, and hypothesis testing. Part III does not only make our work more self contained but also contains important results upon which the results of part I are based.

Part I considers statistical inference for *signal detection*, and contains the formulation of an optimal inference procedure for signal detection based on the main results of Part III. We begin with a brief nontechnical discussion of statistical decision theory in Chapter 2. This is then followed by a detailed discussion of optimal hypothesis testing in the context of signal detection in Chapter 3. Here, we first formulate the optimization problem for convex decision functions and prove a central theorem that can be applied in a variety of distributed detection archi-

tectures. Then, for illustration of application of the results, we derive centralized and distributed sensor network decision rules for Bayesian detection.

Part II deals with *some applications* of the optimal inference procedure of Part I in distributed detection. We summarize the main points of research work on distributed detection based on the methods we have developed in the previous chapters. In some cases, detailed proofs of theorems are not included since they can be found in the references. Each section is an overview of particular research papers. When possible, we indicate the papers that are being summarized, along with the references listed in those papers.

The applications considered in Part II include interactive distributed detection (Chapter 4), optimal two-sensor tandem fusion architecture (Chapter 5), and detection over acyclic graph networks (Chapter 6). In the presentation of each application, we often begin with theoretical results which are essentially corollaries of the main results of Chapter 3. This is then followed by performance analysis. In our case, performance analysis is done simply by plotting the optimal value of the decision function against different observational constraints (i.e., various possible types and qualities of data taken by the sensors), against different network patterns (i.e., the number and distribution of sensing nodes and links), or against different communication constraints (i.e., quality and capacity of the communication links).

We conclude the thesis in Chapter 7, where we summarize our main results and applications, identify possible future directions of research, and briefly comment on why our central results from Chapter 3 can be applied in sequential detection in particular.

Throughout the discussion, we take for granted that the reader is familiar with basic concepts of linear algebra such as spanning, independence, bases, dimension, and matrix representation of linear transformations. We also assume acquaintance

with basic notions of vector calculus in $\mathbb{R}^n$, which include the volume integral, (total) derivative, partial derivative, gradient, and directional derivative of a function from $\mathbb{R}^n$ to $\mathbb{R}$. Some knowledge of basic probability and statistics would be helpful as well.

1.4 Distributed detection

Since this thesis is mainly concerned with distributed detection, we will now briefly introduce distributed detection before proceeding. As we will see in Section 3.2, detection is a means of data compression in which the resulting output directly infers the state of a physical phenomenon (such as the presence or absence of a signal). Detection uses methods of optimization theory, statistical inference, and statistical decision theory. In the distributed detection setting, several detection devices called sensors perform detection separately to achieve a common goal. The main reason for studying distributed detection is contained in the following.

In practice, a distributed sensor network (i.e., a data processing system consisting of several sensors located far apart, in some precise sense) often has limited communication capabilities/resources. This makes distributed processing unavoidable. For example, two or more persons making a single decision together cannot function as a centralized system since they are only capable of exchanging summaries of their thoughts. Distributed detection provides a framework that can enhance data processing by such a system.
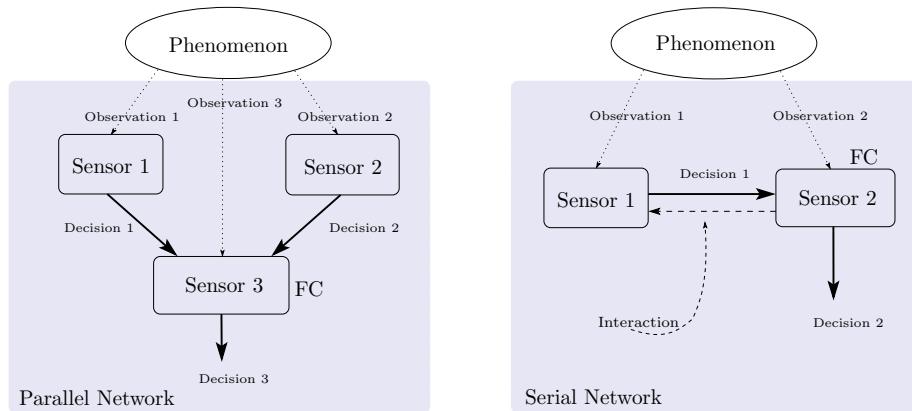
Figure 1.1: Basic decision fusion networks

A distributed sensor network is often specified in the form of a graph consisting of a set of nodes and a set of arrows. Each node represents a sensor making an observation. Each arrow represents a communication link between two sensors and points in the direction in which information must flow. As shown in Fig 1.1, which is based on diagrams found in [22], the simplest nontrivial distributed detection network contains about six basic elements - namely - at least two sensors, a phenomenon accessible to all sensors, sensor observations of the phenomenon as main input, communication links between sensors, sensor decision outputs, and sometimes a fusion center, i.e., a sensor whose output is considered to be "the final decision".

The following are some major benefits and advantages of distributed detection over centralized detection.

1. *Amendable performance*: Detection performance can be improved by increasing inter-sensor connectivity or through interactive processing and feedback.

2. *Robustness or fault tolerance*: If a few sensors fail, the distributed detection

system can still function.

3. *Reduced overload risk*: By distributing responsibility, the risk of over tasking (or overloading) one sensor is reduced.

4. *Reduced communication cost*: Less communication resources/capabilities are required by a distributed detection network, since sensors exchange only summaries of their observations.

The most significant disadvantage of distributed detection is delay in processing, i.e., a distributed system requires a longer processing time. Also, both the design and the performance analysis of a distributed detection system are more complex/challenging when compared with those of its centralized counterpart.

Part I

Detection

Chapter 2

Statistical Decision Theory

2.1 Introduction

This chapter is intended to provide motivation for, as well as improve our understanding of the practical significance of, the analysis to be presented in the next chapter. Since it is a special introduction to Chapter 3, we will be brief and concerned mainly with nontechnical aspects of the basic structure of a simple decision process. The question of how we can actually make certain types of decisions in practice is the subject of Chapter 3. The discussion here will illustrate the usefulness of statistical hypothesis testing in general.

A concise introduction to statistical decision theory is found in [23], and nontechnical introductions to the same are found in [24, 25]. Other useful references include [26, 27, 28].

The main motivation for, as well as the general definition of, decision theory are contained in the following. Making decision under uncertainty is a task that increases in difficulty as society grows and increases in complexity. Decision theory

provides a general structure or framework aimed at simplifying the decision making task.

Statistical decision theory is a method that uses observational data to enhance the decision making process when uncertainty is involved. It is a very interdisciplinary subject with varying perspectives, approaches, and applications. Nevertheless, the basic decision structure is the same in all cases.

The items we will discuss include "basic elements of a simple decision process", "classification of simple decision processes", "extensions to more complex decision processes", and "some applications of statistical decision theory".

2.2   Basic elements of a simple decision process

In a simple decision process, there are three basic elements, the *decision*, the (often unknown) *circumstance*, and the *consequence*.

For concreteness, we will work directly with an example. An example that easily illustrates the statistical aspect of a simple decision process is that of deciding whether an accused person is *guilty*, *partly guilty*, or *not guilty* of a crime.

When given this decision task, we have the following three basic components (See Figure 2.1).

1. *The decision*: This is one of a number of alternative actions to choose from. For our example, these actions or decision values are "The accused is guilty", "The accused is partly guilty", "The accused is not guilty".

2. *The circumstance*: This is the existing one among a number of possible natural conditions/states on which the objective or appropriate decision value depends. The circumstance is often uncertain, i.e., not completely known to

Figure 2.1: Table of consequences and associated costs

the decision maker. For our example, these conditions are "The accused committed the crime", "The accused did not commit the crime".

3. *The consequence*: This is one of a number of (anticipated) decision-circumstance outcomes, to each of which a cost of some sort is assigned. For our example, these outcomes are "Punish a guilty person", "Punish an innocent person", "Warn a guilty person", "Warn an innocent person", "Free a guilty person", "Free an innocent person".

We will refer to the above three elements as *internal elements* of the simple decision process. In order to classify simple decision types, a few more conceptual elements of the simple decision process are necessary.

2.3 Classification of simple decision processes

Our classification of simple decision processes is based on a number of *external elements* – namely – preference, prior information, and data.

1. *Preference*: The decision process is objective if it is based on a function of

the internal elements (i.e., the decision, circumstance, and consequence) that the decision seeks to optimize, otherwise, the decision process is subjective.

2. *Prior information*: The decision process is Bayesian if it uses prior knowledge (i.e., past experience with the circumstances), and it is frequentist or classical otherwise.

3. *Data*: The decision process is statistical if it uses data, which consists of observations from experiment on any systems that are directly or indirectly affected by the existing circumstance. Otherwise, the decision process is non-statistical.

In a statistical decision process, data can reduce uncertainty of the circumstance, i.e., it can partly reveal the circumstance. Consequently, data can improve decision quality. This is the main motivation for considering a statistical decision process.

We will mainly be concerned with objective statistical decision processes, which involve the following. In order to reduce uncertainty in the circumstance, we carry out a statistical investigation or experiment by collecting data from systems whose behaviors depend on the circumstance. The data is then used to improve decision making. The decision as a function of data is called a *decision strategy*. Our decision preference is represented by a function that depends on the decision strategy, the circumstances, and any costs associated with the consequences. Such a function is called a *decision function*, [28]. The *decision objective* is to select an *optimal decision strategy*, which is any decision strategy that minimizes the decision function.

Based on the above discussion, a convenient mathematical tool for handling statistical decision problems is hypothesis testing, which is introduced in Section

B.5. Based on the discussion above and that in Section B.5, a statistical decision process is, equivalently, a hypothesis test. We will see in Chapter 3 that the problem of detecting a signal embedded in corrupted measurements is a statistical decision problem, which can therefore be equivalently expressed as a hypothesis testing problem.

2.4 Extensions to decision processes in practice

In practice a typical decision process can contain several simple decisions, and may also involve several decision makers. For our purpose, these more complex decision processes can take one of the following labels.

1. *Sequential*: A decision process is sequential if it consists of several consecutive simple decisions.

2. *Distributed*: A decision process is distributed (or decentralized) if it involves several decision makers.

3. *Hybrid*: A decision process is hybrid if it is both sequential and distributed.

In Chapters 3, 4, 5, 6, we will encounter applications involving decision processes of the above types. This will be in the context of signal detection.

2.5 Some applications of statistical decision theory

In each of the cases below, the role of statistical decision theory becomes apparent when one attempts to answer the posed questions.

1. *Signal detection*: Is there a signal or no signal? How do we statistically extract it from noisy observations?

2. *Marketing*: Is there demand for a given product? How do we statistically determine it?

3. *Management*: Which task or who needs a resource? How can we be statistically sure?

4. *Forecasting/Prediction*: What is going to happen? How can we find out statistically?

Discussions on various applications of statistical decision theory can be found with the help of [23, 24, 25, 26, 27].

Chapter 3

Optimal Signal Detection

3.1   Introduction

In this chapter, which is a synthesis of preliminary results discussed in some detail in Appendices A, B, we study optimization of convex functions of *decision rules* or of *decision probability functions* for the purpose of distributed signal detection. Here, we should be mindful of the fact that the objective functions we shall deal with in real applications are not merely functions on $\mathbb{R}^n$ as discussed in Appendix A, but functions on function spaces, i.e., functions whose arguments are themselves functions. Such functions are also called functionals.

Routine problems considered in convex optimization mostly involve either the minimization of a convex function or the maximization of a concave function. However, problems that require maximization of a convex function, or minimization of a concave function, also arise. For example, in certain distributed detection problems the Kullback-Leibler distance is a performance metric that is convex in the variables Pr(decision|data), which are the pmf's of local sensor decisions condi-

tioned on data. The optimal decision rules are those that maximize this function.

The goal is to first provide differential relations that serve as necessary and sufficient conditions for the maximum of any detection performance metric that is a differentiable monotonic convex function of Pr(decision|data). Next, we then express optimal local sensor decision rules in terms of these differential relations.

Our approach is based on the following. Consider a real-valued differentiable convex function, defined on the n-dimensional real space, which we wish to maximize over a convex subset of the space (See Appendix A). By carefully studying the geometry of the graph of the function, we can derive *optimality conditions* (i.e., necessary and sufficient conditions for optimality) in the form of differential inequalities involving the derivative of the function at an optimal point (See Theorem A.6). Once this has been done, the problem can then be solved with the help of standard algorithms for solving differential inequalities.

In order to present local sensor decision rules in terms of the optimality conditions, we will first restate the detection problem as a general optimization problem in which the optimization variables in the objective function are Pr(decision|data), i.e., the pdf's of local decision rules conditioned on data. Optimal decision regions will then consist of those data points that satisfy the optimality conditions. The above procedure is presented in Section 3.3.

Notation

For convenience, we will adopt the following conventions from now on. Lower case letters such as $x, y, ...$ denote both random variables and their values. The symbol $\sum_x$ denotes summation when x is a discrete variable, or integration when x is a continuous variable. The expression $\delta_{xx'}$ denotes the Kronecker delta when

x and x' are discrete, or the Dirac delta when x and x' are continuous. Therefore, wherever the identity

$$\partial f(x)/\partial f(x') = \delta_{xx'}$$

appears, the variables x, x' can be tuples $(x_1, ..., x_N)$, $(x'_1, ..., x'_N)$ of several discrete or continuous variables, for which we naturally define $\delta_{xx'} = \delta_{(x_1,...,x_N)(x'_1,...,x'_N)}$ by

$$\delta_{(x_1,x_2,\cdots,x_N)(x'_1,x'_2,\cdots,x'_N)} = \delta_{x_1 x'_1} \delta_{x_2 x'_2} \cdots \delta_{x_N x'_N}.$$

In addition to the above conventions, we will denote sensors by upper case letters and their observations by the corresponding lower case letters. For example, x will denote the observation of sensor X, y the observation of sensor Y, and so on.

3.2 The detection problem

According to the preliminary discussion in Appendix B, the estimation problem refers to any situation in which one is faced with the task of determining the state of a system based on a given data sample, i.e., a set of experimental observations on the system.
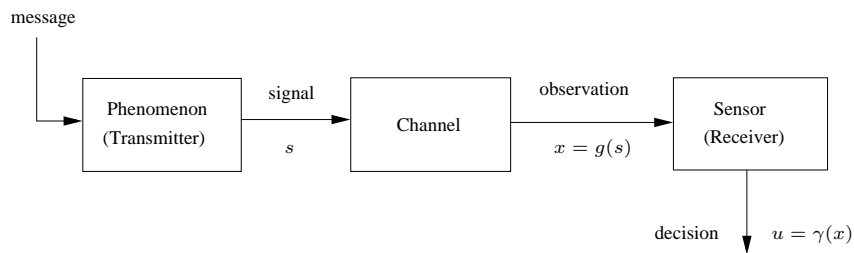
Figure 3.1: Communication system

In the detection problem in particular, we are faced with the task of extracting the value of a discrete signal

$$s \in \mathcal{S} = \{\mu_0, \mu_1, \cdots, \mu_{M-1}\},$$

(representing the unknown state of some system) from a given data sample $x^n = \{x_i \in \mathcal{X} : i = 1, 2, \cdots, n\}$. The system here refers to one or more components, such as the encoder, the channel, or the decoder, of a typical transmitting system. Each observation x in the sample x^n is typically viewed as a (known) *random function* of the signal. That is, the observation can be expressed as

$$x = g(s), \tag{3.1}$$

where the random function $g : s \in \mathcal{S} \mapsto x \in \mathcal{X}$ represents possible effects of known system properties on the signal. Such a function g is often called a *filter*, owing to its role in the signal extraction process.

We will not be dealing with the details of the encoding, channeling, and decoding rules whose composition determines g in general. Instead, for the most part in applications, we will consider the simplest case in which we assume that

$$x = s + b, \tag{3.2}$$

where b is white noise, i.e., a Gaussian-distributed random variable, and the signal s may be random as well. This will be sufficient for our main application interests.

Given the observation x as in (3.2), we wish to know how much noise b there is in x so that we can remove it and be left with s, i.e., we wish to know the value of $b = x - s$. However, we do not know the actual value of s. Since the value of b is known whenever that of s is given, we are therefore faced mainly with the problem

of statistically deciding the true value of s from the given observation x. This is precisely a hypothesis testing problem involving simple hypotheses

$$H_i : s = \mu_i, \quad i = 0, 1, \cdots, M - 1,$$

or equivalently,

$$H_i : x \sim p_i(x) = p(x|s = \mu_i), \quad i = 0, 1, \cdots, M - 1. \tag{3.3}$$

Note that the distribution $p_i(x)$ is known since we have already assumed that the distribution of b in (3.2), or of $g(\cdot)$ in (3.1), is known. With a slight abuse of notation, $p(x|s = \mu_i)$ will also be written as $p(x|H_i)$. Therefore the expressions $p_i(x)$, $p(x|H_i)$, $p(x|s = \mu_i)$ will all mean the same thing. Also, for notational convenience, we will not distinguish between a single observation $x \in x^n$ and the whole sample x^n, i.e., x will stand for a single sample $x \in x^n$ as well as for whole sample x^n sometimes.

In the detection problem as described above, it is sufficient to consider decision rules $\gamma(x)$ that take the same number of values as the number of hypotheses, so that $\gamma(x) = i \in \{0, 1, \cdots, M - 1\}$ stands for acceptance of the jth hypothesis. However, in the more general context of statistical decision theory, discussed in Chapter 2, the number of decision values can be different from the number of circumstance (or hypothesis) values. Moreover, in applications that require data quantization or compression in general, we may sometimes wish to first convert a relatively large set of observations into a smaller data set that represents the original set of observations as best as possible for the purpose at hand.

The above comment is especially true in distributed detection with fusion, where peripheral sensors generally compress their observations and pass them onto a fusing sensor that makes a decision based on the compressed data. In such a

setting, even when all sensors are using the same set of hypotheses, the output of a peripheral sensor may, or may not, be of the same alphabet type as the output of the fusing sensor.

We will therefore take into account the above situations in our analysis of optimal hypothesis testing. In particular, the set of decision values will have a cardinality which is different from that of the set of hypothesis values.

3.3 Optimal hypothesis testing

This section contains an extension of preliminary work found in [18, 19].

Recall that hypothesis testing was introduced in Section B.5, where optimal hypothesis testing was identified as a generalization of the notion of a sufficient statistic. We will now discuss optimal hypothesis testing and derive optimality conditions that are valid for all convex decision functions.

Consider a test of the M simple hypotheses in (3.3). Under H_i, we will denote the probability of a data set $A \subset \mathcal{X}$ by

$$P_i(A) = \sum_{x \in A} p_i(x).$$

The observation space $\mathcal{X}$ may be of arbitrary dimension. As in Definition B.20, a *decision rule* is a mapping defined as

$$\gamma : \ x \ \mapsto \ j \ \in \ \{0, 1, ..., N - 1\},$$

where γ is a deterministic function. We refer to the assignment $\gamma(x) = j$ as a *decision* based on the observation x. In problems where the decision output has the same alphabet as the underlying hypothesis, i.e., $N = M$, the decision $\gamma(x) = j$ may be interpreted as acceptance of the jth hypothesis H_j.

The desired decision rule γ so defined is deterministic in the sense that $p(\gamma(x) = j|x) = \delta_{j,\gamma(x)}$, where $\delta_{a,b} \triangleq 1$ if $a = b$ and 0 otherwise. Therefore, once x is given $\gamma(x)$ is precisely known. As the optimum decision rule is not necessarily deterministic, we consider the larger set containing all deterministic and nondeterministic decision rules. Let us write the generic decision rule as

$$\Gamma : \ x \ \mapsto \ j \ \in \ \{0, 1, ..., N - 1\}, \tag{3.4}$$

and let $u = \Gamma(x)$. Then $\Gamma = \gamma$ denotes a deterministic choice of the decision rule. Recall as in [29] that the set of Γ is the convex hull of the set of γ. Therefore

$$p(\Gamma(x) = j) = \sum_{g} p(g) \, p\big(\gamma_g(x) = j\big) \tag{3.5}$$

where g is a random variable with probability mass, or density, function $p(g)$ and is independent of x. The decision optimization process simply picks the appropriate $p(g)$, and hence the desired $p(\Gamma(x) = j)$.

As u is a random variable, making an optimal guess $u = j$ is equivalent to choosing $p(u = j|x)$ such that some *objective function*, which we denote by S, is optimized. Here S is a function of $p(u = j|x)$ for all $j = 0, 1, ..., N - 1$ and for all data points $x \in \mathcal{X}$. Note that we also refer to S as the *decision function* (See Definition B.20).

In general, $0 \leq p(u = j|x) \leq 1$, for each $j \in \{0, 1, ..., N - 1\}$. A *deterministic decision rule* is one for which $p(u = j|x)$ takes on only the boundary values 0 and 1. For such cases, we will see in Proposition 3.1 that the decision rule can be expressed as a partition of the data space into disjoint decision regions, i.e.,

$$p(u = j|x) \ \triangleq \ p(\gamma(x) = j|x) = I_{R_{u=j}}(x), \tag{3.6}$$

where $I_{R_{u=j}}(x) = \delta_{j,\gamma(x)}$ is the indicator function of the set $R_{u=j} = \{x : \gamma(x) = j\}$, which we call the *decision region* for the jth decision, and

$$R_{u=j} \cap R_{u=j'} = \emptyset, \quad \text{if} \quad j \neq j'.$$

When the number of decision values N equals the number of hypothesis values M, we may also refer to $R_{u=j}$ as the *acceptance region* for the jth hypothesis.

In the following proposition, we establish the general structure of the optimal decision rule for an important class of decision problems namely, those with monotonic convex objective functions. In other words, this proposition solves all monotonic convex versions of the optimization problem

$$\begin{aligned} \text{maximize} \quad & S\big(p(u|x)\big) \\ \text{subject to} \quad & p(u|x) \in Cs \end{aligned} \tag{3.7}$$

Proposition 3.1. *Let x be a random variable or vector, and suppose the objective function S and constraint set Cs in the problem (3.7) satisfy the following conditions.*

1. *Cs is a convex set.*

2. *S is nonconstant, differentiable, and <u>convex</u> on Cs in the multi-variable $\{p(u|x) : u = 0, 1, ..., N-1, \ x \in \mathcal{X}\}$.*

3. *S is <u>monotonic</u> with respect to a point of the boundary ∂Cs, i.e., there is a point $p \in \partial Cs$ such that S is monotonic along every line segment through p in the closure $\overline{Cs} = Cs \cup \partial Cs$. (In other words, S has the property given in Remark 1. immediately after Theorem A.6).*

4. *For each j, the set of data points*

$$C_{u=j} = \bigcup_{j' \neq j} \{x : \ \partial S/\partial p(u = j|x) = \partial S/\partial p(u = j'|x) \neq 0\} \tag{3.8}$$

has zero probability.

Then every optimal decision rule is deterministic, and is uniquely (i.e., necessarily and sufficiently) given by

$$p_{opt}(u = j|x) = I_{R_{u=j}}(x), \quad j = 0, 1, ..., N - 1, \qquad (3.9)$$

where the jth decision region $R_{u=j}$ is specified as

$$R_{u=j} = \bigcap_{j' \neq j} \left\{ x : \ \partial S / \partial p_{opt}(u = j|x) > \partial S / \partial p_{opt}(u = j'|x) \right\}. \qquad (3.10)$$

Note that in (3.10), the expression $\partial S / \partial p_{\mathrm{opt}}(u = j|x)$ denotes the derivative of S evaluated at the optimal point, i.e.,

$$\partial S / \partial p_{\mathrm{opt}}(u = j|x) = \partial S / \partial p(u = j|x)|_{p(u=j|x)=p_{\mathrm{opt}}(u=j|x)}.$$

Proof. Since Cs can be simplified to a polygonal (or simplicial) set by replacing S with a convex Lagrangian, we will assume without loss of generality that Cs is simplicial, i.e., we will choose the constraint set to be the *free* (or extended Cartesian) product $Cs = \prod_{x \in \mathcal{X}} \Delta_N(x)$, where $\Delta_N(x)$ is the N-dimensional probability simplex given by

$$\Delta_N(x) = \left\{ \vec{r}(x) = \big(p(u=0|x), ..., p(u=N-1|x)\big) : p(u=i|x) \geq 0, \sum_{i=0}^{N-1} p(u=i|x) = 1 \right\} \subset [0,1]^N.$$

Moreover, because S is convex in each variable, it is clear that we can proceed by optimizing S over one variable at a time while the others are held constant. Thus, it suffices to optimize S over $\Delta_N(x)$ for an arbitrary $x \in \mathcal{X}$. In other words, we want to maximize S with respect to $\vec{r}(x) = \big(p(u = 0|x), ..., p(u = N - 1|x)\big) \in [0, 1]^N$, with $\sum_{i=0}^{N-1} p(u = j|x) = 1$ and $p(u = j|x) \geq 0$ for $j = 0, \cdots, N - 1$.

If S is convex in $\vec{r}(x)$, then by Remark 4. after Theorem A.6, its maximum value occurs at one or more corner points of $\Delta_N(x)$, i.e.,

$$\vec{r}_{\mathrm{opt}}(x) = \vec{e}_j = (0, ..., 0, \underbrace{1}_{j\text{th spot}}, 0, ..., 0), \quad \text{for some } j \in \{0, ..., N-1\}.$$

For each $x \in X$ and each $j \in \{0, 1, ..., N-1\}$, condition (a) of Theorem A.6 implies

$$\vec{r}_{\mathrm{opt}}(x) = \vec{e}_j \iff p_{\mathrm{opt}}(u = j|x) = 1$$
$$\iff \text{ for all } \vec{r}(x) \in \Delta_N(x) \backslash \{\vec{e}_j\}, \quad (\vec{e}_j - \vec{r}(x)) \cdot \partial S / \partial \vec{r}_{\mathrm{opt}}(x) > 0, \tag{3.11}$$

and,

$$\vec{r}_{\mathrm{opt}}(x) \neq \vec{e}_j \iff p_{\mathrm{opt}}(u = j|x) = 0$$
$$\iff \text{ for some } \vec{r}(x) \in \Delta_N(x) \backslash \{\vec{e}_j\}, \quad (\vec{e}_j - \vec{r}(x)) \cdot \partial S / \partial \vec{r}_{\mathrm{opt}}(x) < 0, \tag{3.12}$$

where $\partial S / \partial \vec{r}_{\mathrm{opt}} = \partial S / \partial \vec{r}|_{\vec{r} = \vec{r}_{\mathrm{opt}}}$, $(\vec{e}_j - \vec{r}(x)) \cdot \partial S / \partial \vec{r}_{\mathrm{opt}}(x)$ is the dot-product of the vectors $\vec{e}_j - \vec{r}(x)$ and $\partial S / \partial \vec{r}_{\mathrm{opt}}(x)$, and $A \backslash B$ denotes the set difference, i.e., $A \backslash B = \{a : a \in A \text{ and } a \notin B\}$.

The region defined by x satisfying $p_{\mathrm{opt}}(u = j|x) = 1$ is

$$
\begin{aligned}
R_{u=j} &= \left\{ x : (\vec{e}_j - \vec{r}(x)) \cdot \partial S / \partial \vec{r}_{\mathrm{opt}}(x) > 0 \text{ for all } \vec{r}(x) \in \Delta_N(x) \backslash \{\vec{e}_j\} \right\} \\
&\overset{(a)}{=} \left\{ x : \partial S / \partial p_{\mathrm{opt}}(u = j|x) > \partial S / \partial p_{\mathrm{opt}}(u = j'|x) \text{ for all } j' \neq j \right\} \\
&= \bigcap_{j' \neq j} \left\{ x : \partial S / \partial p_{\mathrm{opt}}(u = j|x) > \partial S / \partial p_{\mathrm{opt}}(u = j'|x) \right\},
\end{aligned}
\tag{3.13}
$$

which by (3.8) is probability-wise complementary to the region defined by x satisfying $p_{\mathrm{opt}}(u = j|x) = 0$. Therefore, (3.11) covers both cases, and is equivalent to the deterministic rule (3.9). Note that step (a) in (3.13) is due to the following.

For fixed j, let $\Delta_{N-1}^{(j)}(x)$ be the convex hull of $\{\vec{e}_{j'} : \text{ for all } j' \neq j\}$, which is the face of $\Delta_N(x)$ opposite to $\vec{e}_j$. Then every point $\vec{s}(x) \in \Delta_{N-1}^{(j)}(x)$ can be expressed as

$$\vec{s}(x) = \sum_{j' \neq j} \alpha_{j'} \vec{e}_{j'}, \tag{3.14}$$
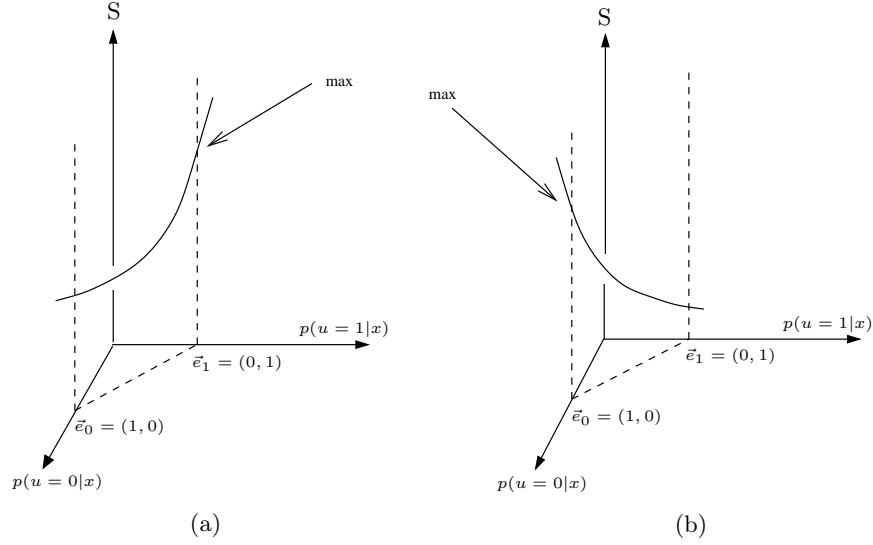
Figure 3.2: Visualization of the decision function S for $N = 2$: In case (a), $x \in R_{u=1}$, and in case (b), $x \in R_{u=0}$.

for some nonnegative numbers $\alpha_{j'} = \alpha_{j'}(x) \geq 0$ such that $\sum_{j' \neq j} \alpha_{j'} = 1$. Now, observe that for any point $\vec{r}(x) \in \Delta_N(x)$, we can write

$$\vec{e}_j - \vec{r}(x) = \vec{e}_j - \left(\lambda \vec{s}(x) + (1 - \lambda)\vec{e}_j\right) = \lambda\left(\vec{e}_j - \vec{s}(x)\right) \overset{(3.14)}{=} \sum_{j' \neq j} \lambda \alpha_{j'}\left(\vec{e}_j - \vec{e}_{j'}\right), \quad (3.15)$$

for some $\lambda = \lambda(x) \in [0, 1]$ and some $\vec{s}(x) \in \Delta_{N-1}^{(j)}(x)$.

$\square$

The following is a series of important remarks regarding applicability and possible extensions of Proposition 3.1.

Remarks.

1. The binary decision rule can be simplified further. In this case, the probability simplex $\Delta_2(x)$ is the single line with equation $p(u = 0|x) + p(u = 1|x) = 1$. Thus, by the chain rule of differentiation, the differential operator $\left(\vec{e}_j - \right.$

$\vec{r}(x)) \cdot \partial/\partial \vec{r}_{\mathrm{opt}}(x)$ along $\Delta_2(x)$ is equivalent to a derivative, which we denote by $\partial^B/\partial p_{\mathrm{opt}}(u = j|x)$, with the property

$$\partial^B p(u|x)/\partial p(u'|x') = (-1)^{u-u'}\delta_{xx'} \tag{3.16}$$

in addition to linearity and the Leibnitz rule. Hence, the binary decision regions take the compact form

$$R_{u=j} = \left\{ x : \ \partial^B S/\partial p_{\mathrm{opt}}(u = j|x) > 0 \right\}, \tag{3.17}$$

where the superscript B in ∂^B serves as a reminder to the reader of the property (3.16) which ensures that the derivative is restricted to the probability simplex $\Delta_2(x)$. The compact form of the binary decision rule as implemented by (3.16) and (3.17) will greatly simplify calculations later on.

2. Notice that (3.9) is an implicit equation in $p_{\mathrm{opt}}(u = j|x)$ since the region $R_{u=j}$ also depends on $p_{\mathrm{opt}}(u = j|x)$. Therefore we must proceed to substitute the equations

$$\left\{ p_{\mathrm{opt}}(u = j|x) = I_{R_{u=j}}(x) : j = 0, 1, ..., N - 1 \right\}$$

into the objective function S, and then compute the optimal threshold values that explicitly determine the decision regions. In the case of distributed networks of sensors where more than one set of local decision rules are involved, the resulting system of equations is often analytically intractable and one has to resort to numerical computation. This is especially the case when sensor observations are conditionally dependent, i.e., they remain dependent or correlated under at least one of the hypotheses.

In sufficiently simple problems for which the decision regions in (3.10) can be specified in terms of a number of threshold parameters, the optimization problem can be *explicitly* solved in a natural way simply by optimizing S over the

threshold variables. This, we will refer to as *the second stage* of the optimization problem, *the first stage* being the *implicit* solution as stated in the proposition.

3. Recall that for the optimal decision rule to be deterministic, as given in (3.9), the data sets (3.8) must be null with respect to the probability measure. Otherwise, the deterministic rule (3.9) is replaced by a randomized version

$$p_{\text{opt}}(u = j|x) = I_{R_{u=j}}(x) + \sum_k \rho_{jk} I_{C_k}(x), \tag{3.18}$$

where $\{C_k\}$ is a partition of the set

$$\bigcup_j C_{u=j}, \quad C_{u=j} = \bigcup_{j' \neq j} \left\{ x : \ \partial S/\partial p_{\text{opt}}(u = j|x) = \partial S/\partial p_{\text{opt}}(u = j'|x) \neq 0 \right\},$$

and $\rho_{jk} \in [0,1]$, $\sum_j \rho_{jk} = 1$, are arbitrary (i.e., free) coefficients but which must be consistent with every constraint of the optimization problem. It is worthwhile to remark that the deterministic rule (3.9) is more easily realized when x is continuous than when x is discrete. Thus, the randomized rule (3.18) is often required when x is a discrete random variable.

The fact that randomization depends on nullness of the sets $C_{u=j}$ generalizes a similar observation that was made in [30, 31, 41] under the Neymann-Pearson framework.

4. Since the Bayes risk is an affine (hence a convex) function of $p(u|x)$, Proposition 3.1 is a direct generalization of the familiar procedure whereby the unconditional Bayes risk

$$R(\gamma) = \sum_{i,j} C_{ij} p(\gamma(x) = i, H_j) = \sum_{i,x} p(\gamma(x) = i|x) R_i(x), \tag{3.19}$$

is minimized over γ simply by separately minimizing the associated conditional

Bayes risks $R_i(x) = \sum_j C_{ij} p(x|H_j) p(H_j)$ over i by means of the choice

$$p(\gamma(x) = i|x) = \begin{cases} 1, & R_i(x) < R_j(x) \text{ for all } j \neq i, \\ 0, & \text{otherwise}, \end{cases}$$

where we have assumed for simplicity that $P\big(R_i(x) = R_j(x)\big) = 0$ for all i and $j \neq i$.

5. If condition 3 of the proposition (i.e., the monotonicity condition) fails, then exactly the same conclusions hold, except for the sufficiency of the decision rule for optimality. That is, if conditions 1, 2, and 4 hold, then the optimal decision rule is deterministic, and satisfies (3.9) and (3.10) as necessary conditions. In that case, we must proceed, according to condition (b) of Theorem A.6, to select a decision rule that maximizes S among all possible decision rules that satisfy (3.9) and (3.10).

   In sufficiently simple problems for which monotonicity fails, the second stage of the optimization (as described in Remark 2. above) will in general suffer from the *local optimum problem* in the sense that S, as a function of the thresholds, possesses two or more local optima from which a global optimum must then be picked by some other method.

6. If both conditions 3 and 4 of the proposition fail, then the randomized rule (3.18) is necessary but not sufficient for optimality. Once again, we must proceed, according to condition (b) of Theorem A.6, to select a decision rule that maximizes S among all possible decision rules that satisfy (3.18).

7. Using standard methods of convex optimization, Proposition 3.1 can be further extended to include non-differentiable convex objective functions by replacing

the derivative $\partial S/\partial p(u = j|x)$ with a subdifferential. The proposition can also be refined to include the optimization of subharmonic objective functions.

We will now state a corollary of proposition for the minimization of a convex function. Consider the problem

$$
\begin{aligned}
\text{minimize} \quad & S\big(p(u|x)\big) \\
\text{subject to} \quad & p(u|x) \in Cs
\end{aligned}
\tag{3.20}
$$

Corollary 3.2. *Let x be a random variable or vector, and suppose the objective function S and constraint set Cs in the problem (3.20) satisfy the following conditions.*

1. *Cs is a convex set.*

2. *S is nonconstant, differentiable, and <u>convex</u> on Cs in the multi-variable $\{p(u|x) : u = 0, 1, ..., N - 1, \ x \in \mathcal{X}\}$.*

3. *For each j, the set of data points*

$$
C_{u=j} = \bigcup_{j' \neq j} \left\{x : \ \partial S/\partial p(u = j|x) = \partial S/\partial p(u = j'|x) \neq 0\right\}
\tag{3.21}
$$

   has zero probability.

Then every optimal decision rule is uniquely (i.e., necessarily and sufficiently) given by

$$
p_{opt}(u = j|x) = I_{R_{u=j}}(x) + p^0(u = j|x)I_{R^0}(x), \quad j = 0, 1, ..., N - 1,
\tag{3.22}
$$

where

$$
R_{u=j} = \bigcap_{j' \neq j} \left\{x : \ \partial S/\partial p_{opt}(u = j|x) < \partial S/\partial p_{opt}(u = j'|x)\right\},
\tag{3.23}
$$

and

$$R^0 = \{x \in \mathcal{X} : \ \partial S/\partial p(u = j|x) = 0 \ \ for \ all \ \ j = 0, 1, \cdots, N - 1\},$$

and for each j, x pair, $p^0(u = j|x)$ is the jth component of the solution of the system of equations

$$\{\partial S/\partial p(u = j|x) = 0, \ \ j = 0, 1, \cdots, N - 1\}.$$

Proof. With the help of Corollary A.7, the proof follows the same arguments as in the proof of Theorem 3.1. We simply need to (1) reverse the inequalities that determine the decision regions, (2) account for the possibility of the minimum occurring at the point where the derivative vanishes, and (3) recognize that monotonicity is not necessary for uniqueness of the solution.   □

Remarks.

1. If the derivative of S with respect to $p(u|x)$ is nonzero for all $x \in \mathcal{X}$, then the second term in (3.22) disappears.

2. If the derivative of S with respect to $p(u|x)$ equals 0 for all $x \in \mathcal{X}$, then the first term in (3.22) disappears.

3. If condition 3 of the corollary fails, then we must include a randomization term in the rule (3.22), as in (3.18), to obtain the randomized version

$$p_{\text{opt}}(u = j|x) = I_{R_{u=j}}(x) + p^0(u = j|x)I_{R^0}(x) + \sum_k \rho_{jk} I_{C_k}(x), \qquad (3.24)$$

where $\{C_k\}$ is a partition of the set $\bigcup_j C_{u=j}$, and $\rho_{jk} \in [0, 1]$, $\sum_j \rho_{jk} = 1$, are arbitrary coefficients but which must be consistent with every constraint of the optimization problem.

Proposition 3.1 will be used in Chapters 4, 5, 6 to determine optimal decision regions with various types of objective functions, including the probability of detection, KL distance, and Bayesian probability of error. First, however, we would like to illustrate how this result can be applied. We will use the Bayesian objective function in this illustration, while noting that the results for more general objective functions are similar. The main purpose is to demonstrate how Proposition 3.1 can be used in practical distributed detection problems. For concreteness, we begin with a discussion of centralized sensor rules in Section 3.4. The centralized detection process is then upgraded to a discussion of distributed sensor rules in Section 3.5.

3.4   Single sensor rules

From Section 3.2, suppose a *discrete* random signal $s \in \mathcal{S}$ is observed by an isolated sensor X as

$$x = s + b \ \in \ \mathcal{X}, \tag{3.25}$$

where $b \in \mathcal{B}$ is a continuous random parameter whose distribution is continuous and known, and s, b are *statistically independent* of one another. Let the signal alphabet be given by $\mathcal{S} = \{\mu_0, \mu_1, ..., \mu_{M-1}\}$. Then we have a set of M *hypotheses*

$$H_j : \ s = \mu_j, \ \text{i.e.,} \ x = \mu_j + b, \qquad j = 0, 1, ..., M - 1,$$

each of which represents a possible value of the signal s. For the special case where $M = 2$ and $\mathcal{S} = \{0, 1\}$, the hypothesis H_0 denoting "target absent" is called the *null* hypothesis, while the hypothesis H_1 denoting "target present" is called the *alternative* hypothesis. As usual, we denote the conditional pdf, $p(x|H_j) = p(x|s =$

μ_j) of x under H_j by $p_j(x)$. We emphasize here that $p_j(x)$ is known for each j since the distribution of b in (3.25) is known.

A *decision rule* of the sensor X is a mapping

$$\gamma : x \in \mathcal{X} \longmapsto u = \gamma(x) \in \mathcal{U} = \{0, 1, ..., N - 1\},$$

where we often denote the rule γ by its value u as a *variable*.

Our objective is to choose the rule γ such that some function $S = S(\gamma(x))$ of $u = \gamma(x)$ is optimized. Because u is a random variable, it is sufficient to treat S as a function of the conditional distributions $p(u|x)$, for all $x \in \mathcal{X}$. For our illustration, we consider the Bayesian objective

$$S = \sum_{i,j} C_{ij} p(u = i, H_j) = \sum_{i,x,j} C_{ij} p(u = i|x) p_j(x) \pi_j, \qquad (3.26)$$

where $p_j(x) = p(x|H_j)$, $\pi_j = p(H_j) = p(s = \mu_j)$ is the probability that H_j is true, and C_{ij} is a nonnegative number denoting the cost of the decision $u = i$ when H_j is true. Note that if $N = M$ (i.e., if the number decision values equals the number of hypothesis values), the decision $u = i$ may be viewed as acceptance of H_i, in which case C_{ij} is the cost of accepting H_i when H_j is true.

Since S is an affine function of the conditional probabilities $p(u|x)$ and the observation x is continuously distributed, by Proposition 3.1 the optimal decision rule is deterministic and, for each $i = 0, 1, ..., N - 1$, is given by

$$p_{\text{opt}}(u = i|x) = I_{R_{u=i}}(x) = \begin{cases} 1, & \text{if } x \in R_{u=i}, \\ 0, & \text{if } x \notin R_{u=i}, \end{cases} \qquad (3.27)$$

where the *decision region* $R_{u=i}$ is given by

$$R_{u=i} = \bigcap_{j \neq i} \left\{ x : \frac{\partial S}{\partial p_{\text{opt}}(u = i|x)} - \frac{\partial S}{\partial p_{\text{opt}}(u = j|x)} < 0 \right\}$$
$$= \bigcap_{j \neq i} \left\{ x : \sum_l C_{il} p_l(x) \pi_l - \sum_l C_{jl} p_l(x) \pi_l < 0 \right\}. \tag{3.28}$$

In terms of probabilities of the decision regions, the optimal value of S is

$$S_{\text{opt}} = \sum_{i,j} C_{ij} \, \pi_j \, p_j(R_{u=i}). \tag{3.29}$$

It is clear in this case that the decision regions, and hence S_{opt}, are determined by a fixed set of known thresholds. These thresholds are directly determined by the costs $\{C_{ij}\}$ and the prior probabilities $\{\pi_i\}$. As we will soon see, the situation is no longer so simple in distributed sensor settings.

3.5   Sensor network rules

Now suppose we have a distributed network of n sensors $\{X_k, \ k = 1, ..., n\}$, each sensor observing the same signal $s \in \{\mu_0, \mu_1, ..., \mu_{M-1}\}$. Here, the main difference with the preceding section is that each X_k must now make its decision u_k based not only on its own observation x_k, but as well on the set of decisions $\tilde{u}_k$ of all sensors forwarding their decisions to X_k, i.e., $u_k = \gamma_k(x_k, \tilde{u}_k)$ for some integer-valued function γ_k. Thus, X_k makes an observation

$$x_k = s + b_k \ \in \ \mathcal{X}_k, \quad b_k \in \mathcal{B}_k, \tag{3.30}$$

considers a set of hypotheses

$$H_j : \ s = \mu_j, \ \ \text{i.e.,} \ \ x_k = \mu_j + b_k, \quad j = 0, 1, ..., M - 1,$$

and applies a decision rule

$$\gamma_k : (x_k, \tilde{u}_k) \in \mathcal{X}_k \times \tilde{\mathcal{U}}_k \longmapsto u_k = \gamma_k(x_k, \tilde{u}_k) \in \mathcal{U}_k,$$

where $\mathcal{U}_k = \{0, 1, ..., N_k - 1\}$, and $\tilde{u}_k \in \tilde{\mathcal{U}}_k$ denotes the *set* of decision variables of all sensors transmitting their decisions to X_k. Here, for each k, we again assume the parameters s, b_k in (3.30) have the same properties as s, b in (3.25). Note that if we fix $N_k = N$ for all $k = 0, 1, ..., n$, then $\tilde{\mathcal{U}}_k = \{0, 1, ..., N - 1\}^{I_k}$, where I_k is the number of sensors transmitting their decisions to X_k.

Without loss of generality, we will let sensor X_1 serve as the fusion center for the network. Once again, our objective is to choose the *decision strategy* $\{\gamma_1, \gamma_2, ..., \gamma_n\}$ such that some function

$$S = S\big(\gamma_1(x_1, \tilde{u}_1), \gamma_1(x_2, \tilde{u}_2), ..., \gamma_k(x_n, \tilde{u}_n)\big)$$

of $u_1 = \gamma_1(x_1, \tilde{u}_1)$, $u_2 = \gamma_2(x_2, \tilde{u}_2)$, $\cdots$, $u_n = \gamma_n(x_n, \tilde{u}_n)$ is optimized. For our illustration, we consider S to be the Bayesian function at the fusion center X_1, i.e.,

$$
\begin{aligned}
S &= \sum_{i,j} C_{ij} p(u_1 = i, H_j) = \sum_{u_1,i} C_{u_1 i} p(u_1, H_i) \\
&= \sum_{u^n, x^n, i} C_{u_1 i} \prod_{k=1}^{n} p(u_k | x_k, \tilde{u}_k) \, p_i(x^n) \pi_i,
\end{aligned}
\tag{3.31}
$$

where $u^n = (u_1, ..., u_n)$, $x^n = (x_1, ..., x_n)$.

Since S is an affine function of the conditional probabilities $p(u_k | x_k, \tilde{u}_k)$ and the observations x_k are continuously distributed, by Proposition 3.1 the optimal decision rule for each sensor X_k is deterministic, and is given by

$$p_{\text{opt}}(u_k | x_k, \tilde{u}_k) = I_{R_{u_k | \tilde{u}_k}}(x_k) = \begin{cases} 1, & \text{if } x_k \in R_{u_k | \tilde{u}_k}, \\ 0, & \text{if } x_k \notin R_{u_k | \tilde{u}_k}, \end{cases} \tag{3.32}$$

where $u_k \in \{0, 1, ..., N_k - 1\}$, and the *decision region* $R_{u_k|\tilde{u}_k}$ is given by

$$R_{u_k|\tilde{u}_k} = \bigcap_{u \neq u_k} \left\{ x_k : \frac{\partial S}{\partial p_{\text{opt}}(u_k|x_k, \tilde{u}_k)} - \frac{\partial S}{\partial p_{\text{opt}}(u|x_k, \tilde{u}_k)} < 0 \right\}. \tag{3.33}$$

The derivative of S in (3.33) can be express as

$$\begin{aligned}
\frac{\partial S}{\partial p_{\text{opt}}(u_k|x_k, \tilde{u}_k)} &= \sum_{\{u^n, x^n\}_k, \ i} C_{u_1 i} \prod_{k' \neq k} p_{\text{opt}}(u_{k'}|x_{k'}, \tilde{u}_{k'}) \ p_i(x^n) \pi_i \\
&= \sum_{\{u^n, x^n\}_k, \ i} C_{u_1 i} \prod_{k' \neq k} I_{R_{u_{k'}|\tilde{u}_{k'}}}(x_{k'}) \ p_i(x^n) \pi_i,
\end{aligned} \tag{3.34}$$

where $\{u^n, x^n\}_k = \{u^n, x^n\} \backslash \{u_k, x_k, \tilde{u}_k\}$. With *conditionally independent observations*, we have $p_i(x^n) = \prod_{k=1}^n p_i(x_k)$. If we further assume there are *no closed processing paths* in the sensor network that can lead to *overlaps* among the decision regions, then (3.34) can be written as

$$\frac{\partial S}{\partial p_{\text{opt}}(u_k|x_k, \tilde{u}_k)} = \sum_{\{u^n, x^n\}_k, \ i} C_{u_1 i} \prod_{k' \neq k} p_i\left(R_{u_{k'}|\tilde{u}_{k'}}\right) \ p_i(x_k) \pi_i, \tag{3.35}$$

in which case, the optimal value of S in terms of probabilities of the decision regions is

$$S_{\text{opt}} = \sum_{u^n, i} C_{u_1 i} \ \pi_i \ \prod_{k=1}^n p_i\left(R_{u_k|\tilde{u}_k}\right). \tag{3.36}$$

We have thus proved the following result.

Theorem 3.3. *Suppose we are given a network of n sensors $X_1, X_2, ..., X_n$ with conditionally independent and continuously distributed observations $x_1, x_2, ..., x_n$. Suppose further that there are no cyclic communication paths in the sensor network. Then the decision rules for detection, based on the Bayes function (3.31), by the sensor network are given by (3.32), (3.33), and (3.35). Moreover, the optimal value of the Bayes function is given by (3.36).*

In further applications in Chapters 4, 5, 6, we will mostly consider *binary* hypothesis and *binary* decisions, i.e., we set $M = N_k = 2$. As already noted in previous remarks, when the problem is sufficiently simple, the decision regions (3.33) can be completely specified in terms of a number of threshold parameters that do not depend on the observations. In that case, we only need to optimize S as a function of the thresholds.

Part II

Applications

Chapter 4

Interactive Distributed Detection

4.1 Introduction

This chapter is based mainly on [18, 19], where detection is done in the Neyman-Pearson (NP) framework, but [20] also contains similar results under the Bayesian framework. We are going to study the impact of interactive fusion on detection performance in tandem fusion networks with conditionally independent observations. Both the fixed sample size NP test and the large sample NP test will be analyzed. There exist related work on fusion architecture in [12, 13, 32], on parallel and noninteractive feedback settings in [34, 35, 36], on various forms of asymptotic in [35, 36, 37, 38], and on parley in [39].

For the fixed sample test in Section 4.2, we will find that interactive distributed detection may strictly outperform the one-way tandem fusion structure. For the large sample test in Section 4.3, however, we will see that interactive fusion and one-way tandem fusion achieve the same asymptotic detection performance. (Note that this conclusion may no longer hold if certain communication constraints are

imposed, [40]). Also, we will find in Section 4.4 that these results remain valid in the following more general settings:

- The two sensors undergo multiple steps of memoryless interaction.

- The peripheral sensor is replaced by multiple peripheral sensors.

- Sensor outputs (before the final output) are multibit.

A simple tandem sensor network is a sequence of two or more sensors in which each sensor makes a single decision using its own observation and the output of its predecessor, and then passes its decision to the next sensor, i.e., its successor. The last sensor serves as a fusion center, and its decision is considered the final decision.

If the sensor outputs are single bit decisions, and sensor observations are independent conditioned on any given hypothesis, the optimal decision rule is determined by a likelihood ratio test [8, 41]. Note that this result assumes that every sensor makes only one decision.
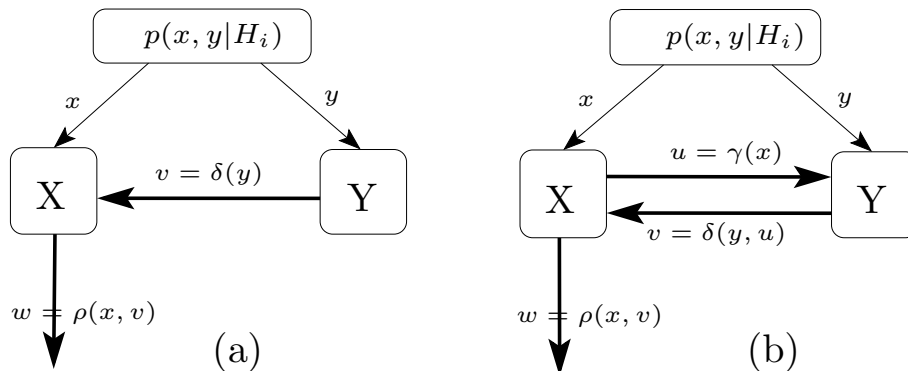
Figure 4.1: (a) One-way tandem fusion (YX process),   (b) Interactive fusion (XYX process).

For a two-sensor tandem network, we will replace the above static message passing with an interactive one: the fusion center (FC) sends an initial bit to the peripheral sensor (PS) based on its observation. The PS then makes a decision based on its own observation as well as the input from the FC and passes it back to the FC as partial input for the final decision. Fig. 4.1 illustrates the difference between the one-way tandem and interactive fusion networks.

In Fig 4.1(b), x is the observation of sensor X, y the observation of sensor Y, u the initial decision of X, v the decision of Y based on x, u, and w the final decision of X based on y, v. The random variables x and y are real-valued and assumed to be conditionally independent with respect to the hypothesis, i.e., $p_i(x, y) = p_i(x)p_i(y)$, meanwhile $u = \gamma(x)$, $v = \delta(y, u)$, $w = \rho(x, v)$ are binary, where γ, δ, ρ are integer-valued mappings. For simplicity, we refer to the fusion architecture in Fig. 4.1(a) as the YX process whereas to that in Fig. 4.1(b) as the XYX process.

Later in Section 4.4, we will also consider more general versions of the above situation, which involve multiple rounds of interaction, multiple peripheral sensors, and exchange of multi-bit decisions.

4.2 The fixed sample size Neyman-Pearson test

Our objective in the NP test is to maximize the probability of detection in such a way that the probability of false alarm does not exceed a given value α, i.e., we have the constrained optimization problem

$$
\begin{aligned}
\text{maximize} \quad & P_d = p_1(w = 1) \\
\text{subject to} \quad & P_f = p_0(w = 1) \leq \alpha
\end{aligned}
\tag{4.1}
$$

The Lagrangian for the problem is

$$L = p_1(w = 1) + \lambda \left(\alpha - p_0(w = 1)\right), \quad \lambda \geq 0. \tag{4.2}$$

For the YX process, L is a function of λ, $p(v|y)$, and $p(w|x, v)$, with

$$p_i(w) = \sum_{x,y,v} p(w|x, v)p(v|y) \, p_i(x)p_i(y), \tag{4.3}$$

while for the XYX process, L is a function of λ, $p(u|x)$, $p(v|y, u)$, and $p(w|x, v)$, with

$$p_i(w) = \sum_{x,v,y,u} p(w|x, v)p(v|y, u)p(u|x) \, p_i(x)p_i(y). \tag{4.4}$$

By applying Proposition 3.1, we obtain the following result, the proof of which is given in [19].

Theorem 4.1. *The optimal decision rules for the NP test with objective (4.2) are as follows. For the YX process, we have*

$$p_{opt}(v|y) = I_{R_v}(y), \quad p_{opt}(w|x, v) = I_{R_{w|v}}(x),$$

with the decision regions given by

$$R_{v=1} = \left\{ y : \frac{p_1(y)}{p_0(y)} > \lambda^{(2)} \right\}, \quad R_{w=1|v} = \left\{ x : \frac{p_1(x)}{p_0(x)} > \lambda_v^{(3)} \right\}, \tag{4.5}$$

where

$$\lambda^{(2)} = \lambda \, \frac{P_0(R_{w=1|v=1}) - P_0(R_{w=1|v=0})}{P_1(R_{w=1|v=1}) - P_1(R_{w=1|v=0})} \quad and \quad \lambda_v^{(3)} = \lambda \, \frac{P_0(R_v)}{P_1(R_v)}.$$

For the XYX process, we have

$$p_{opt}(u|x) = I_{R_u}(x), \quad p_{opt}(v|y, u) = I_{R_{v|u}}(y), \quad p_{opt}(w|x, v) = I_{R_{w|v}}(x),$$

with the decision regions given by

$$R_{u=1} = \left\{ x : \frac{p_1(x)}{p_0(x)} Q(x) > \lambda^{(1)} Q(x) \right\}, \qquad R_{v=1|u} = \left\{ y : \frac{p_1(y)}{p_0(y)} > \lambda_u^{(2)} \right\},$$

$$R_{w=1|v} = \left\{ x : \frac{p_1(x)}{p_0(x)} > \sum_u \lambda_{vu}^{(3)} \, I_{R_u}(x) \right\}, \tag{4.6}$$

where

$$\lambda^{(1)} = \lambda \, \frac{P_0(R_{v=1|u=1}) - P_0(R_{v=1|u=0})}{P_1(R_{v=1|u=1}) - P_1(R_{v=1|u=0})}, \qquad Q(x) = I_{R_{w=1|v=1}}(x) - I_{R_{w=1|v=0}}(x),$$

$$\lambda_u^{(2)} = \lambda \, \frac{P_0(R_{w=1|v=1} \cap R_u) - P_0(R_{w=1|v=0} \cap R_u)}{P_1(R_{w=1|v=1} \cap R_u) - P_1(R_{w=1|v=0} \cap R_u)}, \qquad and \qquad \lambda_{vu}^{(3)} = \lambda \, \frac{P_0(R_{v|u})}{P_1(R_{v|u})}.$$

Observe that in the interactive process, even though sensor observations are conditionally independent, the decision regions at the FC are not determined by simple likelihood ratio tests. A similar phenomenon was noted in [42, 43].

In terms of the obtained decision regions in Theorem 4.1, the Lagrangian in (4.2) can be written as

$$L_{YX} = \sum_v P_1(R_v) P_1(R_{w=1|v}) + \lambda \left[\alpha - \sum_v P_0(R_v) P_0(R_{w=1|v}) \right]$$

for the YX process, and as

$$L_{XYX} = \sum_{u,v} P_1(R_{v|u}) P_1(R_{w=1|v} \cap R_u) + \lambda \left[\alpha - \sum_{u,v} P_0(R_{v|u}) P_0(R_{w=1|v} \cap R_u) \right]$$

for the XYX process.

Example: Constant Signal in White Gaussian Noise

Consider the detection of a constant signal s in white Gaussian noise with observations

$$x = s + z_1, \qquad y = s + z_2, \qquad x, y \in \mathbb{R} = \mathcal{X} = \mathcal{Y}, \tag{4.7}$$

where $z_1 \sim N(0, \sigma_x^2)$, $z_2 \sim N(0, \sigma_y^2)$ and z_1 and z_2 are independent of each other, and the two hypotheses under test are

$$H_0: \ s = 0, \quad H_1: \ s = 1.$$

Fig. 4.2 shows the dependence of the probability of detection on σ_x when σ_y is fixed. The corresponding false alarm probability is $P_f = 0.2$. The figure shows that the XYX process has strictly larger probability of detection compared with the YX process.

The curve corresponding to centralized fusion in Fig. 4.2 is obtained by repeating the same optimization procedure using (4.1) and (4.2), but with the probability of the centralized decision $w = \rho(x, y)$ given by $p_i(w = 1) = \sum_{x,y} p(w = 1|x, y)p_i(x, y)$. Here, the decision rule $p_{\text{opt}}(w = 1|x, y) = I_{R_{w=1}}(x, y)$, the constant false alarm probability constraint $\alpha = p_0(w = 1)$, and the detection probability $P_d = p_1(w = 1)$ can be easily written as

$$R_{w=1} = \left\{ (x, y) : \frac{x}{\sigma_x^2} + \frac{y}{\sigma_y^2} > t = \ln \lambda + \frac{1}{2\sigma_x^2} + \frac{1}{2\sigma_y^2} \right\},$$

$$\alpha = \int_{-\infty}^{\infty} Q\left(\sigma_y t - \frac{\sigma_y}{\sigma_x} \frac{x}{\sigma_x} \right) \frac{e^{-\frac{x^2}{2\sigma_x^2}}}{\sqrt{2\pi\sigma_x^2}} dx,$$

$$P_d = \int_{-\infty}^{\infty} Q\left(\sigma_y t - \frac{\sigma_y}{\sigma_x} \frac{x}{\sigma_x} - \frac{1}{\sigma_y} \right) \frac{e^{-\frac{(x-1)^2}{2\sigma_x^2}}}{\sqrt{2\pi\sigma_x^2}} dx,$$

where the threshold t as a function of α is obtained by solving the constant false alarm probability constraint.

4.3 The asymptotic Neyman-Pearson test

Here, we will use scalar quantization since, as pointed out in [44], it is simpler and more efficient (in terms of a smaller processing delay) than vector quantization.
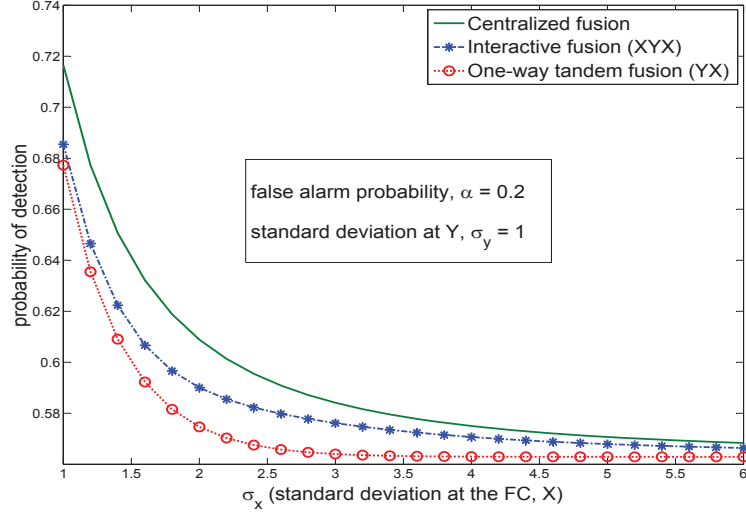
Figure 4.2: Performance of XYX and YX processes

Consider n observation samples $(x_1, y_1), ..., (x_n, y_n)$, and suppose processing is carried out on a sample-by-sample basis. For the XYX process, the two sensors go through, for each $k = 1, \cdots, n$, a decision process with $u_k = \gamma_k(x_k)$, $v_k = \delta_k(y_k, u_k)$. The final decision at node X utilizes the entire observation sequence x^n and the output sequence v^n from node Y, i.e., $w = \rho(x^n, v^n)$. We have, therefore,

$$p_i(w) = \sum_{x^n, v^n} p(w|x^n, v^n) p_i(x^n, v^n), \tag{4.8}$$

where $p(w|x^n, v^n)$ is determined by the final decision rule.

Meanwhile for the YX process, sensor Y sends a decision sequence $v_k = \delta_k(y_k)$, $k = 1, \cdots, n$, and X uses v^n and its own observation x^n to make the final decision $w = \rho(x^n, v^n)$. We again have the relation (4.8).

As shown in [19], and based on [33], the error exponent for the NP test is the

KL distance

$$D(p_0(x,v)\|p_1(x,v)) = \sum_{x,v} p_0(x,v) \log \frac{p_0(x,v)}{p_1(x,v)}. \tag{4.9}$$

This will be our objective function for the asymptotic performance of the NP test.

One-way tandem fusion (YX process)

In the one-way tandem fusion network, as shown in Fig. 4.1(a), Y sends a decision $v = \delta(y)$ to X. The optimal decision v is chosen so as to maximize the KL distance

$$K[x,v] = D\big(p_0(x,v)\|p_1(x,v)\big) \tag{4.10}$$

at sensor X.

Since $p_i(x,v) = p_i(x)p_i(v)$, we have

$$K[x,v] = D\big(p_0(x)\|p_1(x)\big) + \sum_v p_0(v) \, \log\big(p_0(v)/p_1(v)\big) \tag{4.11}$$

where $p_i(v) = \sum_y p(v|y)p_i(y)$.

By application of Proposition 3.1 in the optimization of (4.11), we get the following result, the proof of which is given in [19].

Theorem 4.2. *The optimal decision rule at Y is $p_{opt}(v|y) = I_{R_v}(y)$, where*

$$R_{v=1} = \left\{ y : \frac{p_1(y)}{p_0(y)} > \lambda \right\}, \tag{4.12}$$

$$\lambda = \left(\log \frac{\beta(1-\alpha)}{\alpha(1-\beta)} \right) \Big/ \left(\frac{\beta - \alpha}{\beta(1-\beta)} \right), \tag{4.13}$$

where $\alpha = P_0\left(p_1(y)/p_0(y) > \lambda\right)$ and $\beta = P_1\left(p_1(y)/p_0(y) > \lambda\right)$.

The maximum KL distance is given by

$$K_{\max}^{\text{YX}} = K[x] + \alpha^* \log \frac{\alpha^*}{\beta^*} + (1 - \alpha^*) \log \frac{1 - \alpha^*}{1 - \beta^*}, \tag{4.14}$$

where $K[x] = D(p_0(x)\|p_1(x))$ and α^* and β^* are the values of α and β that maximize the KL distance.

Interactive fusion (XYX process)

For the interactive fusion process in Fig. 4.1(b), X first sends a decision $u = \gamma(x)$ to Y. Then Y makes a decision $v = \delta(y, u)$ and sends it back to X. The optimal decisions u and v are chosen so as to maximize the KL distance $K[x, v] \triangleq K^{\text{XYX}}$ in the final step at X. The KL distance can be written as

$$K^{\text{XYX}} = D\big(p_0(x, v) \| p_1(x, v)\big) = D\big(p_0(x) \| p_1(x)\big) + \sum_x p_0(x) \sum_v p_0(v|x) \, \log \frac{p_0(v|x)}{p_1(v|x)}, \quad (4.15)$$

where $p_i(v|x) = \sum_u p(u|x) \sum_y p(v|y, u) p_i(y)$. Once more, by applying Proposition 3.1 in the optimization of (4.15), we get the following result, the proof of which is given in [19].

Theorem 4.3. *For the XYX process, the optimal decision rule at sensor X is* $p_{opt}(u|x) = I_{R_u}(x)$, *with decision region given by*

$$R_{u=1} = \left\{ x : \sum_u I_{R_u}(x) A_u B_u > 0 \right\}, \quad (4.16)$$

$$A_u = \frac{\beta_u^{(2)} - \alpha_u^{(2)}}{\beta_u^{(2)} (1 - \beta_u^{(2)})}, \qquad B_u = \frac{\beta_1^{(2)} - \beta_0^{(2)}}{\alpha_1^{(2)} - \alpha_0^{(2)}} - \lambda_u^{(2)}, \quad (4.17)$$

and the optimal decision rule at sensor Y is $p_{opt}(v|y, u) = I_{R_{v|u}}(y)$, *with decision regions given by*

$$R_{v=1|u} = \left\{ y : \frac{p_1(y)}{p_0(y)} > \lambda_u^{(2)} \right\}, \quad (4.18)$$

$$\lambda_u^{(2)} = \left(\log \frac{\beta_u^{(2)} (1 - \alpha_u^{(2)})}{\alpha_u^{(2)} (1 - \beta_u^{(2)})} \right) \bigg/ \left(\frac{\beta_u^{(2)} - \alpha_u^{(2)}}{\beta_u^{(2)} (1 - \beta_u^{(2)})} \right), \quad (4.19)$$

where $\alpha_u^{(2)} = P_0(R_{v=1|u})$, *and* $\beta_u^{(2)} = P_1(R_{v=1|u})$.

In terms of the decision regions, the KL distance (4.15) can be expressed as

$$K^{\text{XYX}} = K[x] + \sum_{u,v} P_0(R_u) \, P_0(R_{v|u}) \log \frac{P_0(R_{v|u})}{P_1(R_{v|u})}$$

$$= K[x] + \alpha^{(1)} f(\alpha_1^{(2)}, \beta_1^{(2)}) + (1 - \alpha^{(1)}) f(\alpha_0^{(2)}, \beta_0^{(2)}), \quad (4.20)$$

where $K[x] = D(p_0(x)\|p_1(x))$, $\alpha^{(1)}$ is a constant independent of the thresholds, and

$$f(\alpha, \beta) = \alpha \log \frac{\alpha}{\beta} + (1 - \alpha) \log \frac{1 - \alpha}{1 - \beta}. \tag{4.21}$$

Thus we have the following theorem.

Proposition 4.4. *The YX and XYX processes achieve identical $K[x, v]$. That is,*

$$K_{\max}^{YX} = K_{\max}^{XYX}. \tag{4.22}$$

Proof. The KL distances achieved by the two fusion systems, K^{YX} from (4.14) and K^{XYX} from (4.20), are respectively

$$\begin{aligned} K^{YX} &= K[x] + f(\alpha, \beta), & (4.23) \\ K^{XYX} &= K[x] + \alpha^{(1)} f(\alpha_1^{(2)}, \beta_1^{(2)}) + (1 - \alpha^{(1)}) f(\alpha_0^{(2)}, \beta_0^{(2)}), & (4.24) \end{aligned}$$

where the function $f(\alpha, \beta)$ is defined in (4.21).

Let α^* and β^* be the optimal values that maximize $f(\alpha, \beta)$ in K^{YX}. Comparing (4.12)-(4.13) and (4.18)-(4.19), it is apparent that the same α^* and β^* also maximize both $f(\alpha_1^{(2)}, \beta_1^{(2)})$ and $f(\alpha_0^{(2)}, \beta_0^{(2)})$ in K^{XYX}. This is so since for each value of u, the threshold dependence of the LRT using y is identical to that used in the YX process. Thus, the optimal decision on v at Y for the XYX process simply ignores the input from u, leading to identical LRTs for both values of u.

$\square$

Proposition 4.4 holds for any probability distribution. The results for the constant signal in WGN under hypotheses (4.7) are shown in Fig. 4.3, where the KL distances of YX and XYX processes coincide with each other. Also plotted are the

KL distances of XY and YXY that also coincide with each other. An interesting observation from the plot is that the two sets of curves, each corresponding to making final decision at different nodes, intercept each other at the point when $\sigma_x = \sigma_y = 1$. Thus for this example, it is always better to make the final decision at the sensor with better signal to noise ratio.
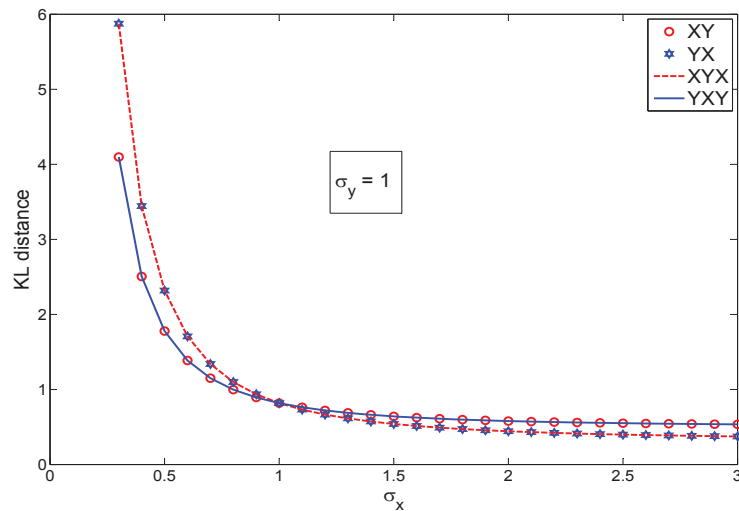
Figure 4.3: Comparison of KL distances of one-way tandem fusion and interactive fusion with different communication directions. For this plot, we fix $\sigma_y = 1$ throughout while varying σ_x.

4.4 Generalizations

In a two-sensor tandem network with a single round of interaction and 1-bit sensor output, we have shown that interactive fusion may strictly improve the detection performance of fixed sample size NP test, but not the asymptotic performance of the large sample NP test. We now consider more realistic settings in which this
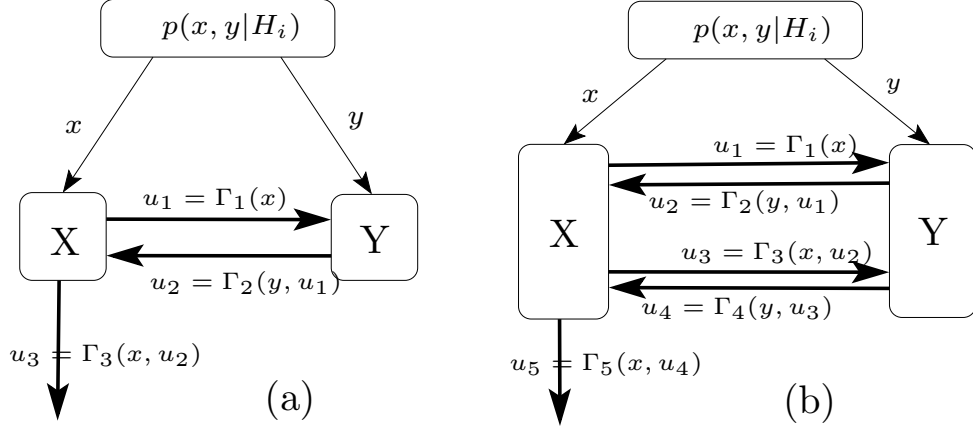
Figure 4.4: Sample MIF processes: (a) $N = 3$ MIF (XYX process), and (b) $N = 5$ MIF (XYXYX process).

result remains valid. These settings involve multiple round iterations, multiple sensors, and soft (i.e., multi-bit) sensor output.

Multiple-step memoryless interactive fusion (MIF)

In multiple round interactive fusion, sensors exchange 1-bit information repeatedly in $N > 3$ steps. Without any restriction on memory, it is not difficult to see that interactive fusion may strictly outperform the one-way tandem fusion asymptotically. Indeed, for N large enough, the performance of interactive fusion with memory will approach that of centralized detection.

However, there might be situations where the multiple round interactive fusion may proceed in a memoryless fashion, which we refer to as *memoryless interactive fusion* (MIF). In this case, a sensor's decision at each step depends on its own observation and the latest decision (but not on earlier decisions) of the other sensor. For this memoryless processing model, we show that with respect to asymptotic

detection performance, multiple-step interactive fusion still has no advantage over the one-way tandem fusion.

We begin with the expansion of the probability $p_i(u_N) = p(u_N|H_i)$ of the final decision u_N. Denote any sequence $s_1, ..., s_N$ by s^N. Let u^N be the sequence of decisions in the MIF process $XYXY \cdots YX$ involving two independent sensors X and Y, and let

$$z^N \equiv (z_1, ..., z_N) = (x, y, x, y, ..., y, x) \tag{4.25}$$

be the corresponding sequence of observations used at processing, as shown in Fig. 4.4 for $N = 3, 5$. Here we assume N is odd, thus, the decision process always starts with and ends at node X. Then because of the dependence structure $u_k = \Gamma_k(z_k, u_{k-1})$, $z_k = x$ when k is odd, and $z_k = y$ when k is even, we obtain

$$
\begin{aligned}
p_i(u_N) &= \sum_{z^N, u^{N-1}} p_i(z^N) \prod_{k=1}^{N} p(u_k|z_k, u_{k-1}) \\
&= \sum_{x,y,u^{N-1}} p_i(x, y) \prod_{r=1}^{(N-1)/2} [p(u_{2r-1}|x, u_{2r-2})p(u_{2r}|y, u_{2r-1})].
\end{aligned}
\tag{4.26}
$$

Now using conditional independence, $p_i(x, y) = p_i(x)p_i(y)$, we get

$$p_i(u_N|x) = \sum_{y, u^{N-1}} p_i(y) \prod_{r=1}^{(N-1)/2} [p(u_{2r-1}|x, u_{2r-2})p(u_{2r}|y, u_{2r-1})]. \tag{4.27}$$

With this expansion of $p_i(u_N)$, the following lemma, based on Proposition 3.1 and proved in [19], gives the peculiar nature of the resulting decision regions which are determined by an observation that is directly involved in the KL distance.

Lemma 4.5 (Degenerate MIF decision regions). *Let $u_N = \Gamma_N(x, u_{N-1})$ be the decision at the final step of a MIF process $XYXY \cdots YX$ with independent observations x and y. Let the objective function be given by the KL distance at the final*

step,

$$K[x, u_{N-1}] = \sum_{x, u_{N-1}} p_0(x, u_{N-1}) \log \frac{p_0(x, u_{N-1})}{p_1(x, u_{N-1})}. \qquad (4.28)$$

Then all decision regions based on x, i.e., in the optimal decision rule

$$p_{opt}(u_{2r-1}|x, u_{2r-2}) = I_{R_{u_{2r-1}|u_{2r-2}}}(x),$$

with decisions $u_{2r-1} = \Gamma_{2r-1}(x, u_{2r-2})$, $r = 1, 2, ..., \frac{N-1}{2}$, have the following general form.

$$R_{u_{2r-1}=1|u_{2r-2}} = \left\{ x : \sum_{\alpha} I_{D_\alpha}(x) A_{\alpha, r, u_{2r-2}} > 0 \right\}, \qquad (4.29)$$

where $\{D_\alpha\}$ is a partition of the data space $\mathcal{X}$, and the coefficients $A_{\alpha, r, u_{2r-2}}$ are independent of x.

Notice that (4.16) is a special case of (4.29). The following are some remarks about the degenerate decision regions (4.29):

- They depend on the distributions $p_0(x)$ and $p_1(x)$ only globally over $\mathcal{X}$, and not pointwise in x. Therefore given a single data point $x \in \mathcal{X}$, they cannot distinguish between H_0 and H_1.

- They are determined by piecewise constant functions with discrete probability distributions, and hence cannot define independent continuous threshold parameters; i.e., they contain no independent thresholds.

- They have piecewise constant probability; i.e., they have the same probability under both hypotheses.

- Their only role is to reparametrize the thresholds of the other regions. Consequently, they cannot improve optimality of the KL distance (as the next lemma shows).
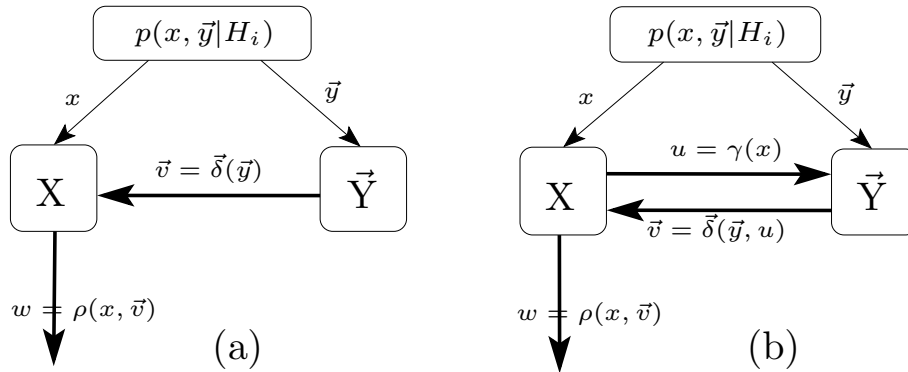
Figure 4.5: (a) = one-way tandem fusion ($\vec{Y}X$ process), and (b) = interactive fusion ($X\vec{Y}X$ process)

The following lemma, proved in [19], shows that the decision regions given by (4.29) are trivial in the sense that they do not participate in the decision process.

Lemma 4.6. *With respect to dependence on thresholds, the decision regions (4.29) of Lemma 4.5 have piece-wise constant probabilities. Moreover, such probabilities play no role at convergence and therefore do not contribute to the overall decision process.*

Therefore, careful analysis of the MIF process shows that whenever a sensor's data is explicitly summed over in the KL distance, the decision process becomes independent of that particular sensor's data. Since repetition of the decision process involving only one sensor's data cannot improve performance, it follows that MIF processing does not improve performance with respect to the KL distance.

Interactive fusion between the FC and multiple peripheral sensors

Consider our main setup in Fig. 4.1 and maintain sensor X as the FC while replacing sensor Y by K different sensors $\vec{Y} = \{Y_1, ..., Y_K\}$, with respective independent observations $\vec{y} = \{y_1, ..., y_K\}$. The resulting system is shown in Fig. 4.5. For the $\vec{Y}X$ process, we have decisions $(\vec{v}, w) \equiv (v_1, ..., v_K, w)$ based on observations $(x, \vec{y}) \equiv (x, y_1, ..., y_K)$, where $\vec{v} = \vec{\delta}(\vec{y}) = \big(\delta_1(y_1), ..., \delta_K(y_K)\big)$ and $w = \rho(x, \vec{v}) \equiv \rho(x, v_1, ..., v_K)$. Similarly, for the $X\vec{Y}X$ process, the decisions $(u, \vec{v}, w) \equiv (u, v_1, ..., v_K, w)$ are based on observations $(x, \vec{y}) \equiv (x, y_1, ..., y_K)$, with $u = \gamma(x)$, $\vec{v} = \vec{\delta}(\vec{y}, u) = \big(\delta_1(y_1, u), ..., \delta_K(y_K, u)\big)$ and $w = \rho(x, \vec{v}) \equiv \rho(x, v_1, ..., v_K)$.

In the fixed sample size NP test with Lagrangian (4.2), $p_i(w)$ is given by

$$p_i(w) = \sum_{x, \vec{y}, \vec{v}} p(w|x, \vec{v}) p(\vec{v}|\vec{y}) \, p_i(x) p_i(\vec{y}) = \sum_{x, \vec{y}, \vec{v}} p(w|x, \vec{v}) \prod_{k=1}^{K} p(v_k|y_k) \, p_i(x) \prod_{k=1}^{K} p_i(y_k) \quad (4.30)$$

for the $\vec{Y}X$ process, and

$$p_i(w) = \sum_{x, \vec{v}, \vec{y}, u} p(w|x, \vec{v}) p(\vec{v}|\vec{y}, u) p(u|x) \, p_i(x) p_i(\vec{y})$$
$$= \sum_{x, \vec{v}, \vec{y}, u} p(w|x, \vec{v}) \prod_{k=1}^{K} p(v_k|y_k, u) \, p(u|x) \, p_i(x) \prod_{k=1}^{K} p_i(y_k) \quad (4.31)$$

for the $X\vec{Y}X$ process. It suffices to find the $X\vec{Y}X$ decision regions only since those for $\vec{Y}X$ can be deduced from them by simply deleting the first decision u. Using Proposition 3.1 and following the same steps as in the proof of Theorem 4.1, we obtain the following. For the initial decision at X, $p_{\text{opt}}(u|x) = I_{R_u}(x)$, with

$$R_{u=1} = \left\{ x : \frac{\partial^B S}{\partial p(u=1|x)} > 0 \right\} = \left\{ x : \frac{p_1(x)}{p_0(x)} Q(x) > \lambda^{(1)} Q(x) \right\}, \quad (4.32)$$

where $\lambda^{(1)} = \lambda^{(0)} \frac{\prod_{k=1}^{K} P_0(R_{v_k=1|u=1}) - \prod_{k=1}^{K} P_0(R_{v_k=1|u=0})}{\prod_{k=1}^{K} P_1(R_{v_k=1|u=1}) - \prod_{k=1}^{K} P_1(R_{v_k=1|u=0})}$,
$Q(x) = -\sum_{\vec{v}} (-1)^{v_1+...+v_K} I_{R_{w=1|\vec{v}}}(x)$, while expressions for $\lambda^{(0)}$ and the objec-

tive function S are found in Appendix A of [19]. For the decision at each $Y_k \in \{Y_1, ..., Y_k\}$, we have $p_{\text{opt}}(v_k|y_k, u) = I_{R_{v_k|u}}(y_k)$, with

$$R_{v_k=1|u} = \left\{ y_k : \frac{\partial^B S}{\partial p(v_k = 1|y_k, u)} > 0 \right\} = \left\{ y_k : \frac{p_1(y_k)}{p_0(y_k)} > \lambda_{k,u}^{(2)} \right\}, \qquad (4.33)$$

where

$$\lambda_{k,u}^{(2)} = \lambda^{(0)} \frac{\sum_{v \backslash v_k} \left[P_0 \left(R_{w=1|v \backslash v_k, v_k=1} \cap R_u \right) - P_0 \left(R_{w=1|v \backslash v_k, v_k=0} \cap R_u \right) \right] \prod_{k' \neq k} P_0 \left(R_{v_{k'}|u} \right)}{\sum_{v \backslash v_k} \left[P_1 \left(R_{w=1|v \backslash v_k, v_k=1} \cap R_u \right) - P_1 \left(R_{w=1|v \backslash v_k, v_k=0} \cap R_u \right) \right] \prod_{k' \neq k} P_1 \left(R_{v_{k'}|u} \right)}.$$

For the final decision at X,

$$R_{w=1|\vec{v}} = \left\{ x : \frac{\partial^B S}{\partial p(w = 1|x, \vec{v})} > 0 \right\} = \left\{ x : \frac{p_1(x)}{p_0(x)} > \sum_u \lambda_{\vec{v}u}^{(3)} I_{R_u}(x) \right\}, \qquad (4.34)$$

where $\lambda_{\vec{v}u}^{(3)} = \lambda^{(0)} \frac{\prod_{k=1}^{K} P_0 \left(R_{v_k|u} \right)}{\prod_{k=1}^{K} P_1 \left(R_{v_k|u} \right)}$.

Similarly, for the asymptotic NP test, the KL distance $K^{X\vec{Y}X} \triangleq K[x, \vec{v}] = D\big(p_0(x, \vec{v})\|p_1(x, \vec{v})\big)$ can be expressed as

$$D\big(p_0(x)\|p_1(x)\big) + \sum_x p_0(x) \sum_{\vec{v}} p_0(\vec{v}|x) \, \log \frac{p_0(\vec{v}|x)}{p_1(\vec{v}|x)}, \qquad (4.35)$$

where $p_i(\vec{v}|x) = \sum_u p(u|x) \sum_y p(\vec{v}|y, u) p_i(y)$. By the same steps as in the proof of Theorem 4.3 for the X$\vec{Y}$X process, the decision rule at sensor X is $p_{\text{opt}}(u|x) = I_{R_u}(x)$, where

$$R_{u=1} = \left\{ x : \frac{\partial^B K[x, \vec{v}]}{\partial p(u = 1|x)} > 0 \right\} = \left\{ x : \sum_u I_{R_u}(x) C_u > 0 \right\},$$

$$C_u = \sum_{\vec{v}} \left(\sum_{u'} (-1)^{u'-1} P_0(R_{\vec{v}|u'}) \, \log \frac{P_0(R_{\vec{v}|u})}{P_1(R_{\vec{v}|u})} - \sum_{u'} (-1)^{u'-1} P_1(R_{\vec{v}|u'}) \frac{P_0(R_{\vec{v}|u})}{P_1(R_{\vec{v}|u})} \right), \qquad (4.36)$$

and the pair of decision regions $R_{v=1|u}$ in the rule $p_{\text{opt}}(v|y, u) = I_{R_{v|u}}(y)$ at sensor Y has the following K analogs corresponding to the sensors $\vec{Y}$; for each $k = 1, ..., K$, we have $p_{\text{opt}}(v_k|y_k, u) = I_{R_{v_k|u}}(y_k)$, with

$$R_{v_k=1|u} = \left\{ y_k : \frac{\partial^B K[x, \vec{v}]}{\partial p(v_k = 1|y_k, u)} > 0 \right\} = \left\{ y_k : \frac{p_1(y_k)}{p_0(y_k)} > \lambda_{ku}^{(2)} \right\},$$

$$\lambda_{ku}^{(2)} = \frac{\sum_{\vec{v}} (-1)^{v_k-1} \prod_{k' \neq k} P_0 \left(R_{v_{k'}|u} \right) \log \frac{P_0 \left(R_{\vec{v}|u} \right)}{P_1 \left(R_{\vec{v}|u} \right)}}{\sum_{\vec{v}} (-1)^{v_k-1} \prod_{k' \neq k} P_1 \left(R_{v_{k'}|u} \right) \frac{P_0 \left(R_{\vec{v}|u} \right)}{P_1 \left(R_{\vec{v}|u} \right)}}, \qquad (4.37)$$
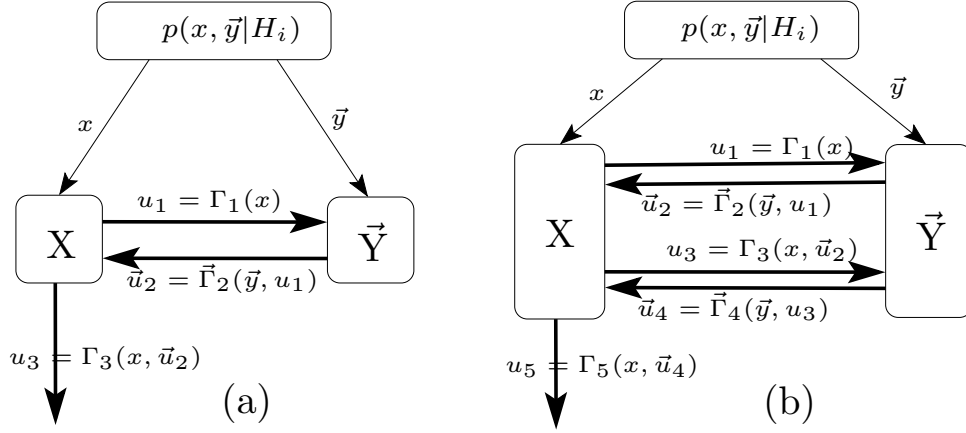
Figure 4.6: Sample MIF processes with K peripheral sensors: (a) $N = 3$ multi-sensor MIF (X$\vec{\text{Y}}$X process), and (b) $N = 5$ milti-sensor MIF (X$\vec{\text{Y}}$X$\vec{\text{Y}}$X process).

where $P_i\left(R_{\vec{v}|u}\right) = \prod_{k=1}^{K} P_i\left(R_{v_k|u}\right)$. The degenerate decision regions of Lemma 4.5 maintain their form as well. Since the decision rules have the same critical features (including threshold structure), our conclusions hold for this more general setup as well. This includes the multiple-step MIF of Section 4.4 with K peripheral sensors, shown in Fig. 4.6 for $N = 3, 5$ steps.

Interactive fusion with soft sensor outputs

We have shown in Section 4.3 that the YX and XYX processes have identical asymptotic detection performance when the output of each sensor is always binary. Now consider the other extreme case where the exchange of information is endowed with unlimited bandwidth. In that case, entire observations can be exchanged between sensors. Thus, both the YX and XYX processes again achieve exactly the same detection performance, namely, that of centralized detection. It remains to see if that is still the case for interactive fusion when soft information is exchanged,

i.e., sensor outputs are of a multiple but finite number of bits.

Consider the case where u and v can take respectively m and l bits. Equivalently, we have $u \in \{0, 1, ..., 2^m - 1\}$ and $v \in \{0, 1, ..., 2^l - 1\}$. Improvement of performance of the fixed-sample NP test by interactive fusion is immediate by induction, since the single bit decisions are a particular case of the multiple bit decisions. Therefore we consider the situation for the asymptotic test.

By Proposition 3.1, the optimal decision rule at X is $p_{\mathrm{opt}}(u|x) = I_{R_u}(x)$, with the decision regions given by

$$
R_{u=k} = \bigcap_{k' \neq k} \left\{ x : \frac{\partial K[x,v]}{\partial p(u=k|x)} - \frac{\partial K[x,v]}{\partial p(u=k'|x)} > 0 \right\},
$$
$$
k = 0, 1, ..., 2^m - 1,
$$

(4.38)

and the optimal rule at Y is $p_{\mathrm{opt}}(v|y, u) = I_{R_{v|u}}(y)$, with the decision regions given by

$$
R_{v=k|u} = \bigcap_{k' \neq k} \left\{ y : \frac{\partial K[x,v]}{\partial p(v=k|y,u)} - \frac{\partial K[x,v]}{\partial p(v=k'|y,u)} > 0 \right\},
$$
$$
k = 0, 1, ..., 2^l - 1,
$$

(4.39)

where the objective function $K[x,v]$ is defined by (4.15). It is straightforward, with the help of equation (3.9), to verify that all of the critical features of our analysis remain unchanged. In particular, by the same procedure as in the proofs of Theorem 4.3 and Lemma 4.5, the decision regions $R_{u=k}$ in (4.38) have the form

$$
R_{u=k} = \bigcap_{k' \neq k} \left\{ x : \sum_{k''=0}^{2^m-1} I_{R_{u=k''}}(x) a_{k''k'} > 0 \right\}, \quad k = 0, 1, ..., 2^m - 1, \qquad (4.40)
$$

which admits a piecewise constant probability. Hence multiple bit passing before the final decision does not alter our results.

Conclusion

We have applied the decision theory developed in Chapter 3 to study two-sensor tandem fusion networks with conditionally independent observations. Based on the optimum decision structure in each case, we have shown that while interactive fusion improves performance of the fixed sample size NP test, it does not affect asymptotic performance as characterized by the error exponent of type II error.

Several extensions of the above result were considered. The lack of improvement in asymptotic detection performance of one-step interactive fusion was shown to extend to multiple-step memoryless interactive fusion. Furthermore, the result was shown to be valid in a more general setting where the FC simultaneously interacts with $K \geq 1$ independent sensors. Finally, the results we also shown to be true in the case of multi-bit sensor output.

Chapter 5

Optimal Fusion Architecture

5.1 Introduction

This section is based on [21], the references in which include [20, 12, 13, 18, 19, 32].

Assume we wish to detect either a deterministic signal, or a Gaussian distributed random signal, in the presence of additive Gaussian noises using a two-sensor fusion system. That is, we have a distributed detection system with two sensors, one serving as a fusion center (FC) while the other as a peripheral sensor (PS) whose output is passed on to the FC for final decision making. As shown in Figure 5.1, a natural question about the preferred communication direction arises. What would be the optimal way of organizing the fusion system, i.e., which of the two sensors must serve as the FC for optimal detection performance?

We will show that for better detection performance at sufficiently low signal to noise ratio (SNR), the better sensor, i.e., the sensor with higher SNR, should serve as the fusion center. For the detection of a constant signal in additive Gaussian noises, it was also found in [18, 20] that under the Neyman-Pearson and Bayesian

criteria with conditionally independent observations, the sensor with lower SNR should serve as the FC. Note that these conclusions are false for general detection problems [32].
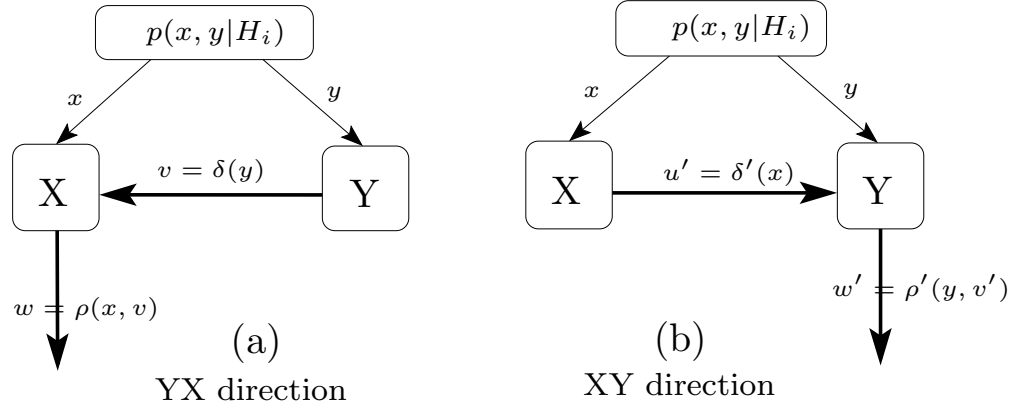
Figure 5.1: Depiction of the communication directions

In what follows, we first derive the optimal Bayesian test with dependent observations, which is valid for any continuous probability distribution that satisfies the assumptions of Proposition 3.1. Next, we consider the model with a random signal in additive Gaussian noises, and describe the topology of the resulting decision regions. Finally, we present computational results for low SNR.

5.2 The optimal Bayesian test

Consider a system of two sensors X and Y, with two possible directions of communication as shown in Fig 5.1. In the YX direction, Fig. 5.1 (a), sensor Y makes the first decision $v = \delta(y)$ with its observation y, and passes v to sensor X. Sensor X, now equipped with v and its own observation x, then makes the final decision $w = \rho(x, v)$. In the reverse direction XY, Fig. 5.1 (b), sensor X makes the first

decision $v' = \delta'(x)$, which is then used by sensor Y to make the final decision $w' = \rho'(y, v')$.

Without loss of generality, we will derive the test for the YX direction only. In addition, we consider only deterministic binary decisions since we are working under the assumptions of Proposition 3.1. The Bayesian cost function is given by

$$S[w] = \sum_{w,i} C_{wi} \, p(w, H_i) = \sum_{w,i} C_{wi} \pi_i \, p_i(w), \tag{5.1}$$

where $p_i(w) = \sum_{x,y,v} p(w|x,v) \, p(v|y) \, p_i(x,y)$, and $\pi_i = p(H_i)$. We note that since we consider only deterministic decisions, after optimization $p_i(w)$ can be written as

$$p_i(w) = \sum_{x,y,v} I_{R_{w|v}}(x) \, I_{R_v}(y) \, p_i(x,y), \tag{5.2}$$

where $R_{w=1|v}$ are the decision regions at X and $R_{v=1}$ is the decision region at Y. We have the following result.

Lemma 5.1. *With simple binary hypotheses $H_i : (x,y) \sim p_i(x,y), \; i = 0, 1$, the optimal decision rule $p_{opt}(v = 1|y) = I_{R_{v=1}}(y), \; p_{opt}(w = 1|x,v) = I_{R_{w=1|v}}(x)$, for the Bayesian decision problem with objective function (5.1) is given by the following decision regions. At Y,*

$$R_{v=1} = \left\{ y : \frac{p_1(y)}{p_0(y)} \frac{\sum_x [I_{R_{w=1|v=1}}(x) - I_{R_{w=1|v=0}}(x)] p_1(x|y)}{P_0(R_{w=1|v=1}) - P_0(R_{w=1|v=0})} > \lambda_0 \right\}, \tag{5.3}$$

and at X,

$$R_{w=1|v} = \left\{ x : \frac{p_1(x)}{p_0(x)} \frac{\sum_y I_{R_v}(y) p_1(y|x)}{P_0(R_v)} > \lambda_0 \right\}, \tag{5.4}$$

where $\lambda_0 = \frac{C_{10} - C_{00}}{C_{01} - C_{11}} \frac{\pi_0}{\pi_1}$, and the summation $\sum_x, \sum_y$ over continuous random variables is by means of integration. For conditionally independent observations, the

decision regions are given by simple likelihood ratio tests as follows:

$$R_{v=1} = \left\{ y : \frac{p_1(y)}{p_0(y)} > \lambda^{(1)} \right\}, \tag{5.5}$$

$$R_{w=1|v} = \left\{ x : \frac{p_1(x)}{p_0(x)} > \lambda_v^{(2)} \right\}, \tag{5.6}$$

where $\lambda^{(1)} = \lambda_0 \frac{P_0(R_{w=1|v=1}) - P_0(R_{w=1|v=0})}{P_1(R_{w=1|v=1}) - P_1(R_{w=1|v=0})}$ *and* $\lambda_v^{(2)} = \lambda_0 \frac{P_0(R_v)}{P_1(R_v)}$.

Proof. By Proposition 3.1, the local decision regions are given by

$$R_{v=1} = \left\{ y : \frac{\partial S[w]}{\partial p(v=1|y)} < 0 \right\},$$

$$R_{w=1|v} = \left\{ x : \frac{\partial S[w]}{\partial p(w=1|x,v)} < 0 \right\}.$$

Due to the constraint $p(v=1|y) + p(v=0|y) = 1$, we must use the differentiation rules

$$\frac{\partial p(v'|y')}{\partial p(v|y)} = (-1)^{v'-v} \delta_{y'y}, \qquad \frac{\partial p(w'|x',v')}{\partial p(w|x,v)} = (-1)^{w'-w} \delta_{x'x} \delta_{v'v}. \tag{5.7}$$

Therefore, at Y

$$\frac{\partial p_i(w)}{\partial p(v=1|y)} = \sum_{x,v} (-1)^{v-1} p(w|x,v) \ p_i(x,y)$$

$$= \sum_{x,v} (-1)^{v-1} I_{R_{w|v}}(x) \ p_i(x,y),$$

$$\Rightarrow \tag{5.8}$$

$$\frac{\partial S[w]}{\partial p(v=1|y)} = \sum_{w,i} C_{wi} \pi_i \frac{\partial p_i(w)}{\partial p(v=1|y)}$$

$$= \sum_{w,i} C_{wi} \pi_i \sum_{x,v} (-1)^{v-1} I_{R_{w|v}}(x) \ p_i(x,y).$$

Similarly, at X,

$$\frac{\partial p_i(w)}{\partial p(w = 1 | x, v)} = (-1)^{w-1} \sum_y p(v|y) \; p_i(x, y)$$

$$= (-1)^{w-1} \sum_y I_{R_v}(y) \; p_i(x, y),$$

$$\Rightarrow \qquad\qquad\qquad\qquad\qquad\qquad\qquad (5.9)$$

$$\frac{\partial S[w]}{\partial p(w = 1 | x, v)} = \sum_{w,i} C_{wi} \pi_i \frac{\partial p_i(w)}{\partial p(w = 1 | x, v)}$$

$$= \sum_{w,i} C_{wi} \pi_i (-1)^{w-1} \sum_y I_{R_v}(y) \; p_i(x, y).$$

Straightforward simplification of (5.8) and (5.9) leads to (5.3) and (5.4).

When the observations x and y are conditionally independent, i.e. $p_i(x, y) = p_i(x)p_i(y)$, then $p_i(x|y) = p_i(x)$ and $p_i(y|x) = p_i(y)$, in which case the decision regions (5.3) and (5.4) reduce to (5.5) and (5.6).

$$\square$$

The topologies of the decision regions (5.3) and (5.4) are in general unknown for an arbitrary distribution, and they do not simplify to LRT's unless under very special circumstances. For the example with normal distributions that follows, we will be able to determine the topologies of the decision regions.

5.3 Random signal in additive Gaussian noise

In this model, the sensor observations are given by

$$x = s + z_1, \quad y = s + z_2, \tag{5.10}$$

where $z_1 \sim N(0, \tau)$, $z_2 \sim N(0, \lambda)$, and the signal s is determined by two hypotheses

$$H_0: \; s = 0, \quad H_1: \; s \sim N(\mu, \sigma_s^2). \tag{5.11}$$

Equivalently,

$$
\begin{aligned}
H_0 : \ (x,y) &\sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \tau & 0 \\ 0 & \lambda \end{bmatrix}\right), \\
H_1 : \ (x,y) &\sim N\left(\begin{bmatrix} \mu \\ \mu \end{bmatrix}, \begin{bmatrix} \sigma_s^2 + \tau & \sigma_s^2 \\ \sigma_s^2 & \sigma_s^2 + \lambda \end{bmatrix}\right).
\end{aligned}
\tag{5.12}
$$

A very important parameter here is the signal variance σ_s^2. Notice that σ_s^2 determines the correlation coefficient of the bivariate normal distribution $p_1(x,y) = p(x,y|H_1)$. Thus, in the limit $\sigma_s^2 \to 0$, the observations become conditionally independent. Also, in the limit of large SNR, $\sigma_s \to \infty$ with τ and λ finite, the YX and XY processes approach identical performance since the decision rule (while still not determined by simple LRT's) becomes independent of τ and λ.

The above random signal model was considered in [12] and [13], where the decision of the first sensor Y was simply assumed to be based on a LRT, i.e.,

$$
R_{v=1} = \left\{ y : \ \frac{p_1(y)}{p_0(y)} > \lambda_0 \right\}, \quad \lambda_0 = \frac{C_{10} - C_{00}}{C_{01} - C_{11}} \frac{\pi_0}{\pi_1}.
\tag{5.13}
$$

The optimal decision rule, as derived in Theorem III.1 below, looks drastically different from the simple LRT. For the decision region at X however, our method (due to Lemma 5.1) gives the decision regions (5.15) and (5.16) which are of the same structure as those of [12] and [13]. Nevertheless, the difference in the decision region at Y alone may lead to remarkably different conclusions about the preferred communication direction. This is because the analysis done in [12] is no longer valid with our decision rule given by Theorem 5.2.

For simplicity, we assume that we have real samples $x, y \in \mathbb{R}$. The proof of the following theorem uses an idea due to [12] to determine the topologies of the decision regions.

Theorem 5.2. *The decision regions for this model are the following. At Y, there are thresholds $T_v^\pm \in \mathbb{R}$, $T_v^- < T_v^+$, such that*

$$R_{v=1} = \left\{ y : \frac{p_1(y)}{p_0(y)} \frac{Q(\frac{T_1^- - \mu_1(y)}{\sigma_1}) - Q(\frac{T_1^+ - \mu_1(y)}{\sigma_1}) - \left[Q(\frac{T_0^- - \mu_1(y)}{\sigma_1}) - Q(\frac{T_0^+ - \mu_1(y)}{\sigma_1}) \right]}{Q(\frac{T_1^-}{\sqrt{\tau}}) - Q(\frac{T_1^+}{\sqrt{\tau}}) - \left[Q(\frac{T_0^-}{\sqrt{\tau}}) - Q(\frac{T_0^+}{\sqrt{\tau}}) \right]} > \lambda_0 \right\},$$

(5.14)

where $\mu_1(y) = \frac{y + \mu\lambda/\sigma_s^2}{1 + \lambda/\sigma_s^2}$, $\sigma_1^2 = \tau + \frac{\lambda}{1 + \lambda/\sigma_s^2}$, *and λ_0 is as defined in Lemma 5.1.*

At X, there are thresholds $t^\pm \in \mathbb{R}$, $t^- < t^+$, such that

$$R_{w=1|v=1} = \left\{ x : \frac{p_1(x)}{p_0(x)} \frac{1 - [Q(\frac{t^- - \mu_2(x)}{\sigma_2}) - Q(\frac{t^+ - \mu_2(x)}{\sigma_2})]}{1 - [Q(\frac{t^-}{\sqrt{\lambda}}) - Q(\frac{t^+}{\sqrt{\lambda}})]} > \lambda_0 \right\},$$

(5.15)

$$R_{w=1|v=0} = \left\{ x : \frac{p_1(x)}{p_0(x)} \frac{Q(\frac{t^- - \mu_2(x)}{\sigma_2}) - Q(\frac{t^+ - \mu_2(x)}{\sigma_2})}{Q(\frac{t^-}{\sqrt{\lambda}}) - Q(\frac{t^+}{\sqrt{\lambda}})} > \lambda_0 \right\},$$

(5.16)

where $\mu_2(x) = \frac{x + \mu\tau/\sigma_s^2}{1 + \tau/\sigma_s^2}$ *and* $\sigma_2^2 = \lambda + \frac{\tau}{1 + \tau/\sigma_s^2}$.

Proof. The general form of the decision regions follows from Lemma 5.1. From (5.3) and (5.4), and following the proof of Lemma B.1, [12], let us define the following functions, which are the logarithms of the left hand sides of the defining inequalities of the decision regions.

$$f(y) = \log \frac{p_1(y)}{p_0(y)} + \log \frac{\sum_x [I_{R_{w=1|v=1}}(x) - I_{R_{w=1|v=0}}(x)] p_1(x|y)}{P_0(R_{w=1|v=1}) - P_0(R_{w=1|v=0})},$$
$$g_v(x) = \log \frac{p_1(x)}{p_0(x)} + \log \frac{\sum_y I_{R_v}(y) \, p_1(y|x)}{P_0(R_v)},$$

where $p_0(y) = n(y|0, \lambda)$, $p_1(y) = n(y|\mu, \lambda)$, $p_0(x) = n(x|0, \tau)$, $p_1(x) = n(x|\mu, \tau)$, $p_1(x|y) = n(x|\mu_1(y), \sigma_1^2)$, and $p_1(y|x) = n(y|\mu_2(x), \sigma_2^2)$ are normal pdfs.

Observe that $f(y)$ and $g_v(x)$ are convex, as can be seen by verifying that $f''(y) \geq$

0 and $g''_v(x) \geq 0$. That is,

$$f'(y) = \frac{y}{\lambda} - \frac{y - \mu}{\sigma_s^2 + \lambda} + \frac{\sum_x \left(\frac{x - \mu'_1(y)}{\sigma_1^2}\right) [I_{R_{w=1|v=1}}(x) - I_{R_{w=1|v=0}}(x)] p_1(x|y)}{\sum_x [I_{R_{w=1|v=1}}(x) - I_{R_{w=1|v=0}}(x)] p_1(x|y)}$$

$$= \frac{y}{\lambda} - \frac{y - \mu}{\sigma_s^2 + \lambda} + \left\langle \frac{x - \mu'_1(y)}{\sigma_1^2} \right\rangle_y,$$

$$\Rightarrow$$

$$f''(y) = \frac{\sigma_s^2}{\lambda(\sigma_s^2 + \lambda)} + \left\langle \left(\frac{x - \mu'_1(y)}{\sigma_1^2} - \left\langle \frac{x - \mu'_1(y)}{\sigma_1^2} \right\rangle_y \right)^2 \right\rangle_y \geq 0,$$

where $\langle \ \rangle_y$ denotes the self-evident conditional expectation involved, and similarly,

$$g'_v(x) = \frac{y}{\tau} - \frac{y - \mu}{\sigma_s^2 + \tau} + \frac{\sum_y \left(\frac{y - \mu'_2(x)}{\sigma_2^2}\right) I_{R_v}(y) p_1(y|x)}{\sum_y I_{R_v}(y) p_1(y|x)}$$

$$= \frac{y}{\tau} - \frac{y - \mu}{\sigma_s^2 + \tau} + \left\langle \frac{y - \mu'_2(x)}{\sigma_2^2} \right\rangle_x,$$

$$\Rightarrow$$

$$g''_v(x) = \frac{\sigma_s^2}{\tau(\sigma_s^2 + \tau)} + \left\langle \left(\frac{y - \mu'_2(x)}{\sigma_2^2} - \left\langle \frac{y - \mu'_2(x)}{\sigma_2^2} \right\rangle_x \right)^2 \right\rangle_x \geq 0.$$

Therefore, the decision regions take the form

$$R_{v=1} = [t^-, t^+]^c, \qquad R_{w=1|v} = [T_v^-, T_v^+]^c, \quad v = 0, 1, \qquad (5.17)$$

where $t^{\pm}$, $T_v^{\pm}$ are thresholds depending on λ_0 and the distribution parameters $\{\mu, \tau, \lambda, \sigma_s^2\}$, and A^c denotes complement of the set A. Substituting (5.17) in (5.3) and (5.4), we obtain the decision regions in terms of Q-functions as stated in the theorem. □

It is not difficult to see that as $\sigma_s^2 \to 0$, the decision regions (5.14), (5.15), and (5.16) are determined by simple LRT's as expected. Based on the derived topologies of the decision regions, we will now estimate and compare detection performance for the two directions, YX and XY.

5.4 Computational results for low signal to noise ratio

Having determined the global nature of the decision regions $R_{v=1} = [t^-, t^+]^c$, $R_{w=1|v} = [T_v^-, T_v^+]^c$, we now compute the threshold values that minimize the cost function $S[w]$. The optimal value of the cost function is given by

$$S_{\text{opt}}[w] = \sum_{w,i} C_{wi}\pi_i \sum_{x,y,v} I_{R_{w|v}}(x)\ I_{R_v}(y)\ p_i(x,y)$$

$$= \pi_0 \sum_w C_{w0} \sum_{x,y,v} I_{R_{w|v}}(x)\ I_{R_v}(y)\ p_0(x,y)$$

$$+ \pi_1 \sum_w C_{w1} \sum_{x,y,v} I_{R_{w|v}}(x)\ I_{R_v}(y)\ p_1(x,y)$$

$$= \pi_0 \sum_w C_{w0} \sum_v p_0(R_{w|v})\ p_0(R_v) + \pi_1 \sum_w C_{w1} \sum_{x,y,v} I_{R_{w|v}}(x)\ I_{R_v}(y)\ p_1(x,y)$$

$$= \pi_0 C_{00} + \pi_0(C_{10} - C_{00}) \sum_v p_0(R_{w=1|v})\ p_0(R_v) + \pi_1 C_{01}$$

$$- \pi_1(C_{01} - C_{11}) \sum_{x,y,v} I_{R_{w=1|v}}(x)\ I_{R_v}(y)\ p_1(x,y)$$

$$= \min_T\ S(T), \quad T = (t^-, t^+, T_0^-, T_1^-, T_0^+, T_1^+).$$

Consider the special but important case of minimizing the error probability (corresponding to $C_{00} = C_{11} = 0,\ C_{10} = C_{01} = 1$). We have

$$P_e[w] = \pi_0 \sum_v p_0(R_{w=1|v} \times R_v) + \pi_1 \sum_v p_1(R_{w=0|v} \times R_v)$$

$$= \pi_0 \sum_v p_0(R_{w=1|v} \times R_v) + \pi_1 - \pi_1 \sum_v p_1(R_{w=1|v} \times R_v)]$$

$$= \pi_0[p_0(R_{w=1|v=1} \times R_{v=1}) - p_0(R_{w=1|v=0} \times R_{v=1})$$

$$+ p_0(R_{w=1|v=0})] + \pi_1 - \pi_1[p_1(R_{w=1|v=1} \times R_{v=1})$$

$$- p_1(R_{w=1|v=0} \times R_{v=1}) + p_1(R_{w=1|v=0})].$$

The inequalities $t^- < t^+$, $T_0^- < T_0^+$, and $T_1^- < T_1^+$ are constraints on the optimization problem.
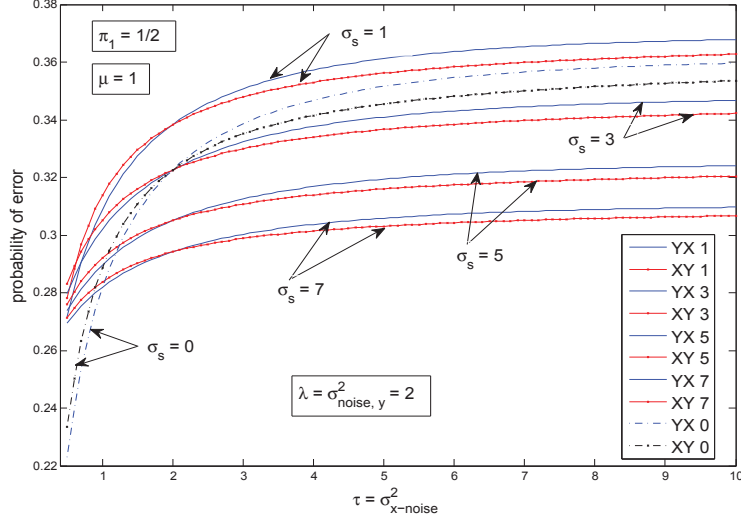
Figure 5.2: Performance of YX and XY directions

Results for signal variance σ_s^2 near 0 are given in Fig. 5.2. In this figure, the prior probability $\pi_1 = 1/2$, but the same behavior is observed for various values of $\pi_1 \neq 1/2$. For various values of $\sigma_s = 1, 3, 5, 7$, the behavior is identical to that for $\sigma_s = 0$ where the observations are conditionally independent. For each value of σ_s, the figure clearly shows that when $\tau < \lambda$ (i.e., when X is the better sensor), YX performs better as it has smaller probability of error, meanwhile XY performs better when $\tau > \lambda$ (i.e., when Y is the better sensor). Hence the better sensor is the preferred FC for $0 \leq \sigma_s \leq 7$. There is no apparent reason why this result should not extend to large σ_s, especially as our decision rules are valid for all σ_s. However, the large σ_s regime requires a more efficient code for numerically computing the optimal threshold values than what is available. Therefore we have

deferred to future work on the large correlation regime.

Conclusion

Based on the results of Chapter 3, have derived the general form of the optimal Bayesian test for a two sensor tandem fusion network with dependent observations. Application of this test to a random signal in additive Gaussian noise shows that for small correlation strength, measured by the signal to noise ratio, the sensor with the cleaner data is still the preferred fusion center. This is in agreement with the case of independent observations.

Chapter 6

Detection over Acyclic Graphs

6.1 Introduction

This section is based on ongoing work on distributed detection over acyclic directed graphs. It is a continuation of the discussion of sensor network rules initiated in Section 3.5. In the Bayesian framework, and under mild assumptions on the dependence structure of the local sensor decisions, we obtain decision rules for arbitrary directed graphs. We also briefly study large sample asymptotic analysis, and derive associated Chernoff and Kullback information measures. These information measures are then used to define online sensor comparison with respect to asymptotic optimality in the network.

It is certainly the case that detection of a phenomenon by a network of distributed sensors is analytically more complex than detection of the same phenomenon by a centralized sensor network. This complexity arises for a number of reasons which include the following.

(i) The decision rules for different sensors are coupled in such a way that exact

analysis may be impossible, especially when the observations of the sensors are conditionally dependent.

(ii) For a given sensor X in the network, the number of decision thresholds for X grows exponentially with the number of sensors that transmit their decisions to X. If I_X is the number of sensors transmitting to X, then X requires at least 2^{I_X} thresholds for binary processing.

(iii) When sensor X is allowed to transmit more than one type of message (i.e., different sensors receive different messages from X), the number of decision channels connecting parents to offsprings of X increases. If O_X is the number of distinct messages X is transmitting, this leads to a total of at least 2^{O_X} more computational steps in binary processing. Moreover, X now requires at least $O_X \times 2^{I_X}$ thresholds for binary processing. An even more interesting consequence of the ability of sensors to transmit multiple messages is that decision processing over a closed path in a directed graph becomes nontrivial.

If we relax (i) and (iii) by considering only conditionally independent observations, and also require that each sensor transmits only one type of message, then the decision rules even for the most elaborate graphical networks closely resemble those of simple networks. Under these assumptions, our main objective is to determine the optimal decision threshold structure in any sensor network for which desired communication and fusion patterns already exist in the form of directed graphs. See Fig 6.1 for example.

Another objective will be to define online sensor comparison based on large sample asymptotic analysis. Let us refer to this type of comparison as AAR (*asymptotic accuracy rate*, or *Chernoff information*) comparison. AAR comparison can serve as an alternative to the usual ROC (receiver operating curve) sensor comparison

in the large sample regime. AAR comparison works for M-ary decisions, and as we will see, it is robust with respect to Bayesian cost structure and prior probability.

We can also use AAR comparison to compare graphical network patterns in the following way. Suppose we are given K sensors $X^n = (X_1, ..., X_K)$. Let $\mathcal{G}(X^n, L)$ be the collection of all network patterns G with at most L direct links between any two sensors from X^n. For any given network pattern $G \in \mathcal{G}(X^n, L)$, define an online fusion center for G to be the sensor X_G in G that has the largest AAR. Then it is natural to define the optimal network pattern to be the network pattern $G \in \mathcal{G}(X^n, L)$ such that the AAR of X_G is larger than that of $X_{G'}$ for all $G' \in \mathcal{G}(X^n, L)$.

Distributed detection over directed graphs with at most one path between any two nodes was studied in [3, 4], where some optimal control techniques were developed and optimal communication architectures were discussed. We consider distributed detection over arbitrary acyclic directed graph networks. Meanwhile the methods of [3, 4] penalized error at every root node in a tree, for simplicity, we consider only the costs at a single fusion center. Because we are interested mainly in the fused decision, all graphs are assumed to be connected. Bayesian detection over graph networks (feedback/memory included) was discussed in Section 4.5 of [8], where decision rules were derived under binary hypothesis testing. We obtain decision rules under M-ary hypothesis testing but, for computational purposes, we likewise restrict to binary decisions. Because our main focus is on the graphical structure, we also assume the sensors take only one set of observations so that the decision process is *static*, i.e., memory and feedback are not included in our discussion, although multiple processing steps can occur if we allow for cyclic communication paths in the network.
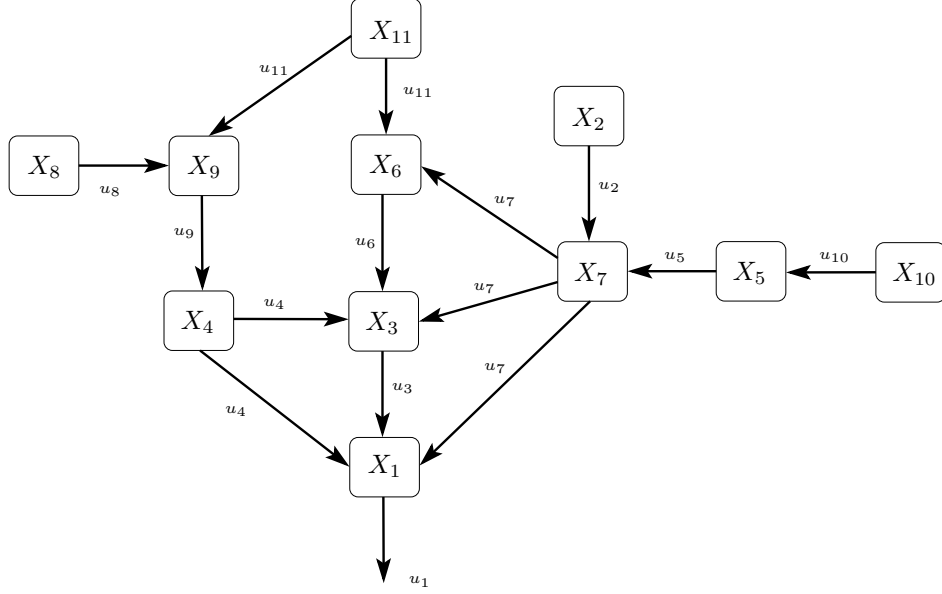
Figure 6.1: An 11-node acyclic directed graph

For an illustration of our notation, consider the 11-node directed graph of Fig 6.1. Every arrow denotes a communication channel and the direction in which information must flow. Each node X_j represents a sensor whose observation we denote by $x_j \in \mathcal{X}_j$, where the alphabet $\mathcal{X}_j$ is for simplicity taken to be the vector space $\mathbb{R}^{n_j}$ for a positive integer n_j. The decision of sensor X_j is denoted by $u_j = \gamma_j(x_j, \tilde{u}_j)$, where γ_j is an integer-valued function and $\tilde{u}_j$ is the set of decisions of all parents of X_j. The decision u_j is passed on to every offspring of X_j. For example, in Fig 6.1, sensor X_7's decision $u_7 = \gamma_7(x_7, \tilde{u}_7)$, where $\tilde{u}_7 = \{u_2, u_5\}$, is transmitted to the sensors X_1, X_3, and X_6. By convention, if there are n nodes in a graph, we consider X_1 to be the fusion center, after which the labeling and placement of all other nodes $X_2, X_3, ..., X_n$ may come in any order.

In the work of [32] and related literature, e.g., Sections 4.4 and 4.5 of [41], two sensors are compared in terms of their stand-alone receiver operating characteristic

(ROC) curves. ROC curve comparison is useful if the sensor quality is determined by more than one parameter, including variance and prior probability. It is also useful when each sensor has a separate local objective and the overall network objective is designed in such a way that it depends on these local objectives. A downside of ROC curve comparison is that it only works for binary decisions, since in that case the detection probability and false alarm probability are related by a single threshold parameter. In our analysis, we assume for simplicity that (i) the overall network objective is independent of any local objectives, (ii) every sensor uses the same prior probability, (iii) a sensor's quality is determined solely by the variance of its observation.

In what follows we first describe graphical networks and provide the threshold structure of decision rules for acyclic graph networks, including effects of communication over nonideal channels. This is followed by a short discussion on large sample analysis, from which the obtained Chernoff and Kullback information measures are then used to define online comparison among individual sensors and comparison of network patterns.

Notation: Let us recall the notation introduced in Section 3.1. Sensors are labeled using upper case letters X, Y, Z, and so on. Random variables, as well as their values, are denoted by lower case letters x, y, z, u, v, w, etc. Also, we do not distinguish between summation and integration symbols, i.e., if u is discrete and x is continuous, we write $\sum_{u,x} = \sum_u \sum_x$, where $\sum_u$ denotes summation over u, and $\sum_x$ denotes integration over x. Similarly, if u, v are discrete and x, y are continuous, we write $\delta_{(u,x)(v,y)} = \delta_{uv}\delta_{xy}$, where δ_{uv} is the Kronecker delta, while δ_{xy} is the Dirac delta.

6.2   Graphical networks

Recall that optimal Bayesian decision rules for acyclic graphs were derived in Section 3.5, and presented in Theorem 3.3. The following observations about an acyclic graph network are what made that result possible.

1. For any given pair of sensors X_i, X_j in a network, multiple directed paths between X_i and X_j do not lead to any intersections between different decision regions as long as these paths have the same sense, i.e., they do not form a loop.

2. Processing in a closed path can be nontrivial (i.e., distributed, or decentralized) only if a sensor on that path is allowed to send multiple messages.

Therefore the simple network decision rules we have seen in the previous chapters, as well as their experimental implementation, can be readily extended to arbitrary directed graphs provided we make one more simplifying assumption about the sensor network, besides conditional independence of observations. The assumption is that every sensor passes the same message to its offsprings. This eliminates networks containing closed processing paths.

Note, however, that a sensor can of course be allowed (if necessary) to send multiple messages in a directed acyclic graph, and that the single-message decision rules we will obtain can easily be extended to multiple-message decision rules for any acyclic directed graph.

Acyclic directed graphs

Consider a directed graph $G = (V, \rho)$, where $V = \{X_1, ..., X_n\}$ are vertices,

$$V \times V = \{(X_i, X_j) : i = 1, ..., n, \; j = 1, ..., n\}$$

is the set of placeholders (i.e., possibilities) for directed arrows connecting the vertices [thus (X_i, X_j) represents the possibility of an arrow directed from X_i to X_j], and the map

$$\rho : V \times V \to \{0, 1\}, \ \ (X_i, X_j) \mapsto a_{ij} = \rho(X_i, X_j)$$

indicates the presence or absence of an arrow, i.e., $a_{ij} = 1$ says an arrow pointing from X_i to X_j exists, while $a_{ij} = 0$ means such an arrow does not exist. For computational purposes, the links in the graph are more conveniently written in matrix form: the matrix of G is given by

$$M_G = \sum_{i=1}^{n} \sum_{j=1}^{n} \rho(X_i, X_j) e_{ij} = \sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij} e_{ij},$$

where e_{ij} is the matrix with (i, j)th entry 1, and 0 for all other entries. For example, the matrix of the graph in Fig 6.1 is

$$\begin{bmatrix}
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & 1 & \cdot & \cdot & \cdot \\
1 & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\
1 & \cdot & 1 & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\
\cdot & \cdot & \cdot & \cdot & \cdot & 1 & \cdot & \cdot & \cdot & \cdot \\
\cdot & \cdot & 1 & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\
1 & \cdot & 1 & \cdot & \cdot & 1 & \cdot & \cdot & \cdot & \cdot \\
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & 1 & \cdot & \cdot \\
\cdot & \cdot & \cdot & 1 & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\
\cdot & \cdot & \cdot & \cdot & 1 & \cdot & \cdot & \cdot & \cdot & \cdot \\
\cdot & \cdot & \cdot & \cdot & \cdot & 1 & \cdot & \cdot & 1 & \cdot & \cdot
\end{bmatrix}$$

Notice that (i) the number of 1-entries is the number of arrows in the graph, (ii) the 1's in the jth column correspond to the parents of X_j, and (iii) the 1's in the ith row correspond to offsprings of X_i. The total number of thresholds is

$$2^3 + 2^0 + 2^3 + 2^1 + 2^1 + 2^2 + 2^2 + 2^0 + 2^2 + 2^0 + 2^0 = 36.$$

These observations are crucial for developing a general optimization code for the graphical networks.

Denoting the observation and decision of X_j by x_j and u_j respectively, the dependence structure of the decisions is given by

$$u_j = u_j(x_j, \tilde{u}_j), \quad \tilde{u}_j = \{u_i : i = 1, ..., n, \ a_{ij} = 1\},$$

where $\tilde{u}_j$ consists of the decisions of the parents of X_j (i.e, all nodes $\tilde{X}_j = \{X_i : i = 1, ..., n, \ a_{ij} = 1\}$ bearing arrows into X_j) in the graph G.

If a node X_k has I_k parents (i.e., in-degree) then its number of thresholds is 2^{I_k}. Since we have assumed that each node passes the same message to all of its offsprings, the total number of thresholds is

$$\sum_{k=1}^{n} 2^{I_k}.$$

As before, let X_1 be the fusion center. Also let $\vec{x} = (x_1, ..., x_n)$ and $\vec{u} = (u_1, ..., u_n)$, and consider the risk function

$$S = \sum_{u_1, h} C_{u_1 h} p(u_1, h) = \sum_{\vec{x}, \vec{u}, h} C_{u_1 h} \prod_{i=1}^{n} p(u_i | x_i, \tilde{u}_i) \ p_h(\vec{x}) \pi_h$$

$$\stackrel{(a)}{=} \sum_{\vec{x}, \vec{u}, h} C_{u_1 h} \prod_{i=1}^{n} p(u_i | x_i, \tilde{u}_i) \prod_{i=1}^{n} p_h(x_i) \pi_h,$$

where step (a) holds for conditionally independent observations.

Theorem 6.1. *The binary decision rule for the network $\{X_1, ..., X_n\}$ viewed as an acyclic directed graph are as follows. We have $p_{opt}(u_k | x_k, \tilde{u}_k) = I_{R_{u_k | \tilde{u}_k}}(x_k)$, with decision regions*

$$R_{u_k = 1 | \tilde{u}_k} = \left\{ x_k : \frac{\partial^B S}{\partial p_{opt}(u_k = 1 | x_k, \tilde{u}_k)} < 0 \right\} = \left\{ x_k : \frac{p_1(x_k)}{p_0(x_k)} > \lambda_{\tilde{u}_k}^{(k)} \right\},$$

$$where \quad \lambda_{\tilde{u}_1}^{(1)} = \lambda \frac{\sum_{\vec{u}\setminus(u_1,\tilde{u}_1)} \prod_{i\neq 1} p_0(R_{u_i|\tilde{u}_i})}{\sum_{\vec{u}\setminus(u_1,\tilde{u}_1)} \prod_{i\neq 1} p_1(R_{u_i|\tilde{u}_i})}, \quad \lambda = \frac{(C_{10}-C_{00})(1-\pi)}{(C_{01}-C_{11})\pi}, \quad and$$

$$\lambda_{\tilde{u}_k}^{(k)} = \lambda \frac{\sum_{\vec{u}\setminus\{u_1,\tilde{u}_k\}} (-1)^{u_k-1} p_0(R_{u_1=1|\tilde{u}_1}) \prod_{i\notin\{1,k\}} p_0(R_{u_i|\tilde{u}_i})}{\sum_{\vec{u}\setminus\{u_1,\tilde{u}_k\}} (-1)^{u_k-1} p_1(R_{u_1=1|\tilde{u}_1}) \prod_{i\notin\{1,k\}} p_1(R_{u_i|\tilde{u}_i})}, \quad for \; k \in \{2,3,...,n\}.$$

The optimal value of the risk function is

$$S_{opt} = \sum_{\vec{x},\vec{u},h} C_{u_1 h} \prod_{i=1}^n I_{R_{u_i|\tilde{u}_i}}(x_i) \prod_{i=1}^n p_h(x_i) \, \pi_h = \sum_{\vec{u},h} C_{u_1 h} \prod_{i=1}^n p_h(R_{u_i|\tilde{u}_i}) \, \pi_h$$

$$\stackrel{(a)}{=} \pi \left[1 - \sum_{\vec{u}\setminus u_1} p_1(R_{u_1=1|\tilde{u}_1}) \prod_{i\neq 1} p_1(R_{u_i|\tilde{u}_i}) \right] + (1-\pi) \sum_{\vec{u}\setminus u_1} p_0(R_{u_1=1|\tilde{u}_1}) \prod_{i\neq 1} p_0(R_{u_i|\tilde{u}_i}),$$

(6.1)

where step (a) holds for 0-1 cost.

Proof. This is the result of Theorem 3.3 simplified for binary decisions. □

For a numerical outcome of Theorem 6.1, consider the following example.

Example. Let the observations of the sensors $(X_1, ..., X_n)$ be given by

$$x_k = s + b_k, \quad k = 1, ..., n,$$

where $b_k \sim N(0, \sigma_k^2)$ and the hypotheses are

$$H_0 : s = 0, \quad H_1 : s = 1.$$

For the above example, Fig 6.3 displays the relative performance of the acyclic graph in Fig 6.1 and the binary tree in Fig 6.2. For this numerical computation, we have assumed the sensors are identical, i.e., $\sigma_k = \sigma_1$ for all k.
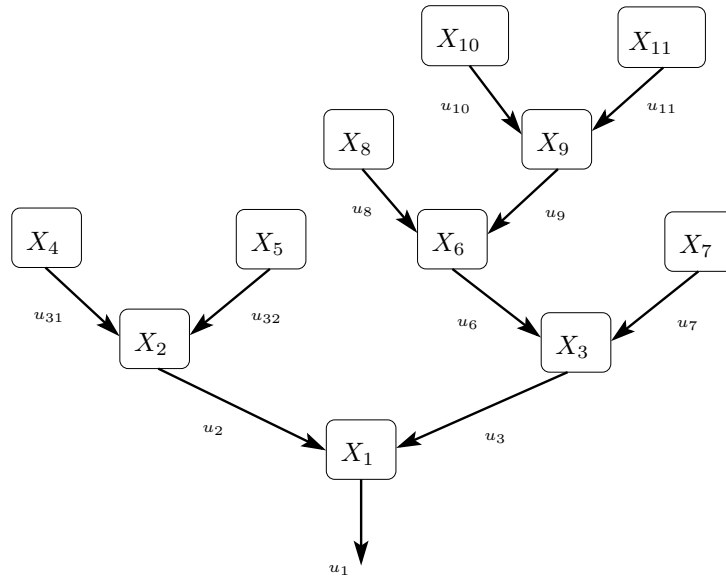
Figure 6.2: An 11-node binary tree network
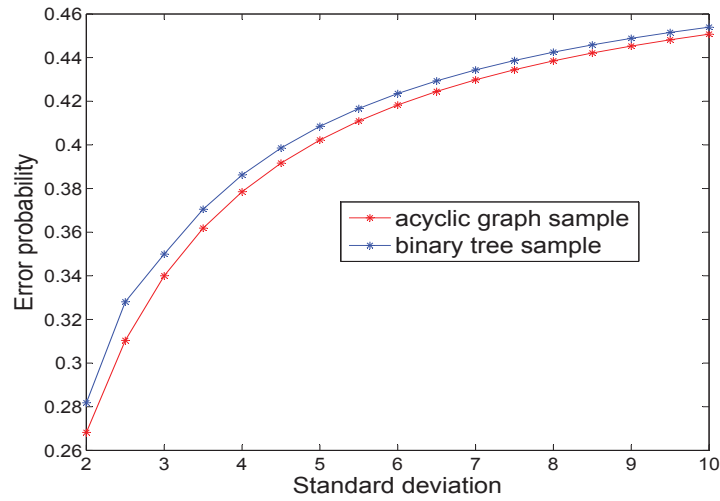
Figure 6.3: Relative performance of the sample acyclic graph of Fig 6.1 and the sample binary tree of Fig 6.2. The graphs each contain 11 sensors but the acyclic graph has greater connectivity and thus performs better.

We also have the following result, which has been obtained in certain fusion settings by [45, 46].

Theorem 6.2 (Nonideal channel effects)**.** *Let $X_1, X_2, \cdots, X_n$ be a sensor network in the form of an acyclic directed graph. Then the Bayesian decision rules for the network have the same form as in Theorem 6.1, except that everywhere, we replace $p_h(R_{u_k|\tilde{u}_k})$ with $\sum_{\tilde{u}'_k} p_h(R_{u_k|\tilde{u}'_k})p(\tilde{u}'_k|\tilde{u}_k)$, where $[p(\tilde{u}'_k|\tilde{u}_k)]$ is a $2^{I_k} \times 2^{I_k}$ multi-channel transition matrix, with I_k the number of parents of X_k. In particular, the objective function (6.1) becomes*

$$S_{opt} = \sum_{\vec{u},\vec{u}',h} \pi_h C_{u_1 h} \prod_{k=1}^{n} p_h(R_{u_k|\tilde{u}'_k}) \; p(\tilde{u}'_k|\tilde{u}_k). \tag{6.2}$$

Proof. If X_j is transmitting its decision u_j to X_i through a channel

$$g_{ji} : u_j \mapsto u'_j = g_{ji}(u_j),$$

where g_{ji} is a random function independent of the observations, then $u'_j = g_{ij}(u_j) \in \tilde{u}'_i$. The objective function can be expanded as

$$S = \sum_{u_1,h} C_{u_1 h} p(u_1, h) = \sum_{\vec{u},\vec{u}',\vec{x},h} C_{u_1 h} p(\vec{u}, \vec{x}, \vec{u}', h)$$

$$= \sum_{\vec{u},\vec{u}',\vec{x},h} \pi_h C_{u_1 h} \prod_{k=1}^{n} p(u_k|x_k, \tilde{u}'_k) \; p(\tilde{u}'_k|\tilde{u}_k) \; p_h(\vec{x}),$$

where $\vec{u} = (u_1, ..., u_n)$ are the decisions of the sensors, meanwhile $\vec{\tilde{u}}' = (\tilde{u}'_1, ..., \tilde{u}'_n)$ are the messages received by the sensors, i.e., $\tilde{u}'_k$ consists of the messages channeled to X_k by its parents. The decisions have the dependence structure $u_k = u_k(x_k, \tilde{u}'_k)$. Thus a straightforward application of the decision procedure developed in Section 3.5 yields the desired result, and in particular, (6.2) holds. $\square$

On directed graphs with cycles
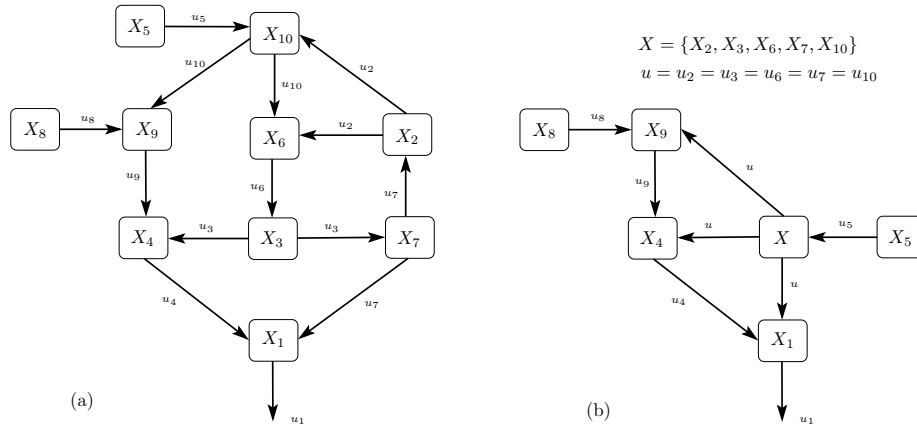
Figure 6.4: A 10-node directed graph containing cycles. The graphs (a) and (b) are equivalent.

Let us again assume that every sensor passes the same message to its offsprings. Then without any further constraint on the *dependence structure* of the decisions (*as determined by the arrows in the graph*), a closed path in a graph is trivial in the sense that it is equivalent to a centralized sub-collection of sensors. The situation may be described more precisely as follows.

Lemma 6.3. *In a graphical sensor network, suppose a closed path consists of r sensors $X_{i_1}, X_{i_2}, ..., X_{i_r}$, with respective observations $x_{i_1}, x_{i_2}, ..., x_{i_r}$. Suppose further that*

- *the dependence structure of the decisions of the sensors is determined only by the directed arrows in the graph, and*

- *every sensor in the network passes the same message to its offsprings.*

Then the subnetwork $Y = \{X_{i_1}, X_{i_2}, ..., X_{i_r}\}$ behaves like a centralized sensor system with all observations $y = \{x_{i_1}, x_{i_2}, ..., x_{i_r}\}$ available at the same location.

Proof. If no constraint (apart from that imposed by the arrows) is placed on the dependence struture of the decisions, then in particular, there is no *timing constraint*. Thus information will continue to flow in the closed path until the decision of very sensor has achieved maximum performance, which is *predicted* to be the performance of a centralized network formed by the r sensors. $\square$

The above discussion shows that directed graphs containing cycles are equivalent to acyclic directed graphs, unless there are extra communication constraints to render the cyclic processing nontrivial.

6.3 Large sample asymptotic analysis

Fix $k \in \{1, ..., n\}$. Let X_k make a T-observations sample, $x_k^T = (x_{k,1}, ..., x_{k,t}, ..., x_{k,T})$, and make a corresponding sequence of decisions

$$u_k^T = u_k^T(x_k^T, \tilde{u}_k^{T-1}) = (u_{k,1}, ..., u_{k,t}, ..., u_{k,T}).$$

The decisions are made in a *sequential* manner, i.e., a decision is made after each sample. Thus for each $t \in \{1, ..., T\}$,

$$u_{k,t} = u_{k,t}(x_k^t, \tilde{u}_k^{t-1}) \in \{0,1\}^{r_k} \simeq \{1, 2, ..., 2^{r_k}\} \tag{6.3}$$

is an r_k-bit random variable. Here, r_k is sensor X_k's transmission rate (in bits per sample).

Let M_k be the size of a linear indexing set for the value set of the decision sequence u_k^T, i.e., as a random variable, u_k^T can take on M_k possible values simply

labeled as $\{u_k^T(m_k) : m_k = 1, ..., M_k\} = \{u_k^T(1), ..., u_k^T(M_k)\}$. Then the number of bits per sample r_k can be written as

$$r_k = \frac{1}{T} \log_2 M_k \quad \text{or} \quad M_k = 2^{r_k T}.$$

Thus for each $m_k \in \{1, ..., M_k\}$, we have

$$u_k^T(m_k) \in (\{0, 1\}^{r_k})^T = \{0, 1\}^{Tr_k} \simeq \{1, 2, ..., 2^{r_k}\}^T$$

$$\simeq \{1, 2, \cdots, 2^{Tr_k}\} = \{1, 2, \cdots, M_k\},$$

$$\Rightarrow \quad u_k^T : (x_k^T, \tilde{u}_k^{T-1}) \overset{\simeq}{\longmapsto} m_k \in \{1, 2, \cdots, M_k\},$$

where all powers of sets are Cartesian, and $\simeq$ denotes equivalence with respect to cardinality (i.e., $A \simeq B$ means "the sets A and B have the same number of elements"). Furthermore, we may wish to impose a rate constraint

$$\frac{1}{T} \log_2 M_k = r_k \leq R_k, \tag{6.4}$$

where R_k is a fixed bit rate representing the maximum number of bits per sample that X_k can transmit to its offsprings.

For simplicity, let the bit rate of the fusion center X_1 be set at $r_1 = 1$. Then the error probabilities (based on the final decision $u_{1,T}$) at the fusion center are

$$\alpha_T = p(u_{1,T} = 1 | H_0) = p_0(u_{1,T} = 1), \quad \text{(False alarm)},$$

$$\beta_T = p(u_{1,T} = 0 | H_1) = p_1(u_{1,T} = 0), \quad \text{(Missed detection)}.$$

With a constraint $\alpha_T \leq \varepsilon_T$ on the false alarm, we define the error exponent for missed detection as

$$b(\varepsilon, R_1, ..., R_n) = \liminf_{T \to \infty} \ -\frac{1}{T} \log \left(\inf_{A_T} \ \beta_T \right), \tag{6.5}$$

where

$$A_T = \left\{ \ \left\{ p\left(u_k^T | x_k^T, \tilde{u}_k^{T-1}\right) \right\}_k : \{r_k \leq R_k\}_k, \ \alpha_T \leq \varepsilon_T \right\}$$

$$\simeq \left\{ \ \left\{ p\left(m_k | x_k^T, \tilde{m}_k\right) \right\}_k : \{r_k \leq R_k\}_k, \ \alpha_T \leq \varepsilon_T \right\}$$

is the set of possible decision strategies. That is, we have

$$\beta_T \approx e^{-b(\varepsilon, R_1, \ldots, R_n)T} \quad \text{for large } T.$$

In the Bayesian formulation, the risk function at X_1 has the expansion

$$
\begin{aligned}
S_T &= \sum_{u_{1,T},h} C_{u_{1,T}h} \, p(u_{1,T}, h) = \sum_{u_{1,T},h} \pi_h C_{u_{1,T}h} \, p(u_{1,T}|h) \\
&= \sum_{u_{1,T},h} \pi_h C_{u_{1,T}h} \, p_h(u_{1,T}) = \sum_{\vec{u}^T, \vec{x}^T, h} \pi_h C_{u_{1,T}h} \, p_h(\vec{u}^T, \vec{x}^T) \\
&= \sum_{\vec{u}^T, \vec{x}^T, h} \pi_h C_{u_{1,T}h} \prod_{k=1}^{n} p\left(u_k^T | x_k^T, \tilde{u}_k^{T-1}\right) \, p_h\left(\vec{x}^T\right) \\
&= \sum_{\vec{m}, \vec{x}^T, h} \pi_h C_{u_{1,T}h} \prod_{k=1}^{n} p\left(m_k | x_k^T, \tilde{m}_k\right) \, p_h\left(\vec{x}^T\right) \\
&= \sum_{\vec{m}, h} \sum_{\left\{x_k^T \in R_{m_k|\tilde{m}_k}\right\}_k} \pi_h C_{u_{1,T}h} \, p_h\left(\vec{x}^T\right),
\end{aligned}
\tag{6.6}
$$

where $\vec{u}^T = (u_1^T, \cdots, u_n^T)$, $\vec{x}^T = (x_1^T, \cdots, x_n^T)$, and $\vec{m} = (m_1, \ldots, m_n)$. The main challenge is to design the decision regions $\{R_{m_k|\tilde{m}_k} : k = 1, \ldots, n\}$ in an optimal way. For large T, the law of large numbers says these regions can be approximated by sets of sequences x_k^T whose *empirical distributions*

$$p(a|x_k^T) = \frac{|\{t : x_{k,t} = a\}|}{T} = \frac{\text{cardinality}\{t : x_{k,t} = a\}}{T}, \quad a \in \mathcal{X}_k,$$

are close (in a certain sense) to the marginal distributions for $p_h\left(\vec{x}^T\right)$. Using this reasoning, it is possible to derive bounds on the error exponent $b(\varepsilon, R_1, \ldots, R_n)$.

We can use a likelihood ratio quantizer to provide an upper bound for S_T as follows. Let

$$R_{m_k|\tilde{m}_k}^{(LR)} = \left\{x_k^T : \lambda_{m_k|\tilde{m}_k}^{(k)} < l_h(x_k^T) < \lambda_{m_k+1|\tilde{m}_k}^{(k)}\right\},$$

where $l_h(x_k^T) = \frac{\max_{h'} p_{h'}(x_k^T)}{p_h(x_k^T)}$, and the thresholds λ depend on T. Then for any $\varepsilon_k > 0$, we can choose T large enough so that

$$S_T \leq \sum_{u_1,h} \pi_h C_{u_1,T h} \prod_{k=1}^{n} e^{-T \left[L_h^{(k)} - \varepsilon_k - r_k \right]},$$

$$L_h^{(k)} = \lim_{T \to \infty} \max_{p(m_k)} -\frac{1}{T} \sum_{m_k} p(m_k) \log p_h \left(\lambda_{m_k|\tilde{m}_k}^{(k)} < l_h(x_k^T) < \lambda_{m_k+1|\tilde{m}_k}^{(k)} \right).$$

Thus, if

$$r_k < L_h^{(k)} - \varepsilon_k,$$

then $S_T \to 0$ as $T \to \infty$. The quantity $L_h^{(k)}$ is an example of an information measure at X_k, since it is a function of the *likelihood ratio statistic* $l_h(x_k^T)$ and determines *estimation accuracy* to some extent. Concrete examples of information measures are the following.

6.4 Chernoff information and Kullback-Leibler distance

The discussion here is related to the discussion in Section 11.8 of [33]. The Chernoff information is derived as follows (Note here that neither the hypothesis h nor the decision $u_{1,T}$ needs to be binary). With the usual optimization variables

$p(u_k|x_k^T, \tilde{u}_k^T)$ in mind, we have

$$\min S_T = \min \sum_{u_{1,T},h} C_{u_{1,T}h}\, p(u_{1,T}, h)$$

$$= \min \sum_{u_{1,T},h,x_1^T,\tilde{u}_1^T} C_{u_{1,T}h}\, p(u_{1,T}|x_1^T, \tilde{u}_1^T) p_h(x_1^T, \tilde{u}_1^T)\pi_h$$

$$\overset{(s1)}{=} \sum_{u_{1,T},h,x_1^T,\tilde{u}_1^T} C_{u_{1,T}h}\, I_{R_{u_{1,T}|\tilde{u}_1^T}}(x_1^T) p_h(x_1^T, \tilde{u}_1^T)\pi_h$$

$$\overset{(s2)}{=} \sum_{u_{1,T},h,x_1^T,\tilde{u}_1^T} C_{u_{1,T}h}\, I_{R_{u_{1,T}}}(x_1^T, \tilde{u}_1^T) p_h(x_1^T, \tilde{u}_1^T)\pi_h$$

$$\overset{(s3)}{=} \sum_{x_1^T,\tilde{u}_1^T} \min_{u_{1,T}} \sum_h \pi_h C_{u_{1,T}h}\, p_h(x_1^T, \tilde{u}_1^T)$$

$$\overset{(s4)}{=} \sum_{x_1^T,\tilde{u}_1^T} \min_{\lambda}\, e^{\sum_{u_{1,T}} \lambda_{u_{1,T}} \log \sum_h \pi_h C_{u_{1,T}h}\, p_h(x_1^T,\tilde{u}_1^T)}$$

$$= \min_{\lambda} \sum_{x_1^T,\tilde{u}_1^T} e^{\sum_{u_{1,T}} \lambda_{u_{1,T}}(x_1^T,\tilde{u}_1^T) \log \sum_h \pi_h C_{u_{1,T}h}\, p_h(x_1^T,\tilde{u}_1^T)},$$

where $\sum_{u_{1,T}} \lambda_{u_{1,T}} = 1$. Step (s1) is due to the optimal decision rule at X_1. Step (s2) is simply step (s1) along with the assumption that the optimal rule at $\tilde{X}_1$ is already given, so that both x_1^T and $\tilde{u}_1^T$ are treated observational data by X_1. Thus, we have introduced a *conditionally optimal* rule at X_1, with decision regions $R_{u_{1,T}}$. Step (s3) is merely a reinterpretation of the (conditionally) optimal rule at X_1. Step (s4) is due to a familiar mathematical identity $\min(|a_1|, |a_2|, ...) = \min_{\lambda:\sum \lambda_i=1} \prod_i |a_i|^{\lambda_i}$.

We define the *Chernoff information* for testing M hypotheses at X_1 to be

$$\mathcal{C}(p_0, p_1, \cdots, p_{M-1}) = \lim_{T\to\infty}\; -\frac{1}{T}\log\; \min(S_T)$$

$$\overset{(s)}{=} -\min_{\lambda}\left[\log \sum_{x_1,\tilde{u}_1} e^{\sum_{u_1} \lambda_{u_1}(x_1,\tilde{u}_1) \log\left(\max_h c_{u_1,h}\, p_h(x_1,\tilde{u}_1)\right)}\right],$$

where $\sum_{u_1} \lambda_{u_1} = 1$,

$$c_{u_1,h} = \lim_{T\to\infty} [\pi_h C_{u_1,T h}]^{1/T} = \lim_{T\to\infty} [C_{u_1,T h}]^{1/T} \; \in \; \{0,1\},$$

and step (s) assumes the independence $p_h(x_1^T, \tilde{u}_1^T) = \prod_{t=1}^T p_h(x_{1,t}, \tilde{u}_{1,t})$, as well as uses the well known identity

$$\lim_{T \to \infty} \left(\sum_i |a_i|^T \right)^{1/T} = \max_i \ |a_i|.$$

Note that if X_k, $k \neq 1$, is viewed as an independent fusion center with risk

$$S_T^{(k)} = \sum_{u_{k,T}, h} C_{u_{k,T}h}^{(k)} \ p(u_{k,T}, h),$$

then the above procedure can be repeated to obtain the Chernoff information at X_k as

$$\mathcal{C}^{(k)}(p_0, p_1, \cdots, p_{M-1}) = \lim_{T \to \infty} \ -\frac{1}{T} \log \ \min(S_T^{(k)})$$

$$= -\min_\lambda \left[\log \sum_{x_k, \tilde{u}_k} e^{\sum_{u_k} \lambda_{u_k}(x_k, \tilde{u}_k) \log \left(\max_h c_{u_k,h}^{(k)} p_h(x_k, \tilde{u}_k) \right)} \right].$$

When h and u_1 are both binary, and we consider 0-1 costs, then $c_{u_1,h} = 1 - \delta_{u_1,h}$, and we obtain the usual expression for Chernoff information,

$$\mathcal{C}(p_0, p_1) = -\min_{0 \leq \lambda \leq 1} \left[\log \sum_{x_1, \tilde{u}_1} p_0(x_1, \tilde{u}_1) e^{\lambda(x_1, \tilde{u}_1) \log \frac{p_1(x_1, \tilde{u}_1)}{p_0(x_1, \tilde{u}_1)}} \right]$$

$$\overset{(s)}{\leq} \sum_{x_1, \tilde{u}_1} p_0(x_1, \tilde{u}_1) \log \frac{p_0(x_1, \tilde{u}_1)}{p_1(x_1, \tilde{u}_1)} = D(p_0|p_1),$$

as given in Section 11.9 of [33], where step (s) is due to Jensen's inequality. Note that because the situation is symmetric, we also have $\mathcal{C}(p_0, p_1) \leq D(p_1|p_0)$. Thus, we have the *Kullback-Leibler (KL) distance* bound

$$\mathcal{C}(p_0, p_1) \leq \min \big(D(p_0|p_1), D(p_1|p_0) \big). \tag{6.7}$$

Recalling that $D(p_0|p_1)$ is the best possible error exponent for the Neyman-Pearson test (Chernoff-Stein Lemma, Section 11.8 of [33]), the above inequality shows

that sacrificing an arbitrarily small false alarm detection performance can improve (asymptotic) missed detection performance.

The KL bound (6.7) can be obtained for more general h and u_1 (again with the help of Jensen's inequality) as follows. Let h, u_1 have the same alphabet of size M, and let $c_{u_1,h} = 1 - \delta_{u_1,h}$. Then

$$
\begin{aligned}
\mathcal{C}(p_0, p_1, \cdots, p_{M-1}) \;&=\; -\min_{\lambda} \left[\log \sum_{x_1, \tilde{u}_1} e^{\sum_{u_1} \lambda_{u_1}(x_1, \tilde{u}_1) \log \left(\max_h \; c_{u_1,h} \; p_h(x_1, \tilde{u}_1) \right)} \right] \\
&\leq\; -\min_{\lambda} \left[\log \sum_{x_1, \tilde{u}_1} \max_h \; e^{\sum_{u_1} \lambda_{u_1}(x_1, \tilde{u}_1) \log \left(c_{u_1,h} \; p_h(x_1, \tilde{u}_1) \right)} \right] \\
&\leq\; (M-1) \sum_{x_1, \tilde{u}_1} \min_{h \neq 0} \; p_0(x_1, \tilde{u}_1) \log \left(\frac{p_0(x_1, \tilde{u}_1)}{p_h(x_1, \tilde{u}_1)} \right).
\end{aligned}
$$

Since the above inequality holds if 0 on the right hand side is replaced by any value of h, we get

$$
\mathcal{C}(p_0, p_1, \cdots, p_{M-1}) \leq (M-1) \min_h \sum_{x_1, \tilde{u}_1} \min_{h' \neq h} \; p_h(x_1, \tilde{u}_1) \log \left(\frac{p_h(x_1, \tilde{u}_1)}{p_{h'}(x_1, \tilde{u}_1)} \right).
$$

Online sensor comparison and asymptotically optimal network patterns

Since Chernoff information does not depend on the prior, we may use it to describe (asymptotically) optimal network patterns. Here, "*optimal*" will mean "*asymptotically optimal*". Note that by definition, every sensor is a *local fusion center* (LFC). On the other other hand, a sensor may or may not be a *global fusion center* (GFC), which is defined as follows.

Consider a set of sensors $X^n = \{X_1, ..., X_n\}$. We may rank sensors according to quality such that X_k is better than $X_{k'}$ if

$$
\mathcal{C}^{(X_k)}(p) \geq \mathcal{C}^{(X_{k'})}(p),
$$

where $\mathcal{C}^{(X_k)}(p)$ denotes $\mathcal{C}^{(X_k)}(p_0, p_1, ..., p_{M-1})$, i.e., the Chernoff information of X_k. Consider a *network pattern* $G = (V_G, A_G)$ over X^n, where $V_G = X^n$ is the set of vertices and $A_G = \{\rho_{kk'} : k, k' = 1, ..., n\}$ is the set of arrows: there is an arrow $X_k \rightarrow X_{k'}$ if and only if $\rho_{kk'} = 1$, and $\rho_{kk'} = 0$ otherwise. Let $\mathcal{C}_G^{(X)}(p)$ denote the chernoff information of a sensor X in G.

X is a *global fusion center* in G, written $G \leq X$, if

$$\mathcal{C}_G^{(X)}(p) = \max_{X' \in G} \mathcal{C}_G^{(X')}(p).$$

That is, a GFC is any sensor with maximal Chernoff information.

Let $\mathcal{G}$ be a set of network patterns over X^n. For example, we may consider $\mathcal{G}$ to be the set of all network patterns over X^n with at most L edges, i.e.,

$$\mathcal{G} = \mathcal{G}(X^n, L) = \{G = (X^n, A_G) : |A_G| \leq L\}.$$

. Let $\mathcal{G}_X = \{G \in \mathcal{G} : G \leq X\}$ be the set of network patterns in each of which X is a GFC. Then we have the following problems:

1. Find an optimal network pattern in $\mathcal{G}$ with X as a GFC.

2. Find an optimal fusion center with respect to $\mathcal{G}$.

For problem 1, G_X is an optimal pattern with X as a GFC if

$$\mathcal{C}_{G_X}^{(X)}(p) = \max_{G \in \mathcal{G}_X} \mathcal{C}_G^{(X)}(p) = \max_{G \in \mathcal{G}: \, G \leq X} \mathcal{C}_G^{(X)}(p).$$

For problem 2, X is an *optimal fusion center* (with respect to $\mathcal{G}$) if

$$\mathcal{C}_{G_X}^{(X)}(p) = \max_{Y \in X^n} \mathcal{C}_{G_Y}^{(Y)}(p) = \max_{Y \in X^n} \max_{G \in \mathcal{G}_Y} \mathcal{C}_G^{(Y)}(p).$$

We note that the optimality defined here is robust with respect to the Bayesian cost structure and prior probability, since the Chernoff information depends neither on cost structure nor on prior probability.

Conclusion

We have studied distributed detection over sensor networks in the form of acyclic directed graphs. It was found that the decision rules for such networks are not more complicated than those for simple networks, provided we assume that each sensor sends the same message to all sensors receiving from it. This is still true regardless of whether sensor observations are conditionally independent or not. Information measures associated with large sample analysis of error probability were used to define sensor comparison and to define asymptotic optimality of network patterns.

Chapter 7

Conclusion

7.1 Main results and application

Based on familiar notions of optimization and statistics (Appendices A,B), we have developed a detection network optimization technique (Chapter 3) that can be applied in a variety of distributed detection systems. The obtained decision procedure provides necessary and sufficient conditions for, i.e., a complete characterization of, optimality in any decision optimization problem for which the underlying decision objective function is differentiable, monotonic, and convex in decision probabilities. This defines the scope of applicability of the result.

Our decision optimization procedure was applied in the following three distributed detection settings.

1. *Interactive distributed detection (Chapter 4)*: Under the Neyman-Pearson framework, we studied effects of a single round of interaction through exchange of 1-bit decisions between two sensors, both in the fixed sample case and in the large sample case. We observed that without any communica-

tion rate constraints, interaction improves performance of the fixed sample test but not its asymptotic performance. These results were generalized to cases involving multiple rounds of memoryless interaction, multiple sensors in parallel, and exchange of multibit decisions.

2. *Optimal fusion architecture (Chapter 5)*: In the Bayesian framework, we derived the optimal decision rule for the detection of a deterministic or Gaussian signal in Gaussian noise by a two-sensor tandem fusion network. We found that for low SNR, the sensor with higher SNR should serve as the fusion center.

3. *Detection by acyclic graph networks (Chapter 6)*: We showed that in a sensor network in the form of an acyclic directed graph, if each sensor transmits the same message to all of its off-springs in the network, then the optimal decision rules for such a network are similar to those of simple tandem and parallel networks. This is still true when the sensors communicate through non-ideal channels. In a brief study of large sample asymptotic analysis we derived Chernoff and Kullback information measures and used these measures to define a scheme for comparison among sensors and among sensor network patterns. This type of comparison scheme can be used, in particular, to determine (asymptotically) optimal sensor distributions within a given class of sensor network patterns.

7.2 Future research

Here we present a number of problems indicating possible directions for future research. These problems come from the major parts of this thesis, and they

include natural extensions of our optimal signal detection procedure as well as additional problems arising from the three main applications we have studied.

Based directly on our decision procedure of Section 3.3, the following are a number of possible research directions.

1. *Sequential detection*: Using our decision optimization method, we would like to describe both centralized and distributed sequential detection from scratch. We expect our method to provide a relatively simple description of the sequential detection problem, especially in the distributed setting.

2. *Deeper study of randomization of decision rules and dependence of detection performance on randomization parameters*: In particular, when randomization does make a difference, we would like to be able to select the best possible (i.e., an optimal) randomized decision rule in an automatic way using our decision procedure.

3. *Optimization of non-differentiable convex decision functions*: For simplicity, we assumed differentiability of the decision function in our analysis. However, we expect that the same decision process should still apply, with minimal adjustments, when the objective function is non-differentiable. In particular, the (partial) derivative of the decision function should be replaced by its subdifferential.

4. *Extension to harmonic decision functions*: Given that some key properties of convex functions which made our analysis possible are possessed by all subharmonic functions, we expect that our optimization procedure for monotonic convex decision functions can be readily extended to a similar optimization procedure for monotonic subharmonic decision functions.

5. *Relation to the optimization of submodular set functions*: We would like to explore connections between our optimization procedure and the optimization of submodular set functions used for the determination of optimal sensor placement within a given network pattern.

Each of our three main applications also gave rise to a research problem as follows.

6. *Interactive distributed detection*: What is the cost incurred by additional rounds of interaction? The main point here is that although interaction can strictly improve performance of the fixed sample test, the additional communication steps involve can be costly. A natural way to account for this would be to modify the original objective function, say by including an additional term in it.

7. *Optimal Fusion architecture*: What is the optimal communication direction for large SNR? What is needed here is simply an efficient computational algorithm, since the optimal decision rule is already available.

8. *Acyclic graph detection*: What is the optimal sensor distribution that achieves uniform reliability (asymptotically) in a sensor network? Here, uniform reliability refers to the situation where the value of the local Chernoff information is the same at each sensor.

7.3 Comments on sequential detection

At first sight, dynamic programming (DP) methods appear to be the ideal choice for sequential detection problems. This is mainly because DP itself is sequential

in nature. The DP approach is an *inductive* approach which, for more practical reasons, typically dictates that optimization should be done incrementally after each observation sample. Unfortunately, however, DP is suboptimal in general. Moreover, despite its built in sequential nature, DP does not necessarily provide the most convenient description of the sequential detection problem.

It is theoretically more convenient to consider a *deductive* approach in which we assume that an exhaustive collection of *sequential detection plans or strategies*, each consisting of a (possibly infinite) sequence of observation samples along with a corresponding randomly generated sequence of decisions, are already available. In such an approach, it is necessary to *explicitly* specify additional constraints ensuring that at any given step, processing continues to the next step if and only if the decision outputs of all previous steps each failed to meet the stopping criterion. (Note that such stopping constraints are *implicit* in an inductive approach such as DP). One then proceeds to select the sequential detection plan that optimizes the underlying objective function of the sequential test.

Note however that the inductive approach is preferred over the deductive approach when knowledge of available sequential detection strategies is severely limited. When knowledge of available sequential detection strategies is unlimited, then with respect to optimality, the deductive approach is always preferred over the inductive approach.

Based on the above discussion, we conclude that our optimization technique developed in Chapter 3 can provide a more convenient description of the sequential detection problem, as compared with a dynamic programming approach. To apply our method in sequential detection, we need to first specify a decision function, a natural choice of which is some measure of stopping time as a function of the probabilities of the decision sequences. Next we specify stopping constraints,

which should greatly reduce the number of nontrivial optimization variables (i.e., probabilities of the decision sequences). Finally, we optimize the decision function by selecting an optimal decision strategy (i.e., a decision sequence or strategy whose probability optimizes the decision function). If the decision function is a monotonic convex function of probabilities of the decision sequences, then either Proposition 3.1 or Corollary 3.2 (depending on whether we are maximizing the decision function or minimizing it) will provide a complete characterization of the solution of the optimization problem.

Part III

Concepts

Appendix A

Optimization of Convex Functions

A.1   The optimization problem

We will briefly discuss optimization problems in general. Our main focus, however, will be on a class of problems called convex problems. For a fixed positive integer d, a real-valued function f on the d-dimensional real vector space $\mathbb{R}^d = \{x = (x_1, x_2, ..., x_d) : x_i \in \mathbb{R}\}$ is a mapping expressed as

$$f : x \in D \subset \mathbb{R}^d \mapsto f(x) \in \mathbb{R},$$

where the domain D is not necessarily all of $\mathbb{R}^d$. For the purpose of optimization however, it is convenient to allow functions to take infinite values, in which case, we simply present every function in the form

$$S : \mathbb{R}^d \to \bar{\mathbb{R}} = \mathbb{R} \cup \{\pm\infty\}, \ x \mapsto S(x),$$

where the *natural domain* of S is separately defined as

$$\operatorname{dom} S = \{x \in \mathbb{R}^d : S(x) \in \mathbb{R}\}.$$

A standard reference for the material in this chapter is [47].

The most basic optimization problem for S can be presented in the form

$$
\begin{aligned}
&\text{optimize} \quad S(x) \\
&\text{subject to} \quad x \in C
\end{aligned}
\tag{A.1}
$$

where C is a subset of $\mathbb{R}^d$ called the *constraint set* of the problem, and S is called the *objective function* of the problem.

In the basic optimization problem (A.1), our *objective* is either to *minimize* (i.e., find the smallest value of) or to *maximize* (i.e., find the largest value of) the function S. However, every maximization problem can be rewritten as a minimization problem, and likewise, every minimization problem is a maximization problem. Consequently, without loss of generality, we will *temporarily* assume for convenience that every optimization problem is in the form

$$
\begin{aligned}
&\text{minimize} \quad S(x) \\
&\text{subject to} \quad x \in C
\end{aligned}
\tag{A.2}
$$

The optimal value of S will be denoted by S_{opt}, and we will write

$$
S_{\text{opt}} = \min_{x \in C} S(x).
$$

We say a point $y \in \mathbb{R}^d$ is an *optimum* (or an *optimal point*) of S if $S(y) = S_{\text{opt}}$, and we write

$$
y \in \operatorname*{argmin}_{x \in C} S(x) = \left\{ z \in \mathbb{R}^d : S(z) = S_{\text{opt}} \right\},
$$

where the set $\operatorname*{argmin}_{x \in C} S(x)$ is called the *solution set* of the problem.

If the restriction $S|_C : C \to \mathbb{R}$ is a *convex function* (Definition A.2), then the problem is called a *convex problem*.

If the constraint set C is not specified in the problem (A.1), then we assume $C = \mathbb{R}^d$, and refer to the problem as *unconstrained*. Otherwise it is a *constrained*

optimization problem. Most practical optimization problems are constrained in nature, and it is often possible to simplify the constraints by adjusting (or redefining) the objective function in some way. Some of these adjustment techniques are discussed next in Section A.2.

A.2 Constrained optimization

Recall that the basic problem (A.2) is constrained if $C \subsetneq \mathbb{R}^d$, i.e., if C is a proper subset of $\mathbb{R}^d$. It is often possible to solve a complex optimization problem by solving a number of simpler optimization problems. However, such a possibility is difficult to uncover or identify when the geometric structure of the constraint set C is sufficiently intricate. By trading the geometric complexity of C for a relatively trivial algebraic refinement of the function S, the problem can become a lot easier to solve.

When the set C is specified in terms of equality or inequality constraints, and the function S satisfies some regularity conditions (e.g., differentiability), then the problem can be rewritten as an equivalent problem

$$\begin{aligned} &\text{minimize} \quad L(x, \lambda) \\ &\text{subject to} \quad (x, \lambda) \in \tilde{C} \end{aligned} \tag{A.3}$$

where the new objective function $L : \mathbb{R}^d \times \mathbb{R}^{\tilde{d}} \to \mathbb{R}, \; (x, \lambda) \mapsto L(x, \lambda)$ depends on the original objective function S, and the new constraint set $\tilde{C}$ is geometrically simpler than the original constraint set C. The function L is called a *Lagrangian* function of the problem. The new optimization variables $\lambda = (\lambda_1, \lambda_2, ..., \lambda_{\tilde{d}})$ are called *Lagrange multipliers*.

We will now make the above discussion more explicit.

Equality constraints and the Lagrangian

Consider an optimization problem with equality constraints:

$$\begin{aligned} \underset{x \in \mathbb{R}^d}{\text{minimize}} \quad & S(x), \\ \text{subject to} \quad & h_i(x) = 0, \quad i = 1, ..., n. \end{aligned} \tag{A.4}$$

Let $C = \{x \in \mathbb{R}^d : h_i(x) = 0, \quad i = 1, ..., n\}$ denote the constraint set as before, and let $\gamma : [0, 1] \to C, \ t \mapsto \gamma(t)$ be any smooth curve in C. For simplicity, we will further make the following assumptions.

1. S and h_i are twice differentiable.

2. S has local minima in C, which we wish to find.

Then the constraints imply that

$$0 = \frac{d}{dt} h_i(\gamma(t)) = \gamma'(t)^T \cdot \nabla h_i(\gamma(t)), \quad i = 1, ..., n,$$

i.e., at the optimum, the hyperplane spanned by the gradients $\{\nabla h_i : i = 1, ..., n\}$ is orthogonal to C. Moreover, because this holds for all γ, the vectors $\{\nabla h_i : i = 1, ..., n\}$ span the *orthogonal complement* of the *tangent space* (i.e., the space of all vectors that are tangent or "parallel") to C at the optimum.

Also, recall that at a local minimum, we have

$$0 = \frac{d}{dt} S(\gamma(t)) = \gamma'(t)^T \cdot \nabla S(\gamma(t)),$$
$$0 \leq \frac{d^2}{dt^2} S(\gamma(t)) = \gamma'(t)^T \cdot \nabla^2 S(\gamma(t)) \cdot \gamma'(t). \tag{A.5}$$

The first of these relations says that at the optimum, ∇S is orthogonal to C. Since the orthogonal complement to C at the optimum is spanned by the gradients

$\{\nabla h_i : i = 1, ..., n\}$, it follows that at the optimum the vector ∇S must lie in the hyperplane spanned by the gradients $\{\nabla h_i : i = 1, ..., n\}$, so that

$$\nabla S(x) + \sum_{i=1}^{n} \lambda_i \nabla h_i(x) = 0, \quad \text{for some} \quad \lambda_i \in \mathbb{R}.$$

The optimization problem (A.4) can now be restated as

$$\underset{(x,\lambda)\in\mathbb{R}^{d+n}}{\text{minimize}} \quad L(x, \lambda) = S(x) + \sum_{i=1}^{n} \lambda_i h_i(x), \tag{A.6}$$

The optimality conditions (for a local minimum) are given by

$$\nabla_{x,\lambda} L(x, \lambda) = 0, \quad \text{(d+n equations)}$$

$$\nabla_{x,\lambda}^2 L(x, \lambda) \succeq 0,$$

or equivalently, by

$$\nabla_x L(x, \lambda) = \nabla S(x) + \sum_{i=1}^{n} \lambda_i \nabla h_i(x) = 0,$$

$$\nabla_{\lambda_i} L(x, \lambda) = h_i(x) = 0, \quad i = 1, ..., n,$$

$$\nabla_{x,\lambda}^2 L(x, \lambda) = \begin{bmatrix} \nabla_x^2 S(x) & 0 \\ 0 & 0 \end{bmatrix} \succeq 0,$$

where $\succeq$ denotes positive definiteness over the constraint set C as implied by the relation (A.5) which holds for every smooth curve γ in C that passes through the optimum.

These optimality conditions show that the problems (A.4) and (A.6) are equivalent for the objective of finding local minima of S.

Inequality constraints and the KKT Lagrangian

Consider a problem with inequality constraints:

$$\begin{aligned}
&\underset{x}{\text{minimize}} && S(x) \\
&\text{subject to} && \begin{matrix} f_i(x) \leq 0, & i=1,...,m \\ h_j(x)=0, & j=1,...,n \end{matrix}
\end{aligned} \tag{A.7}$$

We again assume for simplicity that S, f_i, h_j are twice differentiable, and that S has local minima in the constraint set, $C = \{x \in \mathbb{R}^d : f_i(x) \leq 0, \ h_j(x) = 0, \ i = 1,...,m, \ j = 1,...,n\}$, that we wish to find.

The inequality constraints $f_i(x) \leq 0$ hold if and only if

$$f_i(x) + s_i^2 = 0, \quad \text{for some} \quad s_i \in \mathbb{R}, \tag{A.8}$$

where s_i are known as *slack variables* and their actual values need to be optimal. Therefore the problem becomes

$$\begin{aligned}
&\underset{x,\{s_i\}}{\text{minimize}} && S(x) \\
&\text{subject to} && \begin{matrix} f_i(x)+s_i^2=0, & i=1,...,m \\ h_j(x)=0, & j=1,...,n \end{matrix}
\end{aligned} \tag{A.9}$$

As before, we can write down a Lagrangian

$$L(x, s, \lambda, \nu) = S(x) + \sum_{i=1}^{m} \lambda_i \big(f_i(x) + s_i^2\big) + \sum_{j=1}^{n} \nu_j h_j(x),$$

in terms of which the optimization problem (A.7) becomes

$$\underset{x,s,\lambda,\nu}{\text{minimize}} \quad L(x, s, \lambda, \nu).$$

The optimality conditions (for a local minimum) are given by

$$\nabla_{x,s,\lambda,\nu} L(x, s, \lambda, \nu) = 0,$$

$$\nabla^2_{x,s,\lambda,\nu} L(x, s, \lambda, \nu) \succeq 0,$$

which are equivalent (after $\{s_i\}$ has been completely eliminated) to

$$\nabla S + \sum_{i=1}^{m} \lambda_i \nabla f_i(x) + \sum_{j=1}^{n} \nu_j \nabla h_j(x) = 0,$$

$$f_i(x) \leq 0, \quad i = 1, ..., m,$$

$$\lambda_i \geq 0 \quad i = 1, ..., m,$$

$$\lambda_i f_i(x) = 0, \quad i = 1, ..., m,$$

$$h_j(x) = 0, \quad j = 1, ..., n,$$

$$\nu_j \in \mathbb{R} \quad i = 1, ..., n.$$

The above relations, called *KKT conditions*, show that the original problem (A.7) is equivalent to the problem

$$\operatorname*{minimize}_{x, \{\lambda_i\}, \{\nu_j\}} \quad L(x, \lambda, \nu) = S(x) + \sum_{i=1}^{m} \lambda_i f_i(x) + \sum_{j=1}^{n} \nu_j h_j(x)$$

$$\text{subject to} \quad x \in \mathbb{R}^d, \quad \lambda \in [0, \infty)^m, \quad \nu \in \mathbb{R}^n \tag{A.10}$$

A.3 Convex functions

Many problems that arise in practice are convex. Convex functions possess nice properties which make their optimization relatively easy to handle computationally. We will present some basic properties of convex functions in this section. The optimization of convex functions is considered in Section A.4.

The discussion in this section pays special attention to the following points:

1. The description of a convex set in terms of line segments through the set, and basic operations that preserve set convexity.

2. The behavior of a convex function along line segments through its domain, and basic operations that preserve function convexity.

These points provide a way of understanding maxima and minima of convex functions in terms of two-dimensional geometry. They are also useful for identifying those optimization problems that are convex, as well as constructing convex functions.

To simplify our discussion, we will denote the *oriented line segment* between two points $x, y \in \mathbb{R}^n$ by $[x, y]$. It is convenient to view $[x, y]$ as the image of the parametrization

$$l_{x,y} : [0, 1] \to \mathbb{R}^n, \ t \mapsto l_{x,y}(t) = (1 - t)x + ty. \tag{A.11}$$

Set convexity

Definition A.1 (Convex set). *A set $D \subset \mathbb{R}^n$ is convex if $[x, y] \subset D$ for any two points $x, y \in D$.*

The following are some operations that preserve set convexity, and they are not difficult to check using Definition A.1.

1. *Composition of operations that each preserve set convexity*: It is clear that if $f : U \subset \mathbb{R}^m \to V \subset \mathbb{R}^n$, $g : V \subset \mathbb{R}^n \to W \subset \mathbb{R}^k$ are mappings that each preserve set convexity, then their composition $g \circ f : U \subset \mathbb{R}^m \to W \subset \mathbb{R}^k$ also preserves set convexity.

2. *Set intersection*: If $A, B \subset \mathbb{R}^n$ are two convex sets, let $D = A \cap B$. Then for any $x, y \in D$, $[x, y] \subset A$ and $[x, y] \subset B$, and so $[x, y] \subset D$, i.e., the intersection of convex sets is a convex set.

3. *Affine transformation*: If $D \subset \mathbb{R}^n$ is convex and $f : \mathbb{R}^n \to \mathbb{R}^m, \ x \mapsto Ax + b$ is an affine function, then $f(D) \subset \mathbb{R}^m$ is convex. More precisely, we have the following.

Let $f : U \subset \mathbb{R}^m \to V \subset \mathbb{R}^n,\ x \mapsto Ax + b$, be a surjective affine function, where A is an $n \times m$ matrix with real entries and $b \in \mathbb{R}^n$. Observe that for $x, y \in U$, we have

$$f\big((1-t)(Ax+b) + t(Ay+b)\big) = Af\big((1-t)x+ty\big) + b, \quad \text{for all}\ \ t \in [0,1],$$

and so $f([x,y]) = [f(x), f(y)]$. Thus, if $[x,y] \subset U$, then $[f(x), f(y)] \subset V$. This shows that affine mappings preserve set convexity.

4. *Perspective transformation*: A map of the form $P : \mathbb{R}^{n+1} \to \mathbb{R}^n,\ (x,t) \mapsto x/t$ is called a perspective function. This function simply uses the last component of (x,t) to scale the rest and drops that last component, and thus preserves set convexity.

5. *Fractional linear transformation*: This is the composition of an affine transformation and a perspective transformation P. Let $g : \mathbb{R}^n \to \mathbb{R}^{m+1},\ x \mapsto (Ax + b, c^T x + d)$, where $A \in \mathbb{R}^{m \times n},\ c^T \in \mathbb{R}^{1 \times n},\ b, d \in \mathbb{R}$. Since $P : \mathbb{R}^{m+1} \to \mathbb{R}^m$, we have $\mathbb{R}^n \xrightarrow{g} \mathbb{R}^{m+1} \xrightarrow{P} \mathbb{R}^m$, i.e., we have the composition $P \circ g :\ \mathbb{R}^n \to \mathbb{R}^m$, which is given by

$$P \circ g(x) = P(Ax + b, c^T x + d) = (Ax+b)/(c^T x + d), \quad \text{for} \quad x \in \mathbb{R}^n.$$

Function convexity

Definition A.2 (Convex function). *A function $f : D \subset \mathbb{R}^n \to \mathbb{R}$ is convex if D is a convex set, and for all $x, y \in D$, we have*

$$f\big((1-t)x + ty\big) \leq (1-t)f(x) + tf(y), \quad \text{for all}\ \ t \in [0,1].$$

Remark. It follows immediately from Definition A.2 that a function is convex if and only if it is convex along every line segment through its domain. For this

reason, any characterization of convexity in one dimension may be readily extended to higher dimensions simply by considering it along every line segment through the function's domain.

The following are some operations which preserve function convexity. They are not difficult to check using Definition A.2, but some of them can be more conveniently visualized with the help of simple geometric pictures.

1. *Nonnegative weighted sum:* If $\{f_\alpha(x)\}_\alpha$ is a collection of convex functions, and $w_\alpha \geq 0$ for each α, then the function $\sum_\alpha w_\alpha f_\alpha(x)$ is convex.

2. *Composition with an affine mapping:* If $f : D \subset \mathbb{R}^n \to \mathbb{R}$ is convex, and $L : \mathbb{R}^m \to \mathbb{R}^n$, $x \mapsto Ax + b$, where $A \in \mathbb{R}^{n \times m}$, $b \in \mathbb{R}^n$, then the function

$$g = f \circ L : \ L^{-1}(D) \subset \mathbb{R}^m \to \mathbb{R}, \ x \mapsto f(Ax + b)$$

is convex.

3. *Pointwise supremum:* If $f_\alpha(x)$ is convex for each α then $\sup_\alpha f_\alpha(x)$ is convex over dom $\sup_\alpha f_\alpha = \bigcap_\alpha \text{dom} f_\alpha$. In particular, if $f(x, y)$ is convex in x for each y, then $\sup_{y \in \mathcal{D}} f(x, y)$ is convex for any set $\mathcal{D}$.

4. *Pointwise infimum:* If $f(x, y)$ is convex in (x, y) and C is a nonempty convex set, then $g(x) = \inf_{y \in C} f(x, y)$ is convex if $-\infty < g(x)$ for all x.

5. *Perspective of a function:* If $f : \mathbb{R}^n \to \mathbb{R}$ is convex then the function

$$g : \mathbb{R}^{n+1} \to \mathbb{R}, \ (x, t) \to tf(x/t), \ \ \text{dom} g = \{(x, t) : \ x/t \in \text{dom} f, \ t > 0\},$$

is convex.

6. *Composition of convex functions:* Let $h : \mathbb{R}^m \to \mathbb{R}$ and $g : \mathbb{R}^n \to \mathbb{R}^m$ be twice differentiable. Then $f = h \circ g : \mathbb{R}^n \to \mathbb{R}$ satisfies

$$\nabla_{x_i} \nabla_{x_j} f = \sum_{k=1}^{m} \nabla_{x_i} \nabla_{x_j} g_k(x) \ \nabla_{g_k} h(g) + \sum_{k,k'} \nabla_{x_i} g_k(x) \ \nabla_{g_k} \nabla_{g_{k'}} h(g) \ \nabla_{x_j} g_{k'}.$$

Therefore, if g is convex, and h is both convex and increasing in each of its arguments (or if g is concave, and h is both convex and decreasing in each of its arguments), then f is convex.

Based on the remark following Definition A.2, convexity of a differentiable function of several variables can be described in terms of the following result for a function of a single variable.

Theorem A.3.

(a) If $f : (a, b) \to \mathbb{R}$ is differentiable, then f is convex if and only if f' is monotonically increasing.

(b) If $f : (a, b) \to \mathbb{R}$ is twice differentiable, then f is convex if and only if $f''(x) \geq 0$ for all $x \in (a, b)$.

Proof.

(a) Assume f is differentiable on (a, b).

* ($\Rightarrow$) Let f be convex. Then for all $\lambda \in (0, 1)$ and $x, y \in (a, b)$,

$$f(\lambda x - (1 - \lambda)y) \leq \lambda f(x) + (1 - \lambda)f(y),$$
$$\Rightarrow \quad \frac{f(\lambda[x - y] + y) - f(y)}{\lambda} \leq f(x) - f(y).$$

By taking the limit $\lambda \to 0$, and by interchanging x and y, we obtain

$$f'(y)(x - y) \leq f(x) - f(y), \quad f'(x)(y - x) \leq f(y) - f(x). \quad \text{(A.12)}$$

If $x < y$, then (A.12) implies

$$f'(x) \leq \frac{f(x) - f(y)}{x - y} \leq f'(y),$$

and thus f' is monotonically increasing.

∗ ($\Leftarrow$) Conversely, let f' be monotonically increasing on (a, b). Let $x, y \in (a, b)$ such that $x < y$. For $\lambda \in (0, 1)$, let $z = \lambda x + (1 - \lambda)y$. Then $x < z < y$, and

$$f(\lambda x + (1 - \lambda)y) - [\lambda f(x) + (1 - \lambda)f(y)]$$
$$= f(z) - [\lambda f(x) + (1 - \lambda)f(y)]$$
$$= \lambda[f(z) - f(x)] + (1 - \lambda)[f(z) - f(y)]$$
$$\overset{\text{mvt}}{=} \lambda(z - x)f'(c) + (1 - \lambda)(z - y)f'(d), \quad (x < c < z < d < y)$$
$$= \lambda(1 - \lambda)(y - x)[f'(c) - f'(d)]$$
$$\leq 0,$$

where *mvt* denotes the mean value theorem. Hence f is convex.

(b) Assume f'' exists on (a, b).

∗ ($\Rightarrow$) Let f be convex. Then for $x, y \in (a, b)$, $x < y$,

$$f'(x) \leq f'(y) \quad \Rightarrow \quad f'(x) - f'(y) = (x - y)f''(c) \leq 0, \quad (x < c < y),$$
$$\Rightarrow \quad f''(c) \geq 0.$$

Since x, y were arbitrary, we have that $f''(t) \geq 0$ for all $t \in (a, b)$.

∗ ($\Leftarrow$) Conversely, let $f''(x) \geq 0$ for all $x \in (a, b)$. Then f' is monotonically increasing, and thus f is convex by part (a).

$\square$

By applying Theorem A.3 along every line segment in the function's domain, the following results are immediate.

Corollary A.4.

(a) *If $f : D \subset \mathbb{R}^n \to \mathbb{R}$ is differentiable, then f is convex if and only if*

$$(x - y)^T \cdot \nabla f(y) \leq f(x) - f(y), \quad \text{for all } x, y \in D.$$

(b) *If $f : D \subset \mathbb{R}^n \to \mathbb{R}$ is twice differentiable, then f is convex if and only if*

$$\nabla^2 f(x) \succeq 0, \quad \text{for all } x \in D,$$

i.e., if and only if the Hessian matrix $\nabla^2 f(x)$ is positive semi-definite for all $x \in D$.

In the optimization of convex functions, the following inequality is frequently useful.

Theorem A.5 (Jensen's Inequality). *Let $f : D \subset \mathbb{R}^n$ be integrable, and let $p : D \subset \mathbb{R}^n \to [0, \infty)$ be a probability mass/density function, i.e., $\sum_{x \in D} p(x) = 1$. Let $E_p[f] = \sum_{x \in D} f(x)p(x)$ denote the average of f with respect to p.*

If $S : (a, b) \supset f(D) \to \mathbb{R}$ is a differentiable convex function, then

$$S\big(E_p[f]\big) \leq E_p\big[S(f)\big].$$

Proof. By Corollary A.4,

$$(u - v)S'(v) \leq S(u) - S(v), \quad \text{for all } u, v \in (a, b). \tag{A.13}$$

To obtain the result, we set $u = f(x)$, $v = E_p[f]$, and take the expectation $E_p[\cdot]$ of both sides of the inequality (A.13). $\qquad\square$

Note that the conclusion of the theorem does not require S to be differentiable, i.e., differentiability was included for simplicity only. This is because the definition of convexity of a function implies *a convex function is differentiable almost everywhere in its domain.*

A.4   Maximization and minimization of a convex function

The goal in this section is to determine necessary and sufficient conditions for the maximum, and for the minimum, of a convex function. With respect to optimization, (differentiable) convex functions are nice because they fall in a class of functions whose maxima and minima on any domain (i.e., a simply connected open set) occur either on the boundary of the domain or at points where the derivative equals 0. Such functions are called *subharmonic* functions.

The important thing about subharmonic behavior is the following. Optimization problems involving a large (often infinite) number of optimization variables arise in detection theory. However, mere knowledge of the fact that the maxima and minima of the underlying objective function lie on the boundary of the function's domain (although a necessary condition only) greatly reduces the number of optimization variables. Frequently, the reduction in cardinality of the space of optimization variables is from *infinite* to *at most countable.* (See Theorem 3.1). Moreover, this necessary condition can sometimes directly yield the optimal solution if the objective function and constraints are sufficiently simple. A remark on this last point is given after Theorem A.6, where monotonicity is essential for necessary conditions to become sufficient.

If $f : D \subset \mathbb{R}^n \to R$ is a differentiable convex function, then it is easy to see that the necessary and sufficient condition for $y \in \mathbb{R}^n$ to be the minimum of f is given by

$$\frac{d}{dt} f\big(l_{x,y}(t)\big)\Big|_{t=1} \leq 0 \quad \text{for all } x \in D, \tag{A.14}$$

where $l_{x,y}(t) = l_{x,y}(t) = (1 - t)x + ty$ was defined in (A.11). This condition simply says that while we approach y along any line segment, the derivative of the function at y is either 0 or negative. The condition (A.14) is of course equivalent to

$$(x - y)^T \cdot \nabla f(y) \geq 0 \quad \text{for all } x \in D. \tag{A.15}$$

The necessary and sufficient conditions for the maximum of a convex function are a bit more involved because an additional property, which is *monotonicity* of the objective function over the constraint set, is required to establish sufficiency of basic necessary conditions. The main results are presented in Theorem A.6. Note that in this theorem, the derivative ∇f is denoted by f' for convenience.

Theorem A.6 (Convex maximization theorem). *Let $f : \mathbb{R}^n \to \bar{\mathbb{R}} = \mathbb{R} \cup \{\pm\infty\}$ be differentiable and convex in its natural domain, $\mathrm{dom} f = \{x \in \mathbb{R}^n : f(x) \in \mathbb{R}\}$. Let $D \subset \mathrm{dom} f$ be any convex set. Then for any point $z \in \overline{D} = D \cup \partial D$,*

$$f(z) = \max_{x \in D} f(x) \tag{A.16}$$

if and only if the following conditions hold.

(a) $(x - z) \cdot f'(z) \leq 0$ for all $x \in D$.

(b) $f(z) \geq f(y)$ for every point $y \in \mathbb{R}^n$ satisfying $(y - x) \cdot f'(y) \leq 0$ for all $x \in D$.

Proof. Let $z \in \overline{D}$.

- ($\Rightarrow$): Assume z satisfies (A.16). Let $y \in \overline{D}$. By the convexity of f in D, Corollary A.4 implies

$$(x - y) \cdot f'(y) \leq f(x) - f(y) \quad \text{for all} \ \ x \in D. \tag{A.17}$$

  Setting $y = z$ in (A.17), we see that "$f(x) - f(z) \leq 0$ for all $x \in D$" implies "$(x - z) \cdot f'(z) \leq 0$ for all $x \in D$", which verifies condition (a). Condition (b) is also trivially satisfied.

- ($\Leftarrow$): Assume z satisfies (a) and (b). Let $y \in \overline{D}$ be any point satisfying "$(y - x) \cdot f'(y) \leq 0$ for all $x \in D$". Then for each $x \in D$, the function

$$g_x(t) = f\left(l_{x,y}(t)\right) = f(x + (y - x)t), \quad 0 \leq t \leq 1,$$

  is nondecreasing at y, i.e., $g'_x(1) = f'(y) \cdot (y - x) \geq 0$ for all $x \in D$. This says that as we approach the point y along any line segment, the function cannot decrease. Thus y is a *relative local maximum*, since $f(y) \geq f(x)$ for all $x \in B_\varepsilon(y) \cap D$, where $B_\varepsilon(y)$ is a ball of some radius $\varepsilon > 0$ centered at y.

  Now, by (a), z is a relative local maximum of f on D and, by (b), $f(z) \geq f(y)$ for every relative local maximum, y, of f. This means z is a global maximum of f on D, and so z satisfies (A.16).

  $\square$

Remarks.

1. The condition given in (a) of the theorem is necessary but not sufficient for a global maximum as one can easily verify using a simple quadratic function on the real line. However, if for all $x \in D$ the derivative $g'_x(t) = (y - x) \cdot f'(x + (y - x)t)$ of

$$g_x(t) = f(x + (y - x)t), \quad 0 \leq t \leq 1,$$

has the same sign for all $0 \leq t \leq 1$, then condition (a) is necessary and sufficient for a global maximum. In other words, if f is *monotonic* in D, then (a) is a complete characterization for a maximum of f. In particular, if f is an affine function, then (a) is necessary and sufficient for a global maximum of f.

Monotonicity as described above is too strong. In the following remarks, we will see that the maximum is always a boundary point, and so monotonicity is only required with respect to one point of the boundary ∂D in order for (a) to be necessary and sufficient for a global maximum. I.e., if there is a point $w \in \partial D$ such that f' is monotonic along every line segment through w in $\overline{D}$, then (a) is both necessary and sufficient for a global maximum.

2. Note that the derivative f' does not have to be zero at a relative local maximum. Also, every global maximum is a relative local maximum.

3. Let $L : \mathbb{R}^n \to \mathbb{R}$, $y \mapsto \max_{x \in D \setminus \{y\}} (x - y) \cdot f'(y)$. Then the theorem says $z \in \overline{D}$ is a global maximum of f on D if and only if

$$f(z) = \max_{y \in \overline{D} : L(y) \leq 0} f(y).$$

4. Note that because f is convex, if $y \in \overline{D}$ satisfies $(x - y) \cdot f'(y) \leq 0$ for all $x \in D$, then y must be a boundary point of D. That is, every local maximum lies on the boundary ∂D of D. This can be seen geometrically by recalling that a function is convex if and only if it is convex along each line segment in its domain.

Therefore, $z \in \overline{D}$ is a global maximum of f on D if and only if $z \in \partial D$, and

$$f(z) = \max_{y \in \partial D : L(y) \leq 0} f(y).$$

5. Algorithms exist for solving the optimization problem in Theorem A.6. See [48] for example.

6. In typical problems that arise in detection theory with a huge number of optimization variables, the role of condition (a) is to cut down the space of optimization variables to an at most countable number of threshold variables. Condition (b) then guarantees that (direct) optimization over these threshold parameters will yield an optimal solution, provided the function is monotonic. The above two-step optimization procedure is explicitly carried out in Chapters 3 to 6.

Corollary A.7 (Convex minimization theorem). *A point $z \in \mathbb{R}^n$ is the* global minimum *of the function f given in Theorem A.6 if and only if it satisfies $(x - z) \cdot f'(z) \geq 0$ for all $x \in D$, which is the reverse inequality version of condition (a) of the theorem.*

Proof. This follows the same arguments as in the proof of Theorem A.6. Also see the discussion leading to the conditions (A.14) and (A.15). $\square$

Appendix B

Statistical Information Inference

The term *statistics* refers to a collection of conceptual methods for quantifying and processing experimental observation. Some of these methods include probability in Section B.1, random variables in Section B.2, point estimation in Section B.4, and hypothesis testing in Section B.5.

Given a relatively new physical system, one would like to be able to predict its behavior under certain desired operating conditions. Accordingly, one performs an *experiment* on the system by first subjecting it under specific (external or environmental) conditions, and then monitoring and recording the system's *basic behavioral responses* to the conditions. In a typical experiment, the above process may be repeated as many times as necessary. From a practical standpoint, it is observed that accuracy in predicting the system's behavior using experimental results increases with the number of repetitions.

The basic behavioral responses of the system noted above are called *outcomes* of the experiment. The set of all possible outcomes of the experiment is called the

A more complete discussion of the concepts in this chapter can be found in [49].

sample space of the experiment. The sample space will be denoted by S. Subsets of S are called *events* of the experiment.

B.1 Probability

For computational convenience, the experiment is often specified in the form (S, Σ, P), and called a *probability measure space*, where the entries are defined as follows.

- S is the sample space of the experiment as defined above.

- Σ is a nonempty collection of events (i.e., subsets of S) which is closed under complement and countable union, in the sense that Σ contains the complements and countable unions of its elements. Σ is called a *σ-algebra* (sigma algebra) over S.

- P is a real function of the form $P : \Sigma \rightarrow [0, 1]$, with the following defining properties.

  (i) $P(S) = 1$.

  (ii) $P(U) \leq P(V)$, whenever $U \subset V$.

  (iii) $P(U \cup V) = P(U) + P(V)$, whenever $U \cap V = \emptyset$.

  P is called a *probability measure* over S. Note that property (iii) can be extended to any countable collection of sets in Σ.

The probability $P(U)$ of an event $U \subset S$ is a measure of its *likelihood of occurrence* in the experiment. Since events do intersect (so that the "previous" occurrence of one affects the likelihood of "subsequent" occurrence of another), a

useful concept is that of *conditional probability*, where the probability of an event U given that another event V has already occurred is defined as

$$P(U|V) \triangleq \frac{P(U \cap V)}{P(V)}, \quad \text{or by} \quad P(U \cap V) = P(U|V)P(V).$$

If $\{U_i, \ i = 1, ..., n\} \subset \Sigma$ is a partition of the sample space S, then

$$P(V) = P\left(V \cap \bigcup_{i=1}^{n} U_i\right) = \sum_{i=1}^{n} P(V \cap U_i) = \sum_{i=1}^{n} P(V|U_i)P(U_i)$$

$$\Rightarrow \quad P(U_i|V) = \frac{P(U_i \cap V)}{P(V)} = \frac{P(V|U_i)P(U_i)}{\sum_{j=1}^{n} P(V|U_j)P(U_j)}, \tag{B.1}$$

where the relation (B.1) is known as *Bayes rule*.

B.2   Random variables

Random variables are functions on sample spaces. More precisely, let (S, Σ, P) be the probability measure space representing an experiment. Then any function $X : S \to \mathcal{X}$ is called a *random variable*, where $\mathcal{X}$ is a *vector space*. Note that the probability measure P is seen as summarizing all possible results of the experiment, meanwhile a random variable X is seen as isolating a particular aspect or realization of the experiment.

It is not difficult to observe that every random variable X gives rise to a measure space $(\mathcal{X}, \Sigma_X, P_X)$, where Σ_X is a σ-algebra over $\mathcal{X}$ such that $X^{-1}(A) \in \Sigma$ for all $A \in \Sigma_X$, and the function

$$P_X : \Sigma_X \to [0, 1], \ A \mapsto P(X^{-1}(A))$$

is called the *probability distribution* of X. Therefore, apart from isolating a certain aspect of the experiment, a random variable also summarizes the results of the

experiment through its probability distribution. Note that $P_X(A) = P(X^{-1}(A))$ is often simply written as $P(X \in A)$, or as $P(X = x)$ if $A = \{x\}$ consists of a single point $x \in \mathcal{X}$.

Given $s \in S$, let $x = X(s) \in \mathcal{X}$, and let $U_x = X^{-1}(x) = X^{-1}(X(s))$. We say $X = x$ (in a random manner) with probability

$$P(X = x) = P_X(\{x\}) = P(X^{-1}(x)) = P(U_x).$$

In other words, X can take on any value $x \in X(S)$ but with a certain degree of uncertainty determined by the probability function P. The *expected value* of a random variable X is defined as

$$E_P[X] = \sum_{x \in \mathcal{X}} x \, p_X(x),$$

where the function $p_X : \mathcal{X} \to [0, \infty)$, defined such that

$$P(X \in A) = \sum_{x \in A} p_X(x),$$

is called the *probability mass function* (pmf) of X if X is discrete, or *probability density function* (pdf) of X if X is continuous. Note that $\sum_{x \in A}$ denotes integration over $A \subset \mathcal{X}$ if X is continuous. The existence of the function $\frac{dP_X}{dx} \triangleq p_X$ is determined by the *Radon-Nikodym theorem*.

A function (or *transformation*) of a random variable is again a random variable, in the following sense. If $X : S \to \mathcal{X}$ is a random variable and $f : \mathcal{X} \to \mathcal{Y}$ is any function (or transformation), then the composition $Y = f \circ X : S \xrightarrow{X} \mathcal{X} \xrightarrow{f} \mathcal{Y}$, written simply as $Y = f(X)$, is a random variable. A collection of random variables $X^n = (X_1, ..., X_n)$, $X_i : S \to \mathcal{X}_i$, is again a random variable given by

$$X^n : S \to \mathcal{X}^n = \mathcal{X}_1 \times \cdots \times \mathcal{X}_n, \ s \mapsto X^n(s) = (X_1(s), ..., X_n(s)).$$

Verification of the above claims, based on the definitions, is straightforward.

Note that we can have a possibly continuous collection of random variables, an example of which is the following.

Definition B.1 (Random process). *A random process $\{X(t) : t \in \mathbb{R}\}$ is a collection of random variables indexed by time. That is, for each value of t, $X(t)$ is a random variable.*

Basics of computation with random variables

In this section, for simplicity, we set $\mathcal{X} = \mathbb{R}$. Thus, a (univariate) random variable X is a mapping from the sample space to the reals:

$$X : S \to \mathbb{R}.$$

The measure space associated with X is $(\mathbb{R}, \Sigma_X, P_X)$, where the *probability distribution* $P_X : \Sigma_X \to [0, 1]$, $A \mapsto P_X(A)$ is given by

$$P_X(A) \equiv P(X \in A) \triangleq P\left(X^{-1}(A)\right) = P\big(\{s \in S : X(s) \in A\}\big).$$

A random variable is said to be *discrete* if its image is a discrete set in $\mathbb{R}$, or *continuous* if its image is a continuous set in $\mathbb{R}$. We note however that the description of a continuous random variable is similar to that of a discrete random variable, except that summation $\sum$ is replaced by integration $\int$.

For a discrete random variable X, the evaluation of P_X is often for convenience specified in terms of a *probability mass function* (pmf) f_X for X. Likewise, if X is continuous, the evaluation of P_X is specified in terms of a *probability density function* (pdf) f_X for X. The pmf or pdf is given by

$$P(X \in A) = \sum_{x \in A} f_X(x) \quad \text{or} \quad P(X \in A) = \int_A f_X(x)dx.$$

The *cumulative distribution function* (cdf) of a random variable X is defined as

$$F_X(x) \triangleq P(X \leq x) = \begin{cases} \sum_{x'=-\infty}^{x} f_X(x'), & \text{if } X \text{ is discrete} \\ \int_{-\infty}^{x} f_X(x')dx', & \text{if } X \text{ is continuous} \end{cases}$$
$$\Rightarrow \quad 0 \leq F_X(x) \leq 1.$$

Therefore,

$$f_X(x) = \begin{cases} F_X(x) - F_X(x-1), & \text{if } X \text{ is discrete,} \\ \frac{dF_X(x)}{dx}, & \text{if } X \text{ is continuous.} \end{cases}$$

The *expected value* (or *mean*) and *variance* of a function $g(X)$ of the random variable X are respectively defined by

$$\mu_{g(X)} = E[g(X)] = \begin{cases} \sum_{x=-\infty}^{\infty} g(x)f_X(x), & \text{if } X \text{ is discrete,} \\ \int_{-\infty}^{\infty} g(x)f_X(x)dx, & \text{if } X \text{ is continuous,} \end{cases}$$
$$\sigma_{g(X)}^2 = \text{Var}[g(X)] = E\left[\left(g(X) - E[g(X)]\right)^2\right].$$

Note that $Z = g(X) : S \to \mathbb{R}, \ s \mapsto Z(s) = (g \circ X)(s) = g(X(s))$ is itself a random variable with distribution function given by

$$P_Z(A) = P(Z \in A) = P(Z^{-1}(A)) = P(X^{-1}(g^{-1}(A))) = P_X\left(g^{-1}(A)\right),$$
$$F_Z(z) = P(Z \leq z) = P\left(g(X) \leq z\right) = P\left(X \in g^{-1}((-\infty, z])\right)$$
$$= \begin{cases} \sum_{x \in A_g(z)} f_X(x), & \text{if } X \text{ is discrete,} \\ \int_{A_g(z)} f_X(x)dx, & \text{if } X \text{ is continuous,} \end{cases}$$

where

$$A_g(z) = g^{-1}\left((-\infty, z]\right) = \left\{x \in \mathbb{R} : \ g(x) \in (-\infty, z]\right\}$$
$$= \left\{x \in \mathbb{R} : \ g(x) \leq z\right\}.$$

For simplicity, we assume that the random variables are *continuous* in what follows, while noting that the case of *discrete* as well as *mixed* random variables is much the same. Note that a *mixed random variable* is one whose range in $\mathbb{R}$ has both discrete and continuous subsets that are disjoint. Also, we will sometimes denote a pmf, or pdf, f_X by p_X.

Analogously to the univariate random variable, we define a *bivariate random variable* (X, Y), its *inherited probability distribution* $P_{X,Y}$, its pdf $f_{X,Y}$, and its cdf $F_{X,Y}$ as follows:

$$(X, Y) : S \to \mathbb{R}^2, \ s \mapsto (X(s), Y(s)), \qquad P_{X,Y} : \Sigma_{X,Y} \to [0, 1], \ A \mapsto P_{X,Y}(A),$$

where

$$P_{X,Y}(A) \equiv P((X, Y) \in A) \triangleq P\left((X, Y)^{-1}(A)\right) = P\big(\{s \in S : (X(s), Y(s)) \in A\}\big),$$

$$P((X, Y) \in A) = \int_A f_{X,Y}(x, y)dxdy,$$

$$F_{X,Y}(x, y) \triangleq P(X \le x, Y \le y) = \int_{-\infty}^{y} \int_{-\infty}^{x} f(x', y')dx'dy', \quad \Rightarrow \quad 0 \le F_{X,Y}(x, y) \le 1,$$

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

$P_{X,Y}$ is said to be the *joint probability distribution* for the pair of random variables (X, Y), while the *component random variables* X and Y are said to have *marginal distribution functions*

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x, y)dy = \frac{F_{X,Y}(x, \infty)}{dx} - \frac{F_{X,Y}(x, -\infty)}{dx},$$

$$f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x, y)dx = \frac{dF_{X,Y}(\infty, y)}{dy} - \frac{dF_{X,Y}(-\infty, y)}{dy}$$

associated with (i.e., due to) the joint distribution. The expected value and dis-

tribution of a new random variable $Z = g(X, Y)$ are given by

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

$$F_Z(z) = P(Z \leq z) = P\big(g(X, Y) \leq z\big) = \int_{A_g(z)} f_{X,Y}(x, y) dx dy,$$

where $A_g(z) = \{(x, y) \in \mathbb{R}^2 : g(x, y) \leq z\}$.

We can similarly proceed to describe *multivariate random variables* $X = (X_1, ..., X_n)$, where the *mean vector* M_X and *covariance matrix* Σ_X of X are defined as

$$M_X = E[X] = (E[X_1], ..., E[X_n]),$$

$$\Sigma_X = E\left[(X - E[X])(X - E[X])^T\right] = \left[E\big[(X_i - E[X_i])(X_j - E[X_j])\big]\right]_{ij}.$$

If X is a Gaussian-distributed real multivariate random variable, then a basic fact is that its pdf is completely determined by the pair (M_X, Σ_X) and given by

$$f_X(x) = \frac{e^{-\frac{(x-M_X)^T \Sigma_X^{-1}(x-M_X)}{2}}}{\sqrt{(2\pi)^n \det \Sigma_X}}.$$

It is also useful to note that if $\{X_i\}$ are jointly Gaussian-distributed, then so is any collection of variables $\{Y_k = \sum_i a_{ki} X_i\}$ that each depend linearly on the variables $\{X_i\}$.

Given any two random variables X and Y (each of which may be multivariate), it is often convenient to write

$$f_{X,Y}(x, y) = \frac{f_{X,Y}(x, y)}{f_Y(y)} f_Y(y) = f_{X|Y}(x|y) f_Y(y), \quad f_{X|Y}(x|y) = \frac{f_{X,Y}(x, y)}{f_Y(y)},$$

where $f_{X|Y}(x|y)$ is referred to as the *conditional pdf* of X given Y. Equivalently, in a shorthand notation where $X \sim f_X(x)$ means X is distributed according to the density function $f_X(x)$, we can also write

$$X|Y \sim f_{X|Y}(x|y).$$

Definition B.2 (Characteristic function, Moment generating function)**.** *The characteristic function of a random variable X is defined to be the function*

$$M_X(t) = E[e^{tX}] = \sum_{x \in \mathcal{X}} f_X(x) e^{tx},$$

for every complex number $t \in \mathbb{C}$ for which the expectation exists. The restriction of $M_X(t)$ to $t \in \mathbb{R}$ is called the moment generating function (mgf) of X.

Note that the characteristic function of a random variable can be used to determine its distribution.

Definition B.3 (Independence, Conditional independence, Identical distribution, iid sequence)**.**
Let $X^n = (X_1, X_2, ..., X_n)$, Y be random variables. We say the random variables $X_1, X_2, ..., X_n$ are independent, or that the sequence of random variables $X^n = (X_1, X_2, ..., X_n)$ is independent, if

$$p(x^n) = p(x_1)p(x_2) \cdots p(x_n).$$

Similarly, we say $X_1, X_2, ..., X_n$ are conditionally independent (or that X^n is conditionally independent) with respect to Y if

$$p(x^n|y) = p(x_1|y)p(x_2|y) \cdots p(x_n|y).$$

A sequence of random variables $X_1, X_2, \cdots$ is identically distributed if for all i, j, we have $p_{X_i} = p_{X_j}$, i.e., $p_{X_i}(x) = p_{X_j}(x)$ for all $x \in \mathcal{X}$. We say the sequence $X_1, X_2, \cdots$ is iid if it is both independent and identically distributed.

Definition B.4 (Convergence in distribution, Convergence almost surely, Convergence in probability)**.**

A sequence of random variables $X_1, X_2, \cdots$ converges in distribution to a random variable X if

$$\lim_{n \to \infty} F_{X_n}(x) = F_X(x)$$

whenever F_X is continuous at $x \in \mathcal{X}$.

The sequence $X_1, X_2, \cdots$ converges almost surely to X if for any $\varepsilon > 0$, we have

$$P\left(\lim |X_n - X| < \varepsilon\right) \triangleq P\left(\limsup E_n^\varepsilon\right) = 1,$$

where $E_n^\varepsilon = \{s \in S : |X_n(s) - X(s)| < \varepsilon\}$, and $\limsup E_n^\varepsilon = \cap_{m=1}^\infty \cup_{n=m}^\infty E_n^\varepsilon$.

The sequence $X_1, X_2, \cdots$ converges in probability to X if for any $\varepsilon > 0$, we have

$$\lim P(|X_n - X| < \varepsilon) \triangleq \lim P(E_n^\varepsilon) = 1.$$

Theorem B.5 (Central limit theorem). *Let $X_1, X_2, \ldots$ be a sequence of iid random variables. Then the random variable $Y_n = \frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma}$, where $\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i$, converges in distribution to the standard normal random variable, i.e.,*

$$\lim_{n \to \infty} f_{Y_n}(x) = N(0,1)(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}, \quad x \in \mathbb{R}.$$

Alternatively, for sufficiently large n we have

$$f_{\bar{X}_n}(x) \approx N(\mu, \sigma^2/n)(x) = \frac{1}{\sqrt{2\pi\sigma^2/n}} e^{-\frac{(x-\mu)^2}{2\sigma^2/n}}. \tag{B.2}$$

Proof. For simplicity, assume the mgf of the X_i's exists near $t = 0$. Then the mgf

of $Y_n = \frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{X_i - \mu}{\sigma}$ is given by

$$M_{Y_n}(t) = M_{\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{X_i - \mu}{\sigma}}(t) = M_{\sum_{i=1}^{n} \frac{X_i - \mu}{\sigma}}\left(\frac{t}{\sqrt{n}}\right) = \left[M_{\frac{X - \mu}{\sigma}}\left(\frac{t}{\sqrt{n}}\right)\right]^n$$

$$= \left[\sum_{k=0}^{\infty} \frac{1}{k!} M^{(k)}_{\frac{X - \mu}{\sigma}}(0) \frac{t^k}{n^{\frac{k}{2}}}\right]^n = \left[1 + 0 + \frac{t^2}{2n} + \sum_{k=3}^{\infty} \frac{1}{k!} M^{(k)}_{\frac{X - \mu}{\sigma}}(0) \frac{t^k}{n^{\frac{k}{2}}}\right]^n$$

$$\overset{n \to \infty}{\longrightarrow} e^{\frac{t^2}{2} + \lim_{n \to \infty} \sum_{k=3}^{\infty} \frac{1}{k!} M^{(k)}_{\frac{X - \mu}{\sigma}}(0) \frac{t^k}{n^{\frac{k-2}{2}}}} = e^{\frac{t^2}{2}},$$

where $\lim_n (1 + a_n)^n = e^{\lim_n n a_n}$, and the limiting mgf is that of the standard normal distribution $N(0, 1)$. $\qquad \square$

Lemma B.6 (Chebychev-Markov inequality). *Let X be a random variable, and let $g : \mathcal{X} \to (0, \infty)$ be an integrable function. Then*

$$P(g(X) \geq c) \leq \frac{1}{c} E[g(X)], \quad \text{for any } c > 0.$$

Proof.

$$P(g(X) \geq c) = \sum_{x \in \mathcal{X} : g(x) \geq c} f_X(x) \leq \frac{1}{c} \sum_{x \in \mathcal{X} : g(x) \geq c} g(x) f_X(x) \leq \frac{1}{c} \sum_{x \in \mathcal{X}} g(x) f_X(x).$$

$\qquad \square$

Theorem B.7 (Laws of large numbers). *Let $X_1, X_2, \ldots$ be iid random variables with $EX_i = \mu$, $VarX_i = \sigma^2 < \infty$, and $\bar{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i$. Then we have the following.*

1. *Strong law: $\bar{X}_n$ converges almost surely to μ.*

2. *Weak law: $\bar{X}_n$ converges in probability to μ.*

Proof. Observe that

$$P\left(\lim |\bar{X}_n - \mu| < \varepsilon\right) = 1 \quad \Longleftrightarrow \quad P\left(\lim |\bar{X}_n - \mu| \geq \varepsilon\right) = 0.$$

1. Let

$$A_n = \{s \in S : |\bar{X}_n(s) - \mu| \geq \varepsilon\} = \{s \in S : |[X_1(s) + \cdots + X_n(s)]/n - \mu| \geq \varepsilon\}$$

$$\simeq \{(x_1, ..., x_n, 0, 0, \cdots) \in \mathbb{R}^\infty : |(x_1 + \cdots + x_n)/n - \mu| \geq \varepsilon\},$$

where $\simeq$ denotes equivalence of sets with respect to cardinality, i.e., $A \simeq B$ if A and B have the same cardinality. Then

$$P\left(\lim |\bar{X}_n - \mu| \geq \varepsilon\right) = P\left(\limsup A_n\right) = P\left(\bigcap_{m=1}^\infty \bigcup_{n=m}^\infty A_n\right) \leq \inf_{m \geq 1} P\left(\bigcup_{n=m}^\infty A_n\right)$$

$$\leq \inf_{m \geq 1} \sum_{n=m}^\infty P(A_n) = \inf_{m \geq 1} \sum_{n=m}^\infty P(|\bar{X}_n - \mu| \geq \varepsilon).$$

Thus, it is sufficient to show that $\sum_{n=1}^\infty P(|\bar{X}_n - \mu| \geq \varepsilon) < \infty$. This finiteness easily follows from the central limit theorem, i.e., from the fact that $\bar{X}_n$ is distributed according to (B.2) for large n. Hence,

$$0 \leq P\left(\lim |\bar{X}_n - \mu| \geq \varepsilon\right) \leq \inf_{m \geq 1} \sum_{n=m}^\infty P(|\bar{X}_n - \mu| \geq \varepsilon) = 0.$$

2. Almost sure convergence implies convergence in probability for the following reason: Let $B_n = \bigcup_{k=n}^\infty A_k$, where A_k is as defined in part 1 above. Then $P(|\bar{X}_n - \mu| \geq \varepsilon) = P(A_n) \leq P(B_n)$, which implies

$$\lim P(|\bar{X}_n - \mu| \geq \varepsilon) = \lim P(A_n) \leq \lim P(B_n) = P(\limsup A_n)$$

$$= P(\lim |\bar{X}_n - \mu| \geq \varepsilon),$$

$$\Rightarrow \quad \lim P(|\bar{X}_n - \mu| \geq \varepsilon) \leq P(\lim |\bar{X}_n - \mu| \geq \varepsilon).$$

Alternatively, by the Chebychev-Markov inequality, we get

$$P(|\bar{X}_n - \mu| \geq \epsilon) = P((\bar{X}_n - \mu)^2 \geq \epsilon^2) \leq \frac{E(\bar{X}_n - \mu)^2}{\epsilon^2} = \frac{\sigma^2}{n\epsilon^2} \to 0 \quad \text{as} \quad n \to \infty.$$

$\square$

B.3 Statistical information

A function of several random variables is called a *statistic*. If $X^n = (X_1, ..., X_n)$ is a collection of random variables and $f : \mathcal{X}^n \to \mathcal{Y}$ is any function, then the random variable $Y = f \circ X^n : S \xrightarrow{X^n} \mathcal{X}^n \xrightarrow{f} \mathcal{Y}$, written $Y = f(X^n) = f(X_1, ..., X_n)$, is said to be a *statistic* based on X^n. In particular, X^n is a statistic based on X^n.

A basic property of every random variable is *uncertainty* or *entropy*, and is defined as a measure of the amount of *randomness* in the variable. Therefore, we can view the space $\mathcal{R} = \{X : S \to \mathcal{X}\}$ consisting of all random variables as a "*field of uncertainty*", with the random variables being the points of the space. *Information* is a measure of how much two variables in $\mathcal{R}$ are separated in randomness. That is, information is randomness distance, or distance with respect to randomness, between random variables in $\mathcal{R}$. Therefore, information is *relative uncertainty*, and one may of course loosely refer to the randomness of a random variable as the "*information content*" of the variable.

A basic example of an information measure $d_P : \mathcal{R} \times \mathcal{R} \to \mathbb{R}$ is given by

$$d_P(X, Y) = P(X \neq Y), \tag{B.3}$$

which is a familiar quantity known as *error probability* in a context where one of the variables is viewed as an *estimate* of the other. Other examples, called *distortions*, are given by

$$d_P(X, Y) = E_P[d(X, Y)], \tag{B.4}$$

where $d : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ is a *deterministic "distance"* function.

In general, *information metrics* are real-valued functions of statistics. Examples include asymptotic detection and estimation performance metrics such as

Shannon, Kullback, Chernoff, and Fisher information. Note that all of these metrics are special instances of the quantities in (B.3) and (B.4), which have certain convenient properties, including *additivity* over independent random variables in particular.

B.4 Estimation I: Point estimators and sufficiency

Consider a sequence of random variables $X^n = (X_1, ..., X_n) : S \to \mathcal{X}^n$, $s \mapsto (X_1(s), ..., X_n(s))$. Any value $x^n = (x_1, ..., x_n) \in \mathcal{X}^n$ of X^n is called a *data sample*, where n is the *sample size*. If the sequence x^n is generated according to the distribution of X^n, it is called a *random data sample*. Consequently, we may loosely refer to X^n itself as a "*random sample*". Because X^n summarizes the results of a *composite experiment* in the form of a *series of experiments*, each variable X_i in the random sample X^n is called an (experimental) *observation*.

Assume we have a system with a property θ that can take values in a set Θ, but we do not know its true (current) value. Then in order to determine the true value of θ, we further assume that we have conducted an experiment on the system and made observations $X^n = X_1, ..., X_n$. The observations are presumed to have been randomly (and independently) generated from the system according to a distribution $p_\theta(x) = p(x|\theta)$, written $X_i \sim p(x_i|\theta)$, so that $p(x^n|\theta) = \prod_{i=1}^{n} p(x_i|\theta)$, which is just another way of saying that the sample X^n summarizes the results of an experiment on the system (by means of its distribution $p(x^n|\theta)$). Consequently, X^n contains information about the true value of θ. Accordingly, we have the following definitions.

Definition B.8 (Point estimator)**.** *Any statistic $T(X^n)$ for the purpose of inferring the true value of θ is called a* point estimator *of θ.*

By the reasoning in (B.4), estimation performance of an estimator $T(X^n)$ can be investigated using a metric of the form

$$D = E_{p(x,\theta)}\big[d\big(T(X^n), \vartheta\big)\big], \tag{B.5}$$

where ϑ is θ viewed as a random variable.

Definition B.9 (Sufficient statistic). *$T(X^n)$ is a sufficient statistic for θ if X^n, as an estimator of θ, is no better than $T(X^n)$, i.e., if*

$$E_{p(x,\theta)}d\big(T(X^n), \vartheta\big) \leq \inf_{T'} \ E_{p(x,\theta)}d\big(T'(X^n), \vartheta\big), \tag{B.6}$$

where the infimum is taken over all possible estimators of θ based on X^n.

Proposition B.10. *The following are equivalent.*

1. $T(X^n)$ is a sufficient statistic for θ.

2. We have a Markov chain $\vartheta \to T(X^n) \to X^n$, i.e.,

$$p(\theta, x^n | T(x^n)) = p(\theta | T(x^n)) p(x^n | T(x^n)).$$

Equivalently,

$$p(\theta | x^n, T(x^n)) = p(\theta | T(x^n)).$$

3. The conditional distribution

$$h(x^n) = p_\theta(x^n | T(x^n)) \triangleq p(x^n | T(x^n), \theta)$$

is independent of θ.

4. *For any points $x^n, y^n, z^n \in \mathcal{X}^n$ satisfying the redundancy condition $T(x^n) = T(y^n) = T(z^n)$, the function*

$$h(x^n, y^n, z^n) = \frac{p_\theta(y^n|T(x^n))}{p_\theta(z^n|T(x^n))}$$

is independent of θ, where

$$p_\theta(y^n|T(x^n)) = p_\theta(X^n = y^n|T(X^n) = x^n) = p(X^n = y^n|T(X^n) = x^n, \theta)$$
$$= \frac{p_\theta(y^n, T(x^n))}{p_\theta(T(x^n))}.$$

5. *For all $x^n, y^n \in \mathcal{X}^n$,*

$$T(x^n) = T(y^n) \quad \Rightarrow \quad \frac{\partial}{\partial \theta} \frac{p(x^n|\theta)}{p(y^n|\theta)} = 0.$$

(Note that some continuity and differentiability are assumed in this case)

Proof. The equivalences $2 \iff 3 \iff 4 \iff 5$ are straightforward.

The main challenge is with $1 \iff 2$. For this case, we must choose the function d in (B.6) in such a way that the following conditions hold.

(a) An estimator $T_1(X^n)$ is closer in randomness to ϑ than another estimator $T_2(X^n)$, i.e.,

$$E_{p(x,\theta)}d\big(T_1(X^n), \vartheta\big) \leq E_{p(x,\theta)}d\big(T_2(X^n), \vartheta\big),$$

if and only if we have a Markov chain $\vartheta \to T_1(X^n) \to T_2(X^n)$.

(b) For every estimator $T'(X^n)$, we have a Markov chain $\vartheta \to X^n \to T'(X^n)$, i.e., X^n is the best possible estimator.

It then follows that a statistic $T(X^n)$ satisfies a Markov chain $\vartheta \to T(X^n) \to X^n$ [in addition to a Markov chain $\vartheta \to X^n \to T(X^n)$] if and only if it satisfies (B.6). $\qquad\square$

Theorem B.11 (Factorization theorem). *A statistic $T(X^n)$ is sufficient for θ if and only if there are functions α, β_θ (with α independent of θ) such that*

$$p(x^n|\theta) = \alpha(x^n)\,\beta_\theta\big(T(x^n)\big). \tag{B.7}$$

Proof.

- ($\Leftarrow$): If $p(x^n|\theta)$ satisfies (B.7), then it follow immediately from the definitions that $T(X^n)$ is a sufficient statistic for θ.

- ($\Rightarrow$): Assume $T(X^n)$ is a sufficient statistic for θ. Then for all $x^n, y^n \in \mathcal{X}^n$,

$$T(x^n) = T(y^n) \quad \Rightarrow \quad \frac{\partial}{\partial\theta}\frac{f(x^n|\theta)}{f(y^n|\theta)} = 0.$$

Since

$$\frac{\partial}{\partial\theta}\frac{p(x^n|\theta)}{p(y^n|\theta)} = 0 \quad \Longleftrightarrow \quad \frac{\partial_\theta p(x^n|\theta)}{p(x^n|\theta)} = \frac{\partial_\theta p(y^n|\theta)}{p(y^n|\theta)},$$

we have

$$T(x^n) = T(y^n) \quad \Rightarrow \quad \partial_\theta \ln p(x^n|\theta) = \partial_\theta \ln p(y^n|\theta), \quad \text{for all } x^n, y^n \in \mathcal{X}^n.$$

This means $\partial_\theta \ln p(x^n|\theta) = g(\theta, T(x^n)) \equiv g_\theta(T(x^n))$ for some function g_θ. Integration of this relation with respect to θ yields a formal solution of the form

$$p(x^n|\theta) = e^{\int^\theta d\theta'\,g_{\theta'}(T(x^n))+K(x^n)} \equiv \alpha(x^n)\,\beta_\theta(T(x^n)),$$

which is (B.7).

$\square$

Definition B.12 (Necessary statistic). *$T(X^n)$ is a necessary statistic for θ if for all $x^n, y^n \in \mathcal{X}^n$,*

$$\frac{\partial}{\partial \theta} \frac{p(x^n|\theta)}{p(y^n|\theta)} = 0 \quad \Rightarrow \quad T(x^n) = T(y^n).$$

Definition B.13 (Efficient statistic). *A statistic $T(X^n)$ is an efficient statistic for θ if it is both a necessary and a sufficient statistic for θ, i.e., if for all $x^n, y^n \in \mathcal{X}^n$,*

$$T(x^n) = T(y^n) \quad \Longleftrightarrow \quad \frac{\partial}{\partial \theta} \frac{p(x^n|\theta)}{p(y^n|\theta)} = 0.$$

Note that an efficient statistic is also called a minimal sufficient statistic.

Theorem B.14. *(Efficient statistic formula) An efficient statistic $T(X^n)$ has the form*

$$T(X^n) = h_\theta\big(\partial_\theta \log p(X^n|\theta)\big), \tag{B.8}$$

where h_θ is any invertible function which is at least capable of removing all of the θ dependence from $\partial_\theta \log p(X^n|\theta)$ as its argument.

Proof. Recall that $T(X^n)$ is an efficient statistic iff for all $x^n, y^n \in \mathcal{X}^n$,

$$T(x^n) = T(y^n) \quad \Longleftrightarrow \quad \frac{\partial}{\partial \theta} \frac{f(x^n|\theta)}{f(y^n|\theta)} = 0.$$

Since

$$\frac{\partial}{\partial \theta} \frac{p(x^n|\theta)}{p(y^n|\theta)} = 0 \quad \Longleftrightarrow \quad \frac{\partial_\theta p(x^n|\theta)}{p(x^n|\theta)} = \frac{\partial_\theta p(y^n|\theta)}{p(y^n|\theta)},$$

we have

$$T(x^n) = T(y^n) \quad \Longleftrightarrow \quad \partial_\theta \ln p(x^n|\theta) = \partial_\theta \ln p(y^n|\theta), \quad \text{for all} \ \ x^n, y^n \in \mathcal{X}^n.$$

This means $\partial_\theta \ln p(x^n|\theta) = g(\theta, T(x^n)) \equiv g_\theta(T(x^n))$, where g_θ is an invertible function such that the quantity $g_\theta^{-1}(\partial_\theta \log p(x^n|\theta))$ is independent of θ. Setting $h_\theta = g_\theta^{-1}$, we obtain the formula (B.8). $\square$

B.5 Estimation II: Set estimators and hypothesis testing

Recall that a point estimator is a statistic of the form $T(x^n) \in \Theta$, $x^n \in \mathcal{X}^n$. In general, estimators of the form $T(x^n) \subset \Theta$, $x^n \in \mathcal{X}^n$, are more practical. These are called *set estimators* (or *confidence sets*).

Definition B.15 (Set estimator). *Any statistic $T(X^n)$ for the purpose of inferring a reasonably small subset of Θ containing the true value of θ is called a* set estimator *of θ.*

With a point estimator, one reports the result of estimation based on $x^n \in \mathbb{R}^n$ by saying "given $\theta \in \Theta$, we have $\theta = T(x^n)$ with probability $P(T(X^n) = \theta)$". With a set estimator, one similarly reports the result of estimation based on $x^n \in \mathbb{R}^n$ by saying "given $\theta \in \Theta$, we have $\theta \in T(x^n)$ with probability $P(\theta \in T(X^n))$".

Note that a point estimator is a special case of a set estimator. This implies, in particular, that the notion of sufficiency discussed earlier for point estimators can, at least formally, be extended to set estimators. Moreover, concepts we will introduce for interval estimators apply to point estimators as well.

Definition B.16 (Degree of confidence, Percentage of confidence). *The degree of confidence (or confidence coefficient) of a set estimator $T(X^n)$ is $c = \min_{\theta \in \Theta} P(\theta \in T(X^n))$. The percentage of confidence of $T(X^n)$ is $100c\,\%$, and we say $T(X^n)$ is a $100c\,\%$ confidence set for θ.*

A method of statistical estimation that involves set estimators in a natural way is called *hypothesis testing*. Hypothesis testing, like point estimation, is a method of inference (of a parameter θ) based on observations. In the discussion

that follows, it is assumed we have observations $\{x^n \in \mathcal{X}^n\}$ based on at least one *known* family of probability distributions $\{p(x^n|\theta) : \theta \in \Theta\}$.

Definition B.17 (Hypothesis, Simple hypothesis). *A hypothesis, denoted by H, is a statement about the inference parameter θ (i.e., a parameter whose value we wish to infer), which is in the form of a constraint or restriction R_H on the value of θ. By convention, we write*

$$H : R_H,$$

which reads "H stands for, or represents, the value restriction R_H on θ".

A simple hypothesis, H, is a statement of the form

$$H : \theta = \theta_0,$$

for some fixed value $\theta_0 \in \Theta$.

We will say that two statements are *mutually exclusive* if they cannot be both valid simultaneously.

Definition B.18 (Hypothesis test). *Given a set of (mutually exclusive) hypotheses on θ, one and only one of which is valid, a hypothesis test is a method for deciding (based on observations $x^n \in \mathcal{X}^n$) the hypothesis that is most likely to be the valid one.*

Remarks.

I. Observe that by definition, the hypothesis test is determined by statistics which are real valued functions of the observations. Consequently, every hypothesis test can be specified as a solution of some optimization problem, as discussed in Appendix A. Moreover, in practice, the decision involved in

the test is of course made so as to meet a given objective, which is often the optimization (minimization or maximization) of some information measure.

II. Note that the above discussion indicates that the notion of an optimal hypothesis test (Definition B.19) is a natural generalization of the notion of a sufficient statistic (Definition B.9).

Definition B.19 (Objective hypothesis test, Optimal hypothesis test). *A hypothesis test is objective if it is specified as a solution of some optimization problem. An optimal hypothesis test is an objective hypothesis test for which the decision on the valid hypothesis is optimal with respect to the underlying objective of the test.*

The following is a preview of some basic points which are relevant in *statistical decision theory* (the subject of Chapter 2) and *optimal hypothesis testing* (the main subject of Chapter 3).

Remark. Although the eventual or end objective in a hypothesis test is to *decide* the valid hypothesis among a set of say M hypotheses, it is often more useful to consider intermediate decision operations that can take values in a set whose cardinality is different from M. This is important in distributed detection where some local sensors may only need to forward quantized versions of their observations to a fusion center which actually decides the true hypothesis based on the quantized observations. In general therefore, the intermediate decision output from a local sensor may not have the same alphabet as the hypothesis.

The above remark is also emphasized in Chapters 2 and 3.

Definition B.20 (Decision rule, Decision function). *A decision rule is a point estimator of the form*

$$\gamma : x^n \in \mathcal{X}^n \mapsto u = \gamma(x^n) \in \{0, 1, \cdots, N-1\}, \tag{B.9}$$

i.e., an estimator that takes on a discrete set of values.

In objective hypothesis testing, a decision function is an information measure that depends on both the decision rule and the hypothesis.

We can, for example, consider a decision function of the form

$$S = E_{p(u,h)} d(U, H), \tag{B.10}$$

where H denotes the hypothesis and u the decision. The primary objective of a hypothesis test is often to select a decision rule that optimizes an underlying decision function such as the function S in (B.10).

Definition B.21 (Binary hypothesis test, Null hypothesis, Alternative hypothesis). *A hypothesis test involving two complementary hypotheses is the simplest type of hypothesis test, and is called a binary hypothesis test. One of the hypotheses is denoted by $H_0 : \theta \in \Theta_0$ and called the* null hypothesis, *while the other is denoted by $H_1 : \theta \in \Theta_1$ and called the* alternative hypothesis, *where $\Theta_1 \cup \Theta_2 = \Theta$.*

Remarks (Computational Setup and Results)**.**

1. ***Indicator variables***: If we let $s = s(\theta) = I_{\Theta_1}(\theta) = \left\{ \begin{array}{ll} 0, & \theta \in \Theta_0 \\ 1, & \theta \in \Theta_1 \end{array} \right\}$, then the binary hypothesis test becomes a problem of estimating the value of the binary variable s. The hypotheses become $H_0 : s = 0$, $H_1 : s = 1$. The family of distributions $\{p(x^n|s) : s = 0, 1\}$ associated with s, i.e., the distribution of the observation conditioned on s, is given by

   $$p(x^n|s = i) = \frac{p(x^n, s = i)}{p(s = i)} = \frac{p(x^n, \theta \in \Theta_i)}{p(\theta \in \Theta_i)} = \frac{\sum_{\theta \in \Theta_i} p(x^n|\theta)p(\theta)}{\sum_{\theta \in \Theta_i} p(\theta)}, \quad i = 0, 1,$$

   where $p(\theta)$ is a *prior probability* distribution on Θ. Note that if $p(\theta)$ is unknown, then it must be treated as an optimization variable in the objective function of the test.

2. For notational convenience, we will often write $p_i(x^n) = p(x^n|s = i)$. Also, the hypotheses

$$H_0 : s = 0, \quad H_1 : s = 1$$

are equivalently expressed in terms of the conditional distribution of the observations as

$$H_0 : x^n \sim p(x^n|s = 0), \quad H_1 : x^n \sim p(x^n|s = 1),$$

where $x^n \sim p(x^n|s = i)$ means "x^n *is distributed according to* $p(x^n|s = i)$".

3. **Neyman-Pearson lemma**: Now suppose the binary hypothesis test satisfies the following two conditions.

   (a) The decision rule (B.9) is binary, i.e., $M = 2$, where the decision $u = 0$ is interpreted as acceptance of H_0 (or rejection of H_1) while $u = 1$ is acceptance of H_1 (or rejection of H_0).

   (b) The test's underlying objective function, such as (B.10), to be maximized is a convex function of the conditional probabilities $p(u|x^n)$, viewed as the main optimization variables.

   (c) The observations X^n are continuous variables and the distribution $p_i(x^n)$ is continuous.

   Then it can be shown (see Proposition 3.1) that under these conditions, the *optimal decision rule* for the binary hypothesis test takes the form

$$p_{\text{opt}}(\theta \in \Theta_i|x^n) = p_{\text{opt}}(u = i|x^n) = I_{R_{u=i}}(x^n), \quad i = 0, 1, \qquad \text{(B.11)}$$

   where the *decision regions* $R_{u=i}$ are given by

$$R_{u=1} = \{x^n \in \mathcal{X}^n : p_1(x^n)/p_0(x^n) > \lambda\}, \quad R_{u=0} \cup R_{u=1} = \mathcal{X}^n.$$

Here, $\lambda \in \mathbb{R}$ is a threshold parameter. A formal statement of this particular result is well known as the *Neyman-Pearson lemma*.

4. **A set estimator of** θ: Let us fix $\varepsilon > 0$, and let $\theta_0 \in \Theta$. Denote by the pair (ε, θ_0) the binary hypothesis test with

$$H_0 : \theta \in B(\theta_0, \varepsilon), \quad H_1 : \theta \notin B(\theta_0, \varepsilon),$$

where $B(\theta_0, \varepsilon) \subset \Theta$ is the open ball of radius ε centered at θ_0. Let the acceptance region for H_0 be

$$R_{(\varepsilon, \theta_0)} = R_{u=0}^{(\varepsilon, \theta_0)} = \left\{ x^n \in \mathcal{X}^n : \; p_1^{(\varepsilon, \theta_0)}(x^n) / p_0^{(\varepsilon, \theta_0)}(x^n) < \lambda_{(\varepsilon, \theta_0)} \right\},$$

where

$$p_0^{(\varepsilon, \theta_0)}(x^n) = \frac{\sum_{\theta \in B(\theta_0, \varepsilon)} p(x^n | \theta) p(\theta)}{\sum_{\theta \in B(\theta_0, \varepsilon)} p(\theta)}, \quad p_1^{(\varepsilon, \theta_0)}(x^n) = \frac{\sum_{\theta \notin B(\theta_0, \varepsilon)} p(x^n | \theta) p(\theta)}{\sum_{\theta \notin B(\theta_0, \varepsilon)} p(\theta)}.$$

Then a natural set estimator for θ is given by

$$T^{(\varepsilon)}(x^n) = \left\{ \theta_0 \in \Theta : x^n \in R_{(\varepsilon, \theta_0)} \right\}.$$

Given $\theta \in \Theta$, we have

$$P\left(\theta \in T^{(\varepsilon)}(X^n) \right) = P\left(X^n \in R_{(\varepsilon, \theta)} \right) = P\left(R_{(\varepsilon, \theta)} \right).$$

Thus, the confidence coefficient of $T^{(\varepsilon)}(x^n)$ satisfies

$$c^{(\varepsilon)} = \min_{\theta \in \Theta} P\left(\theta \in T^{(\varepsilon)}(X^n) \right) = \min_{\theta \in \Theta} P\left(R_{(\varepsilon, \theta)} \right).$$

5. **Hypothesis test sequences**: Although the result of a single binary hypothesis test does not necessarily yield a direct estimate for the true value of θ (except when Θ is a binary set), it does reduce the *search space* for the

true value of θ from Θ to $\Theta_0 \subsetneq \Theta$ or $\Theta_1 \subsetneq \Theta$. Thus, if we consider a sequence of consecutive binary hypothesis tests $\tau_1, \tau_2, \tau_3, \cdots$ and let $\Theta^{(\tau_k)}$ denote the search space in the kth test τ_k, then we have

$$\Theta = \Theta^{(\tau_1)} \subsetneq \Theta^{(\tau_2)} \subsetneq \Theta^{(\tau_3)} \subsetneq \cdots \tag{B.12}$$

That is, the result of a sufficiently long sequence of binary hypothesis tests will yield a reasonable estimate for the true value of θ. Of course if we consider a sequence of tests with more than two hypotheses, then the length N' of a sequence of such tests required to reach a certain desired level of accuracy will be smaller than the length N of a sequence of binary hypothesis tests required to reach the same level of accuracy.

6. ***Multiple hypothesis tests***: The description of binary hypothesis tests given above extends in a straightforward way to *multiple hypothesis tests*. The basic idea remains the same: To split up Θ into N disjoint subsets $\Theta_0, \Theta_1, ..., \Theta_{N-1}$, consider hypotheses $H_i : \theta \in \Theta_i$, and then find the conditional probabilities $p(\theta \in \Theta_i | x^n)$ that best suite a given objective.

   Once again, the multiple hypothesis test for θ with hypotheses $H_i : \theta \in \Theta_i$ is equivalent to a multiple hypothesis test for a discrete indicator variable s with hypotheses $H_i : s_i = \mu_i$, where $s = s(\theta) = \sum_{i=0}^{N-1} \mu_i I_{\Theta_i}(\theta) \in \{\mu_0, \mu_1, ..., \mu_{N-1}\}$, and the distribution of s is computed as

   $$p(x^n | s = \mu_i) = \frac{p(x^n, s = \mu_i)}{p(s = \mu_i)} = \frac{p(x^n, \theta \in \Theta_i)}{p(\theta \in \Theta_i)} = \frac{\sum_{\theta \in \Theta_i} p(x^n | \theta) p(\theta)}{\sum_{\theta \in \Theta_i} p(\theta)},$$

   for each $i \in \{0, 1, ..., N-1\}$.

7. Note that a multiple hypothesis test can be approximated by a number of binary hypothesis tests. Also, a binary hypothesis test can be approximated

by a *union-intersection* (or *intersection-union*) test, which is a combination of a number of binary hypothesis tests.

Generalized likelihood ratio tests

Consider a general binary hypothesis test

$$H_0 : \theta \in \Theta_0, \quad H_1 : \theta \in \Theta_1, \quad \Theta_0 \cup \Theta_1 = \Theta. \tag{B.13}$$

For a fixed $\theta_0 \in \Theta_0$ and a fixed $\theta_1 \in \Theta_1$, we have the simple test

$$H_0^{(\theta_0,\theta_1)} : \theta = \theta_0, \quad H_1^{(\theta_0,\theta_1)} : \theta = \theta_1.$$

Thus, we have a family of simple tests $\left\{ \left(H_0^{(\theta_0,\theta_1)}, H_1^{(\theta_0,\theta_1)} \right) : \theta_0 \in \Theta_0, \theta_1 \in \Theta_1 \right\}$. For each pair $(\theta_0, \theta_1) \in \Theta_0 \times \Theta_1$, let the decision region for the test $\left(H_0^{(\theta_0,\theta_1)}, H_1^{(\theta_0,\theta_1)} \right)$ be

$$R_{u=1}^{(\theta_0,\theta_1)} = \{ x^n \in \mathcal{X}^n : p(x^n|\theta_1)/p(x^n|\theta_0) > \lambda_{\theta_0,\theta_1} \}.$$

If we choose to accept a data point x^n under H_1 in (B.13) whenever it falls in any one of the regions $R_{u=1}^{(\theta_0,\theta_1)}$, then the test (B.13) has a (suboptimal) decision rule given by the decision region

$$R_{u=1} = \bigcup_{(\theta_0,\theta_1) \in \Theta_0 \times \Theta_1} R_{u=1}^{(\theta_0,\theta_1)} = \left\{ x^n \in \mathcal{X}^n : \sup_{\theta_0,\theta_1} \frac{p(x^n|\theta_1)/\lambda_{\theta_0,\theta_1}}{p(x^n|\theta_0)} > 1 \right\}$$

$$\subset \left\{ x^n \in \mathcal{X}^n : \frac{\sup_{\theta_1} p(x^n|\theta_1)}{\sup_{\theta_0} p(x^n|\theta_0)} > \sup_{\theta_0,\theta_1} \lambda_{\theta_0,\theta_1} \right\} = \tilde{R}_{u=1}. \tag{B.14}$$

Tests with a decision region of the form $\tilde{R}_{u=1}$ in (B.14) are called generalized likelihood ratio tests (GLRT's). Although such tests are clearly suboptimal in general, a remark we made earlier says that if such a test is repeated a sufficiently large number of times, it can yield very good results that may even be judged to be *asymptotically optimal* depending on the underlying objective of the test.

Union-Intersection (or Intersection-Union) tests

There are situations where a binary test (H_0, H_1) is seen to be composed of elements of a family of binary hypothesis tests $\{(H_{0\alpha}, H_{1\alpha}) : \alpha \in \mathcal{A}\}$, where $\mathcal{A}$ is an index set.

Consider a test with hypotheses $H_0 : \theta \in \Theta_0$, $H_1 : \theta \in \Theta_0^c$. Suppose that $\Theta_0 = \bigcap_{\alpha \in \mathcal{A}} \Theta_{0\alpha}$. Then

$$H_0 : \; \theta \in \Theta_0 = \bigcap_{\alpha \in \mathcal{A}} \Theta_{0\alpha}, \quad H_1 : \; \theta \in \Theta_0^c = \bigcup_{\alpha \in \mathcal{A}} \Theta_{0\alpha}^c.$$

Notice that the test involves separate tests of the form

$$H_{0\alpha} : \; \theta \in \Theta_{0\alpha}, \quad H_{1\alpha} : \; \theta \in \Theta_{0\alpha}^c, \quad \alpha \in \mathcal{A}. \tag{B.15}$$

Thus, if $R_{u=1}^{(\alpha)} = \{x^n : L_\alpha(x^n) > \lambda_\alpha\}$ is the decision region of the test $(H_{0\alpha}, H_{1\alpha})$ for each $\alpha \in \mathcal{A}$, then a (suboptimal) decision region for the test (H_0, H_1) is given by

$$R_{u=1} = \bigcup_{\alpha \in \mathcal{A}} R_{u=1}^{(\alpha)} = \bigcup_{\alpha \in \mathcal{A}} \{x^n \in \mathcal{X}^n : L_\alpha(x^n) > \lambda_\alpha\} = \left\{x^n \in \mathcal{X}^n : \sup_{\alpha \in \mathcal{A}} L_\alpha(x^n)/\lambda_\alpha > 1\right\}.$$

Similarly, if we consider a test with hypotheses $H_0 : \theta \in \Theta_0$, $H_1 : \theta \in \Theta_0^c$, and suppose that $\Theta_0 = \bigcup_{\alpha \in \mathcal{A}} \Theta_{0\alpha}$, then

$$H_0 : \; \theta \in \Theta_0 = \bigcup_{\alpha \in \mathcal{A}} \Theta_{0\alpha}, \quad H_1 : \; \theta \in \Theta_0^c = \bigcap_{\alpha \in \mathcal{A}} \Theta_{0\alpha}^c.$$

We again observe that the test involves separate tests of the form (B.15). Thus, if

$$R_{u=1}^{(\alpha)} = \{x^n \in \mathcal{X}^n : L_\alpha(x^n) > \lambda_\alpha\}$$

is the decision region of the test $(H_{0\alpha}, H_{1\alpha})$, then a (suboptimal) decision region for the test (H_0, H_1) is

$$R_{u=1} = \bigcap_{\alpha \in \mathcal{A}} R_{u=1}^{(\alpha)} = \left\{x^n \in \mathcal{X}^n : \inf_{\alpha \in \mathcal{A}} L_\alpha(x^n)/\lambda_\alpha > 1\right\}.$$

References

[1] Z. B. Tang, K. R. Pattipati, and D. L. Kleinman, "An algorithm for determining the decision thresholds in a distributed detection problem," *IEEE Trans. Sys., Man, and Cyb.*, Vol. 21, No. 1, pgs 231-236, Jan./Feb. 1991.

[2] Z. B. Tang, K. R. Pattipati, and D. L. Kleinman, "A distributed m-ary hypothesis testing problem with correlated observations," *IEEE Trans. Aut. Control*, Vol. 37, No. 7, pgs 1042-1046, July 1992.

[3] Z. B. Tang, K. R. Pattipati, and D. L. Kleinman, "Optimization of detection networks: Part I – Tandem structures," *IEEE Trans. Syst., Man, Cybern.*, vol. 21, no. 5, pp. 1044-1059, 1991.

[4] Z. B. Tang, K. R. Pattipati, and D. L. Kleinman, "Optimization of detection networks: Part II – Tree Structures," *IEEE Trans. Sys., Man, Cybern.*, vol. 23, no. 1, pp. 211-221, 1991.

[5] S. Alhakeem and P. K. Varshney, "A Unified Approach to the Design of Decentralized Detection Systems," *IEEE Trans. Aero. Elec. Sys.*, Vol. 31, No. 1 January 1995.

[6] J. A. Gubner, L. L. Scharf, and E. K. P. Chong, "Optimization of Exponential Error Rates for a Suboptimum Fusion Rule in Wireless Sensor Networks," *2011 Conference Record of the Forty Fifth Asilomar Conference on Signals, Systems and Computers (ASILOMAR)*, Pacific Grove, CA, Nov. 2011

[7] Z. Zhang, E. K. P. Chong, A. Pezeshki, W. Moran, and S. D. Howard, "Submodularity and Optimality of Fusion Rules in Balanced Binary Relay Trees," *51st IEEE Conference on Decision and Control*, Maui, Hawaii, USA, Dec. 2012.

[8] P. K. Varshney, *Distributed Detection and Data Fusion*. New York: Springer, 1997.

[9] V. V. Veeravalli, T. Basar, and H. Poor, "Decentralized sequential detection with a fusion center performing the sequential test," *IEEE Trans. Inf. Theory*, vol. 39, pp. 433-442, Mar. 1993.

[10] Y. M. Zhu, R. S. Blum, Z. Q. Luo, and K. M. Wong, "Unexpected properties and optimum distributed sensor detectors for dependent observation cases", *IEEE Transactions on Automatic Control*, 45(1): 62-72, 2000.

[11] Y. M. Zhu and X. R. Li, "Unified fusion rules for multisensor multihypothesis network decision systems", *IEEE Trans. Systems, Man and Cybernetics Part A*, 33(4): 502-513, 2003.

[12] E. Song, X. Shen, J. Zhou, Y. Zhu, and Z. You, "Performance analysis of communication direction for two-sensor tandem binary decision system",

IEEE Transactions On Information Theory, Vol. 55, No. 10, October 2009.

[13] E. Song, Y. Zhu, J. Zhou, "Some progress in sensor network decision fusion", *Jrl Syst Sci* & *Complexity* (2007) 20: 293-303.

[14] D. Teneketzis and Y. C. Ho, "The decentralized wald problem," *Information and Computation*, vol. 73, pp. 23-44, 1987.

[15] A. LaVigna, A. M. Makowski, and J. S. Baras, "A continuous-time distributed version of the wald's sequential hypothesis testing problem," *Lecture Notes in Control and Information Sciences*, vol. 83, pp. 533-543, 1986.

[16] V. V. Veeravalli, Topics in decentralized detection, *ProQuest Dissertations and Theses*, 1992.

[17] V. V. Veeravalli, "Sequential decision fusion: Theory and applications," *Journ. Franklin Inst.*, vol. 336, pp. 301-322, Feb. 1999.

[18] E. Akofor and B. Chen, "Interactive fusion in distributed detection: architecture and performance analysis," *Proc. of IEEE International Conference on Acoustic, Speech, and Signal Processing, Vancouver, Canada*, May 2013.

[19] E. Akofor and B. Chen, "Interactive distributed detection: architecture and performance analysis", *IEEE Trans. Info. Theory*, Vol. 60, Issue 10, pgs. 6456-6473, 2014.

[20] S. Zhu, E. Akofor, and B. Chen, "Interactive distributed detection with conditionally independent observations," *IEEE WCNC, Shanghai China*, April 2013.

[21] E. Akofor and B. Chen, "On optimal fusion architecture for a two-sensor tandem distributed detection system", *IEEE GlobalSIP, Austin, TX*, Dec. 2013.

[22] R. R. Tenney and N. R. Sandell, "Detection with distributed sensors," *IEEE Trans. Aero. and Elect. Systems*, Vol. AES-17, Issue 4, 501 - 510, July 1981.

[23] G. Parmigiani and L. Inoue, *Decision Theory - Principles and Approaches*, Wiley & Sons, 2009.

[24] D. W. North, "A tutorial introduction to decision theory", *IEEE Trans.*, 1968.

[25] S. O. Hansson, "Decision Theory - A Brief Introduction", 1994.

[26] A. Mazumder, "Statistical decision theory: concepts, methods and applications", http://probability.ca/jeff/ftpdir/anjali0.pdf

[27] J. O. Berger, *Statistical Decision Theory and Bayesian Analysis*, 2nd. Ed, Springer-Verlag, New York, 1980.

[28] A. Wald, "Statistical decision functions", *Ann. Math. Stat*, 1949.

[29] J. N. Tsitsiklis, "Extremal properties of likelihood-ratio quantizers," *IEEE Trans. Comm.*, vol. 41, no. 4, pp. 550-558, April. 1993.

[30] P. Willett and D. Warren, "Randomization in Distributed Detection," *Proceedings of the 29th Conference on Decision and Control*, Honolulu, Hawaii, Dec. 1990.

[31] P. Willett and D. Warren, "The Suboptimality of Randomized Tests in Distributed and Quantized Detection Systems," *IEEE Trans. Inf. Theory*, Vol. 38, No. 2, pgs 355-361, March 1992

[32] J. D. Papastavrow and M. Athans, "On optimal distributed decision architectures in a hypothesis testing environment," *IEEE Trans. Autom. Control*, vol. 37, no. 8, pp. 1154-1169, Aug. 1992.

[33] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. New York, Wiley, 2nd edition, 2006.

[34] S. Alhakeem and P. K. Varshney, "Decentralized Bayesian detection with feedback," *IEEE Trans. Sys., Man, Cyb. - A: Sys. Hum.*, Vol. 26, No. 4, pp. 503 - 513, July 1996.

[35] W. P. Tay and J. N. Tsitsiklis. "The value of feedback for decentralized detection in large sensor networks." *Proceedings of the 6th International Symposium on Wireless and Pervasive Computing (ISWPC)*, 2011. 1-6.

[36] H. Shalaby and A. Papamarcou, "A note on the asymptotics of distributed detection with feedback," *IEEE Trans. Inf. Theory*, vol. 39, no. 2, pp. 633-640, Mar. 1993.

[37] J. D. Papastavrou and M. Athans, "Distributed detection by a large team of sensors in tandem," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 28, no. 3, pp. 639-653, 1992.

[38] W. P. Tay, J. N. Tsitsiklis, and M. Z. Win, "On the sub-exponential decay of detection error probabilities in long tandems," *IEEE Trans. Inf. Theory*, vol. 54, no. 10, pp. 4767-4771, Oct. 2008.

[39] P. F. Swaszek, P. Willett, "Parley as an Approach to Distributed Detection", *IEEE Trans. Aero. & Elec. Sys.*, Vol. 31, No. 1, Jan. 1995.

[40] Y. Xiang and Y. Kim, "Interactive hypothesis testing with communication constraints," *Proc. Allerton Conference on Communication, Control, and Computing, Montecillo, IL*, Sept. 2012.

[41] J. N. Tsitsiklis, "Decentralized detection", *Advances in Statistical Signal Processing*, Volume 2, pages 297-344, 1993.

[42] S. Marano, V. Matta, F. Mazzarella, "Refining decisions after losing data: the unlucky broker problem," *IEEE Transactions on Signal Processing*, vol.58, no.4, pp.1980,1990, April 2010.

[43] S. Marano, V. Matta, F. Mazzarella, "The Bayesian unlucky broker", *18th European Signal Processing Conference (EUSIPCO-2010), Aalborg, Denmark*, pp.154-158, August 2010.

[44] M. Longo, T. D. Lookabaugh, and R. M. Gray, "Quantization for decentralized hypothesis testing under communication constraints", *IEEE Trans. Inf. Theory*, vol. 36, no. 2, pp. 241-255, Mar. 1990.

[45] B. Chen and P. K. Willett, "On the Optimality of the Likelihood-Ratio Test for Local Sensor Decision Rules in the Presence of Nonideal Channels," *IEEE Trans. Inf. Theory*, Vol. 51, No. 2, Feb. 2005.

[46] H. Chen, B. Chen, P. K. Varshney, "Further Results on the Optimality of the Likelihood-Ratio Test for Local Sensor Decision Rules in the Presence of Nonideal Channels," *IEEE Trans. Inf. Theory*, Vol. 55, No. 2, Feb. 2009.

[47] S. Boyd and L. Vandenberghe, *Convex Optimization* (Cambridge University Press).

[48] R. Enhbat, "An algorithm for maximizing a convex function over a simple set", *Journal of Global Optimization* 8:379-391, 1996.

[49] G. Casella and R. L. Berger, *Statistical Inference* (2nd Edition).

VITA

Name of author: Earnest Akofor

Place of birth: Awing, Cameroon

Date of birth: March 18, 1977

Degrees awarded:

Ph.D. Physics, Syracuse University, 2010

Diploma in Mathematical Sciences, AIMS, 2004

Diploma in High Energy Physics, ICTP, 2003

B.Sc. Physics, University of Buea, 2001

Professional experience:

Peer Review, IEEE Transactions, 2015-2016

Research Assistant, Syracuse University, 2011-2016

Adjunct Instructor, ITT Technical Institute, 2011

Teaching & Research Assistant, Syracuse University, 2004-2009