STRUCTURAL AND METRICAL INFORMATION IN LINEAR SYSTEMS

A THESIS

SUBMITTED TO THE DEPARTMENT OF ELECTRICAL AND

ELECTRONICS ENGINEERING

AND THE INSTITUTE OF ENGINEERING AND SCIENCES

OF BILKENT UNIVERSITY

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

FOR THE DEGREE OF

MASTER OF SCIENCE

By

Ayça Özçelikkale

August 2006

I certify that I have read this thesis and that in my opinion it is fully adequate, in scope and in quality, as a thesis for the degree of Master of Science.

Prof. Dr. Haldun M. Özaktaş (Supervisor)

I certify that I have read this thesis and that in my opinion it is fully adequate, in scope and in quality, as a thesis for the degree of Master of Science.

Prof. Dr. Erdal Arıkan

I certify that I have read this thesis and that in my opinion it is fully adequate, in scope and in quality, as a thesis for the degree of Master of Science.

Prof. Dr. Mustafa Pınar

Approved for the Institute of Engineering and Sciences:

Prof. Dr. Mehmet Baray
Director of Institute of Engineering and Sciences

ABSTRACT

STRUCTURAL AND METRICAL INFORMATION IN LINEAR SYSTEMS

Ayça Özçelikkale
M.S. in Electrical and Electronics Engineering
Supervisor: Prof. Dr. Haldun M. Özaktaş
August 2006

We present a systematic approach to understand the information-theoretic relationships in linear systems. Our main aim is to understand what kind of information the output of a linear system carries about the input of the system and how much of this information is preserved in the measurement process. We recognize structural and metrical information as two fundamental concepts for classifying the information content of signals. We base our understanding of the problem on information-theoretic concepts like entropy, mutual information and channel capacity. We present our results as trade-offs between cost and performance, yielding insights about different aspects of the information flow in a linear system. We especially focus on building a framework which indicates how accurately, and how many measurements must be made and how the measurement locations should be selected.

Keywords: inverse problems, signal recovery, structural information, metrical information, experiment design, measurement problem, information theory, fractional Fourier transform, wave propagation, optical information processing

ÖZET

DOĞRUSAL SİSTEMLERDE YAPISAL VE ÖLÇEVSEL BİLGİ

Ayça Özçelikkale
Elektrik ve Elektronik Mühendisliği Bölümü Yüksek Lisans
Tez Yöneticisi: Prof. Dr. Haldun M. Özaktaş
Ağustos 2006

Bu tezde doğrusal sistemlerdeki bilişim kuramı ilişkilerini anlamak için sistematik bir yaklaşım sunuyoruz. Temel amacımız doğrusal bir sistemin çıktısının girdisi hakkında ne çeşit bir bilgi taşıdığını ve bu bilginin ne kadarının ölçüm alma sürecinde korunduğunu anlamaktır. Bu amaçla, yapısal bilgi ve ölçevsel bilgi kavramlarını işaretlerin bilgi içeriklerini sınıflandırmak için iki temel kavram olarak kullanıyoruz. Yaklaşımımızı entropi, karşılıklı bilgi ve kanal kapasitesi gibi bilişim kuramına ait kavramlara dayandırıyoruz. Sonuçlarımızı doğrusal sistemlerdeki bilgi akışının farklı yönlerini kavramamızı sağlayan maliyet ve performans arasındaki ödünleşimler olarak sunuyoruz. Özellikle de ölçümlerin ne kadar doğrulukla ve kaç tane yapılması gerektiği ve ölçüm noktalarının seçilmesini kapsayan bir teori kurmak üstünde yoğunlaşıyoruz.

Anahtar Kelimeler: ters problemler, işaret geri kazanımı, yapısal bilgi, ölçevsel bilgi, deney tasarımı, ölçüm problemi, bilişim kuramı, kesirli Fourier dönüşümü, dalga yayılımı, optik bilgi işleme

ACKNOWLEDGMENTS

I would like to express my sincere gratitude to Prof. Dr. Haldun M. Özaktaş for his supervision and guidance throughout the development of this thesis. I am obligated to express my indebtedness for his keen concern and encouragement.

I would like to thank the members of the thesis committee Prof. Dr. Erdal Arıkan and Prof. Dr. Mustafa Pınar for accepting to read the manuscript and commenting on the thesis.

Special thanks to Kıvanç Köse and A. Polat Ay for their support in terms of catering, accommodation and encouragement during writing of this thesis. I would like to thank Gökhan Bora Esmer and Erdem Ulusoy for the fruitful discussions. I would like to thank my office mates Namık Şengezer, Ahmet Serdar Tan, Kaan Doğan and Mehmet Köse for their support and endless tolerance.

It is a great pleasure to express my gratitude to my family, Gülsen, Levent, Altuğ Özçelikkale and H. Volkan Hünerli. Without their love and patience, this work would not be possible. Finally, I would like to thank M. Volkan for being himself, which has made life joyful even in the period of writing the manuscript of this thesis.

Contents

List of Figures

Dedicated to my family

Chapter 1

Introduction

The use of linear systems in modelling physical phenomena is a common practice in engineering. Linear systems are used in various areas such as signal processing, control theory, and communication theory. In this thesis we focus on linear systems from an information-theoretic point of view. We develop a framework for information-theoretic interpretation of input-output relationships in linear systems. Our main intended area of application is optical fields, but our approach is not based on any property specific to this area.

Our basic goal is to understand what happens to the information contained in a signal after it passes through a linear system. How much of the information that was originally present in the input signal is preserved in the output of the linear system is a question of central importance. We are also interested in the practical limitations regarding information that can be recovered from the output signal. We would like to understand the effect of finite precision measurement devices on the quality of the recovered information.

To achieve these goals, we focus on building a framework where these questions can be formulated in their most natural terms. For this purpose, we reconsider an interpretation of information mentioned in [1]. This work distinguishes

Figure 1.1: Block diagram of a process with its input and output

structural and *metrical* aspects of information as two fundamental concepts. These concepts provide the framework for the analysis in this thesis.

1.1 Model

In this thesis, we focus on the model in Figure 1.1, which shows a system and its input and the output, constituting the three components of interest.

Usually there is information on some of these components and with this information we want to extract some information about the other components. In a typical framework, the process parameters are assumed to be known a priori, the output signal is observed possibly with some error, and the input signal which explains the observed signal best under the given process parameters is investigated. This problem is referred to as an inverse problem or the signal recovery problem.

1.2 Possible Approaches to the Problem

To understand the information-theoretic relationships in such systems, different approaches may be adopted.

One approach is to focus on numerical experimentation in a brute force manner. One may assume the process is completely known and focus on the signal

recovery problem. Observing the change in the quality of the recovered input signal while the location and number of the samples of the output change is a possible method. Descriptive conclusions about the nature of information flow can be drawn from these simulations.

Another approach is to focus on the process. One assumes the process itself determines the form of information flow. In a discrete framework it is possible to represent the process by a system matrix and focus on its algebraic properties such as Singular Value Decomposition (SVD), rank, and condition number. In a continuous framework concepts like eigenfunction decomposition and bandwidth may be useful.

Instead of these two approaches, we adopt an information-theoretic approach. We express the relationships between physical quantities with information theoretical concepts like entropy, mutual information, and channel capacity. While modelling the problem, we pay special attention to preserving generality. As a result, in our framework the unknown parameters and given parameters can be related to all of the three main components: input, process, and output. We systematically define the problems to exploit the link between the information contained in the unknown parameters and the known parameters.

1.3 Classification of Problem Parameters

Each problem is designed to understand the relationship between the information contained in the known parameters and the unknowns. We define a problem parameter as any quantity that is a function of the three main components of the problem (input, output, process).

Problem parameters can be classified as quantities on a scale of varying degree of freedom. At one end there is the extreme of given parameters. Given

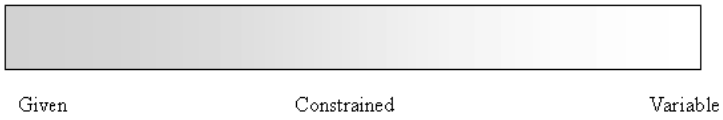

Figure 1.2: Classification of problem parameters

parameters are the ones whose values cannot be changed. At the other end, there are the variable parameters. These are the ones whose values can be changed to achieve certain goals in the problem. Constrained parameters are considered as an intermediate group between given and variable parameters. More strictly constrained parameters are closer to the given parameter end of the scale and less constrained parameters are closer to variable parameter end of the scale. This classification is illustrated in Figure 1.2.

Every problem has an objective to be optimized. It is given as function of the problem parameters. An objective should not be a given parameter of the problem, since if it was so, there would be no point in optimizing it.

A broad range of problems related to performing measurements and estimating unknowns from them can be stated in this framework. The relationships between the problem parameters are mostly exploited by trying to express the trade-offs between these parameters. These trade-offs will reveal the relationship between the information contained in the problem parameters in a systematic manner.

We investigate a restricted but very important class of problem parameters in section 1.4. This class of parameters, although very small, provides insight for a considerable number of problems in the literature. In section 1.5, some of the possible trade-offs that can be interesting are illustrated.

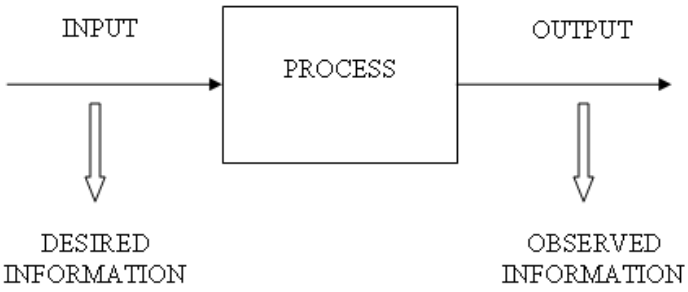

Figure 1.3: The desired information and the observed information are subsets of the input signal and the output signal respectively

1.4 Structural and Metrical Information

Let us assume that the input signal is unknown and output signal is partially known (observed information). We desire to partially or wholly obtain the input signal (desired information). These definitions are illustrated in Figure 1.3.

In this thesis, we will distinguish between structural information and metrical information. Structural information is related to the inherent structure of information in space, time, or another coordinate variable. Within this framework, information is assumed to be distributed over independent coordinates and the emphasis is on description of how this information is distributed over these coordinates. Metrical information is related to the values of the quantities that carry the information.

The following sections discuss our understanding of structural and metrical information. While interpreting these concepts we classify signals into two groups: a) signals whose existence is independent of our observations or interest in them b) signals that are related to our efforts to obtain information. For instance, consider the temperature distribution in a room. The three-dimensional signal which gives the values of the temperature in this room is a signal of the first kind. It exists whether we observe it or not. When we put sensors at several

locations in this room and measure the temperature, the measurement values constitute a set of samples obtained as a result of our measurements. These samples constitute a signal of the second kind. If we use these measurements to reconstruct the actual temperature signal with a particular resolution, this desired information will also be of the second kind. In our scheme the input signal and the output signal are of the first kind and desired information and observed information are of the second kind.

1.4.1 Structural Information

The structural information contained in a signal of the first kind may be interpreted as the structure and the number of independent quantities that should be known to uniquely characterize the signal. Hence the structural information is a description of the set of signals that the signal is a member of. From this point of view, structural information represents our a priori knowledge about the signal. As a simple example consider a point about which we know that it lies on a circle with a known radius and known center in two-dimensional space. This a priori information is the structural part of the information associated with the exact position of the point. With this structural information at hand, we know it is sufficient to learn the angle instead of the two independent coordinates to possess all the information.

For signals of the second kind, the structural information is strongly related to our method of obtaining this signal. For example, consider a signal which is formed by sampling another signal. Then the structure of information conveyed by this observation will be determined by our sampling method. For a given sampling rate with a uniform sampling strategy, the observed signal will not possess information about details smaller than a predefined value. Hence structural information is closely connected to the concept of resolving power. Similarly, if the sampled signal does not have details smaller than a predefined value, no matter

how closely we take samples the observed signal will not exhibit more than that level of detail.

1.4.2 Metrical Information

Metrical information is related to the amplitude values of individual variables. It is the answer to the question "what is the value of this individual variable?"

The most important feature of metrical information is the accuracy or the resolution of the amplitude values of individual variables. For the signals of the first kind, this feature answers the question "how many distinguishable levels are there in the values of this signal?" The answer to this question is related to the inherent noise present in all kinds of physical phonemena. This noise is independent of our attempt to measure values of these signals. For the signals of the second kind, this feature is related to the accuracy of the values of these signals. For the observed information it is the answer to the question "with how much uncertainty do we observe each value?" Therefore this aspect of metrical information is closely related to the precision of measurement devices and the quantization of the results of a measurement, which is also known before doing a measurement. For the desired information, this feature is the answer to the question "with how much uncertainty do we want to learn each value of the input signal?" The answer to this question is constrained by the method we use for signal recovery as well as the accuracy of the observed information.

These concepts may be also interpreted from the point of view of σ-algebras, for which an introduction can be found in [2]. Looking at information from a structural-metrical perspective enables us to state a large class of problems in our framework in a systematic way. In this thesis, we explore the trade-offs between the cost of measurements and the extracted information (performance) in terms of these concepts.

1.5 Examples of Experiment Design Problems

By assigning the structural and metrical problem parameters related to desired information and observation, into either given, variable, or constrained classes, it is possible to express a broad class of different and interesting trade-off problems.

We assume the process is completely known. We want to extract information about the input signal by the help of observations. The following four points outline typical structural and metrical constraints on the desired information and observed information.

- Certain structural constraints may be imposed on the desired information: One may want to learn the input signal with a resolution that is at least as good as a certain predefined value. It is also possible not to feel the need to learn the signal with a resolution better than a value. This case may occur in situations where the postprocessing of the recovered signal will be performed with a limited bandwidth. It is also possible to specify a completely arbitrary organization of samples where the locations of samples correspond to the points one wants to learn the values of the input signal.

- Certain structural constraints may be imposed on the observed information: It may be desired to have the observations as close as possible in the situation that there is a travel cost in moving from one location to another. It may also be desired to have the observations as far as possible in cases where some a priori information about the process indicates that samples taken too close will not contain new information.

- Certain metrical constraints may be imposed on the desired information: One may be interested to learn the input signal values with an accuracy that is not less than a predefined level. Similarly, an accuracy greater than a certain value may be unnecessary for some applications.

- Certain metrical constraints may be imposed on the observed information: The available measurement devices may constrain the accuracy that the observations can be made.

Several interesting problems can be expressed as trade-off problems between these parameters. An important class is the relationship between structural information content of the input signal and the output signal under a given set of metrical constraints. The problem of determining the optimal locations of sensors at the output end of the process, in order to learn the input with a predetermined resolution is an example. One special case of this example is the problem whose result is stated as Nyquist-Shannon sampling theorem. This case focuses on a particular relationship between the structural constraints on the desired information and the output signal when the process is taken to be identity. The desired information is taken to be the input signal, no uncertainty in the reconstructed signal and observations is allowed, and a sampling strategy with equal intervals is adopted on the output signal. The theorem states the sufficient sampling interval length to satisfy these constraints.

Similarly, an interesting class of problems is to investigate the relationship between the metrical information content of desired signal and observations under a given set of structural constraints. Questions like "with how much accuracy should each observation be made, in order to obtain a given accuracy in the values of the desired signal?" can be answered in this framework.

1.6 Information Measure

To be able to understand the relationship between the information contained in the input signal and samples of the output signal, we should have a measure of information.

The most natural information-theoretic concept to use as a measure of information is mutual information. Mutual information can be roughly defined as a measure of the average reduction in the uncertainty of one random variable due to knowledge of another. This concept is discussed in detail in section 3.5.2. To understand the relationship between the unknown vector and the measured vector, it is natural to investigate the mutual information between the unknown vector and the measured vector.

One alternative is to use the quality of the recovered input signal as a measure of the information. As the quality of the recovered signal gets better, we derive the conclusion that the observations preserved more information about the input signal. This approach is highly dependent on the estimation technique. Different estimation techniques can recover different types of information about the input signal. Different approaches may be summarized as follows:

The explanations below assume an unknown vector f and a vector of observations s with a process represented by the matrix H. (These may be also used as a measure of information even in the case where the known parameters and unknowns are different.)

- Non-Probabilistic Approaches:
  - Norm-Approximation ($f_{\text{est}} = \arg\min_f \|Hf - s\|$)
  - Weighted Norm Approximation ($f_{\text{est}} = \arg\min_f \|W(Hf - s)\|$ where W is the weighting matrix)
  - Singular Value Decomposition ($f_{\text{est}} = H^+ s$ where H^+ is pseudo-inverse of H — this is the minimum length least-squares solution)
- Probabilistic Approaches:
  - Maximum Likelihood (ML) Estimation (if the noise is iid Gaussian, ML estimate is the same as the solution of the least squares problem)

- Maximum A Posteriori Probability (MAP) Estimation
- Minimum Mean-Square Error (MMSE) Estimator (for jointly Gaussian random variables, MMSE estimate and MAP estimate are identical)
- Cramer-Rao Bound (Cramer-Rao Bound provides a lower bound for the variance of unbiased estimators)

If the noise is assumed to be Gaussian, ML problem is the same as weighted norm approximation problem where weighting matrix is found by Cholesky factorization of inverse of noise covariance.

In this thesis, we work with MMSE estimation case. Since the input is assumed Gaussian and the process is modelled by a linear system, the input and the output are jointly Gaussian. Hence MAP case is also covered by our MMSE estimation formulation.

1.7 Illustrative Example

This section presents an example which illustrates the measurement design problem, with the purpose of making the concepts mentioned so far more concrete.

We consider the system in Figure 1.4. An optical system alters the distribution of light in the input plane and produces the distribution of light in the output plane. We assume the rule of this mapping is known. We would like to get information about the distribution in the input plane, but we have access only to the output plane. We will make some measurements on the output plane with sensors varying in the precision and cost. As the precision offered by a device increases, its cost also increases.

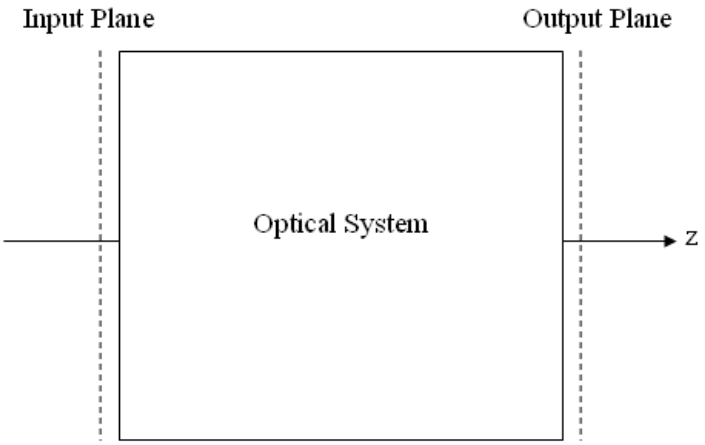

Figure 1.4: Optical system

Since we have a limited budget, we can use a finite number of sensors. We want to choose the places to put the sensors. We cannot put them too close because of the physical dimensions of the sensors. As a matter of fact, we tend to believe that putting the sensors too close will not be beneficial, since the data collected by sensors that are too close will probably be redundant. (although in some cases this redundancy may compensate for the effect of measurement noise.)

Other than deciding the sensor locations we also want to decide what the precision of each device should be. We would prefer to use the highest precision devices available, but we have limited budget and high precision devices cost more.

We might want to learn the answers to questions such as the following:

- What is the best sampling strategy, given a total number of bits (corresponding to cost) to represent all of our measurements?
- To satisfy a given distortion constraint, with what resolution should each measurement be done?
- What should be the number of detectors to satisfy a distortion constraint?

- Where and with which resolution should the detectors be placed?
- What are the trade-offs between sampling rate and sampling accuracy? Which is better: a small number of high precision devices or a large number of low precision devices?
- How can we compare the value of lower and higher significant bits among different samples?

This is an example of the problems that have motivated us to study the information-theoretic interpretation of input-output relationships in linear systems. Although intution and commonly used techniques can guide us through some of these decisions, existing knowledge on this problem does not seem to be consolidated and unified. This thesis aims to provide the groundwork towards this end.

1.8 Contributions

Several problems related to information flow in linear systems have been studied in various contexts in earlier works, including the limits of the information transfer capability of optical systems, the problem of sensor placement in control systems, and the problem of coding of outputs of sensors. However, in these contexts either the emphasis is not on the problem of understanding what happens to the information contained in a signal after it passes through a linear system, or the approach adopted is not as general or systematic as one might wish. This thesis directly focuses on this problem and presents a novel framework where this problem can be systematically investigated. We use the concepts of structural and metrical information while building our framework. These concepts have been proposed before, but their usage as the basis of such a systematic framework is, to the best of our knowledge, new.

In this thesis, we also study the practical limitations regarding the information that can be recovered from the output of a linear system. To be able to model the act of performing a measurement in an abstract manner, we associate a cost with every measurement. The proposed cost function is consistent with the properties which we believe a plausible cost function should have and it is a new approach in understanding a measurement.

To understand the information-theoretic relationships in linear systems, we formulate different trade-off problems. The trade-off problem stated in section 4.3 is illustrative of the kinds of problems which can be formulated in our framework. In this problem the MMSE estimation of the unknown vector from the observation vector, when we are allowed to vary the measurement accuracy of components of the observation vector, is studied. The trade-off problem stated in section 4.6 investigates the mutual information between the input and the output of a channel when a part of the channel has a limited capacity. The trade-offs illustrated in section 5.4 focus on the location of samples in space. Finally, the numerical results shown in section 6.3 illustrate the trade-off between error and resolution in space and accuracy in amplitude.

1.9 Outline

A brief overview of related work is presented in chapter 2. Our model of measurement and proposed definition of measurement cost is given in chapter 3. This chapter also exploits the proposed cost function's relationship with the concept of number of distinguishable levels and information theory. In chapter 4, we focus on the purely metrical problem, dealing with the accuracy of measurements and estimation error. In this chapter, we also investigate the relationship between the accuracy of the measurements and the mutual information. The purely structural problem, focusing on resolution in space is formulated in chapter 5. In

chapter 6, the metrical and structural problems are unified. Finally, chapter 7 presents the conclusions of this thesis and outlines directions for future work.

Chapter 2

Related Work

Information transmission capability of optical systems has been an important area of research. Although it is a broad area which can be dated as far back as the 1910s [3], the 1950s are the times the subject has been intensively investigated. This section presents the history of the subject focusing on theoretical developments. A treatment of the history with special emphasis on research which leads to practical progress can be found in [4] and [5]. This section also reviews a collection of works that are related to selection and coding of measurements in signal processing, control theory, and information theory.

Research in optical transmission of information in the 1970s focuses on the concept of number of degrees of freedom (DOF). DOF is interpreted differently in different contexts. Signals, systems, communication channels, number of elements of signal sets are some examples of the concepts to which a definition of DOF is associated. An illustrative definition for signals given by Von Laue is mentioned in [3] as the number of independent real parameters necessary to describe a scalar wave field completely.

In [3], Lukosz compares DOF and space-bandwidth product and concludes that DOF is the fundamental invariant of optical systems. In [6], ideas presented

in [3] are illustrated and a method for obtaining spatial super resolution by sacrifice of temporal resolution is introduced.

In [7], Toraldo di Francia derived the conclusion that an image formed by a finite pupil has finite degrees of freedom using the sampling theorem. In [8], the author recognized the inconsistencies of the results based on sampling theorem and investigated practical limitation of DOF by applying the theory of the prolate spheroidal functions.

The concept of DOF is extensively studied in [9], [10], [11], [12], [13], [14], [15]. In [11], DOF from point-like element pupils using eigenfunctions of integral equation is found. In [16], DOF in the presence of noise is illustrated. In [12], DOF for scatterers with circular cross section in the presence of noise with eigenfunction technique is studied. In [13], DOF without noise with eigenfunction technique for spherical scatterers is investigated.

The results presented in the mentioned works are mostly based on the scalar approximations and paraxial approximations and studied for specific optical systems. An analysis of DOF for transmission of information with electromagnetic waves between domains in three-dimensional space is given in [14].

Different approaches to the problem are also pursued. In [1] MacKay offers the terminology of *structural information* and *metrical information* to the engineering community. These concepts are used as a basis for understanding the information transmission capability of optical systems in [17]. This idea is also reviewed in [18]. Reference [17] develops the concept of an information-flow vector assigned to each point of the wave field to understand the flow of structural information. Reference [19] discusses whether two fields with different coherence properties can produce the same optical intensity everywhere in the space and investigates the differences in one-dimensional and two-dimensional case.

To understand the relationship between the nature of optical information transmission and information theory, attempts to connect concepts from optics and information theory have been done. Reference [20] investigates the entropy of a point-spread function as a measure of its effective area. This work shows how some drawbacks of the definition of entropy in information theory can be interpreted as natural consequences of properties of the optical diffraction integral. In [21], field propagation in terms of communication modes is studied. Reference [22] uses information theory concepts to describe and analyze physical properties of coherent and partially polarized light. Reference [23] proposes a method for using Shannon number and information capacity to provide compact performance measures of integral imaging systems. Reference [24] studies laser beam characterization based on Shannon's information-entropy formula. Reference [25] presents a new variational principal that concerns both the phase and intensity of a wave in the framework of geometrical-optics approximation of the wave equation which may be of use understanding the nature of information transmission.

The practical information transmission limitation of optical laws has also been studied with a sampling approach. In [15], Gori gives an account of the uses of sampling in optics. DOF, fundamental properties of Fresnel transform and their optical significance, Mellin transform and exponential sampling, role of sampling in coherence theory are the main subjects reviewed in this work. Reference [26] focuses on the convolution kernel describing the Fresnel diffraction and provides a reconstruction method. In [27], reconstruction of Fresnel fields sampled with nonideal sampling devices is studied.

To exploit the relationships between information contained in optical fields in different areas of space, it is possible to focus on the signal recovery problem with a numerical approach. Reference [28] gives an overview of the method of Projection onto Convex Sets (POCS) and other iterative methods for image

recovery. Reference [29] provides a generic introduction to image recovery by the method of POCS. Reference [30] provides an application of this method to optics in the context of resolution enhancement. Reference [31] presents another application of method of POCS. In this work, the authors assume the optical field is known at some random points in space and reconstruct the optical field at other points by POCS.

In our numerical examples, we will employ the fractional Fourier transform (FRT) as an example system because it captures the essence of wave propagation in a mathematically pure way. Reference [32] provides a comprehensive account of FRT and its history. The FRT, which is a generalization of the ordinary Fourier transform implies a more general formulation of the area of optical information processing. In [32], references to the milestones of the development of FRT and its applications are given. This book also presents an overview of basic concepts and tools which have been important in the history of optical information processing such as DOF, Wigner distribution and Gabor expansion. A review which clarifies the concept of DOF as the area of the space-frequency support and which emphasizes its difference from the space-bandwidth product is also given.

References [33] and [34] present a general overview of the relationship between information theory and optics. To describe the optical spatial channel and its information theoretic characteristics, these texts provide introductory material on information theory, diffraction and signal analysis. The relationship between the concept of entropy in thermodynamics and entropy in information theory is extensively studied. Information provided by observations is discussed with a strict connection to the wave nature of light and quantum theory. Several applications in the area of optical information processing including image restoration, wavelet transforms, pattern recognition, computing with optics and fiber-optic communication are also covered.

Optical systems are frequently modelled as linear systems. The relationship between samples of the output of linear shift-variant systems and the input are also studied within a signal processing and communication framework. Reference [35] has shown that a bandlimited signal of finite energy passing through a single-input multiple-output system can be uniquely reconstructed from the samples of outputs of the system under some conditions on the system. Reference [36] studies recovery of input from finitely many noisy output data where the system is driven by a differential equation. Reference [37] investigates recovery of a signal from a channel modelled as known linear time-invariant system from nonuniform sampling of outputs. Reference [38] presents an approach based on Gabor time-frequency space.

The problem of finding the optimal placements of sensors is investigated in specific applications in several contexts including power systems and power delivery, robotics and automation and magnetics [39], [40], [41], [42]. In [43], the importance of a framework for the general signal reconstruction problem is emphasized. This research focuses on developing efficient methods for determining optimal combination of observations rather than on understanding information flow in measurement process in an abstract manner.

The measurement selection problem is extensively investigated in the framework of control theory with a special emphasis on controllability and observability [44], [45]. In [45] Fisher information matrix is used as a tool for understanding the nature of optimal measurement strategy problem.

In an information theory framework, sensors and the information content of the output of sensors is an important subject. This subject is investigated in the context of distributed sensing systems, noisy source coding, multi-terminal source coding, and the CEO problem. In these works the emphasis is on coding of observations. Several different scenarios are considered with a coding approach in [46], [47], [48], [49], [50]. Another related work in the information theory

framework is the subject of hypothesis testing. A hypothesis testing problem under communication constraints is investigated in [51]. This work is similar to the problem we have introduced in chapter 4 in the sense that the information retrieval problem under communication constraints is investigated. However in this problem the focus is on hypothesis testing, which is quite different from our problem, where the estimation of the unknown vector is considered.

Chapter 3

Preliminaries

3.1 Metrical Information

To understand the properties of metrical information provided by an observation, it is necessary to understand how a measurement is made. This section and the following sections present our understanding of a measurement and proposes a mathematical model. It also proposes a measure of cost for doing a measurement and exploits its relationship with the concept of number of distinguishable levels and information theory.

3.2 Metrical Information and Measurement Devices

This section discusses the relation between the metrical information in an observed variable and a measurement device. It also proposes a mathematical model for measurement devices.

As stated earlier metrical information is related to the measure of the uncertainty in the value of an observation. In an experiment which does not involve quantum effects, the basic source of uncertainty is the measurement device. With this in mind, we ignore the other sources of impreciseness in an obtained value and consider the relationship between the metrical aspect of a measurement and finite precision measurement devices.

When a physical quantity is measured, the result of the measurement is not exactly the true value of the observed variable. Very small changes in the original variable do not necessarily produce detectable output changes in the readings of a measurement device. Even when the exactly same value is measured a number of times, the measurement device will output different values concentrated around the actual value. This measurement error is unavoidable in the case of both analog and digital devices. In the case of an analog device the resolution of the analog display is an important source of uncertainty. In the case of a digital measurement device, intrinsic quantization in the digital display is an important source of finite-precision.

Hence, we see the act of performing a measurement as a process which adds uncertainty to a signal value. The following model is adopted:

$$s = g + m, \tag{3.1}$$

where g is the original value to be measured, m models the uncertainty introduced by the measurement device and s is the result of the measurement. With this point of view, the only distinctive property of a measurement is statistical properties of the noise introduced by it.

3.3 Cost of Doing a Measurement

In our model, a cost is associated with every measurement. We propose the following function for the cost of a measurement

$$C = \log\left(\frac{\sigma_s^2}{\sigma_m^2}\right) = \log\left(1 + \frac{\sigma_g^2}{\sigma_m^2}\right), \tag{3.2}$$

where σ_s^2 is the variance of the observation and σ_m^2 is the variance of the noise introduced by the measurement device. In writing these equations, it is assumed that original value to be measured and noise introduced by the measurement device are uncorrelated.

C can be interpreted as a measure of the number of distinguishable levels which can be resolved by the measurement device. Moreover, this definition can be seen as a measure of the information transfer capacity of the measurement channel. These interpretations are discussed in sections 3.4 and 3.5.

3.4 Discussion of the Model

This section discusses the plausibility of the proposed model for the measurement devices and the cost associated with a measurement.

3.4.1 Number of Distinguishable Levels

Most of the measurement devices are characterized by their accuracy, i.e. the number of input levels that they can distinguish. A measurement device, digital or analog, effectively has finite number of distinguishable output levels. Another characteristic of the measurement devices is scalability of their ranges. That is, once you buy a measurement device you can arrange it to different ranges and

use it to measure variables with different ranges. We want our model to reflect these properties of physical measurement devices as much as possible.

Let a measurement device have one distinctive property: the number of input levels that it can distinguish, i.e. its dynamic range. It is assumed that the total range R of a measurement device can be adjusted. For example, let our device be able go distinguish 10 levels. We assume we can freely use it to measure a range of 100 Volts with 10 Volts accuracy or a range of 10 Volts with 1 Volt accuracy.

Assume that the measurement error introduced by a measurement device can be modelled as Gaussian additive noise. Our assumption of range scalability implies that the variance of this additive noise should change as the range R of the measurement device is scaled.

To illustrate this idea, let us consider a Gaussian random variable with a known variance σ^2 to be quantized with uniform quantization. Let the number of quantization intervals be N_{q}. To have the minimum mean-square error (MMSE) between the quantized variable and the original continuous variable, there is a best quantization interval Δ for each N_{q} for a given σ^2 [52]. The range covered by this quantization is given as $\Delta \times N_{\mathrm{q}}$. It is possible to plot curves of N_{q} versus Δ for different σ^2 values. A figure illustrating this idea is given in Figure A.1 in Appendix A. Within this scheme Δ can be considered as a measure of uncertainty in each quantized variable. This interpretation may seem implausible for the values near the ends of the ranges, but these values are probably on the tails of the distribution, hence they are already unlikely.

We associate a measurement scenario with this quantization scenario as follows: We consider a digital measurement device with number of distinguishable levels N_{q}. This device will measure the value of a Gaussian random variable with a known variance. The device is arranged to a range of $\Delta \times N_{\mathrm{q}}$ where Δ is determined by the plot of N_{q} versus the best Δ for this particular variance.

Numerical results show that for a given number of distinguishable levels the ratio of $\sigma^2$to Δ^2 is roughly constant and increases with increasing N_{q}. Figure A.2 in Appendix A illustrates this observation. Since in the quantization scheme Δ is considered as a measure of uncertainty on the quantized variable and in the measurement scheme a noise is associated with uncertainty of measured values, Δ can be interpreted as a characteristic of the uncertainty of the noise introduced by the measurement device. Since it is the ratio of σ^2 to Δ^2 which is constant, Δ is interpreted as the standard deviation of the measurement noise. This observation suggests using ratio of variance of the observed signal to variance of the noise as a measure of the number of distinguishable levels for a measurement device.

Another observation can be made by considering measurement of a uniform random variable which is in the range $[-R/2, R/2]$ [53]. Consider a device which outputs the sum of the value of this variable with a noise term that is uniform in the range $[-\Delta/2, +\Delta/2]$. Then the output will be in the range $[-(\Delta + R)/2$, $+ (\Delta + R)/2]$ and the number of distinguishable levels in the output will be given by $(\Delta + R)/\Delta = 1 + R/\Delta$. This is the output range divided by the range of the noise term. This idea is illustrated in Figure 3.1. Comparison of this result with the argument of the log term in equation 3.2 is instructive in understanding the general form of $1 + \cdots$. This observation supports using the ratio of observed signal's characteristics to noise characteristics rather than the ratio of original signal's characteristics to noise characteristics as a measure of number of distinguishable levels. For a Gaussian random variable, range is thought to be proportional to standard deviation. Hence it is plausible to use the ratio of standard deviation of output to standard deviation of noise as a measure of number of distinguishable levels in the case of Gaussian random variables.

With this observation and the motivation supplied by the mentioned simulations, let ρ be defined as a measure of the number of distinguishable levels of a

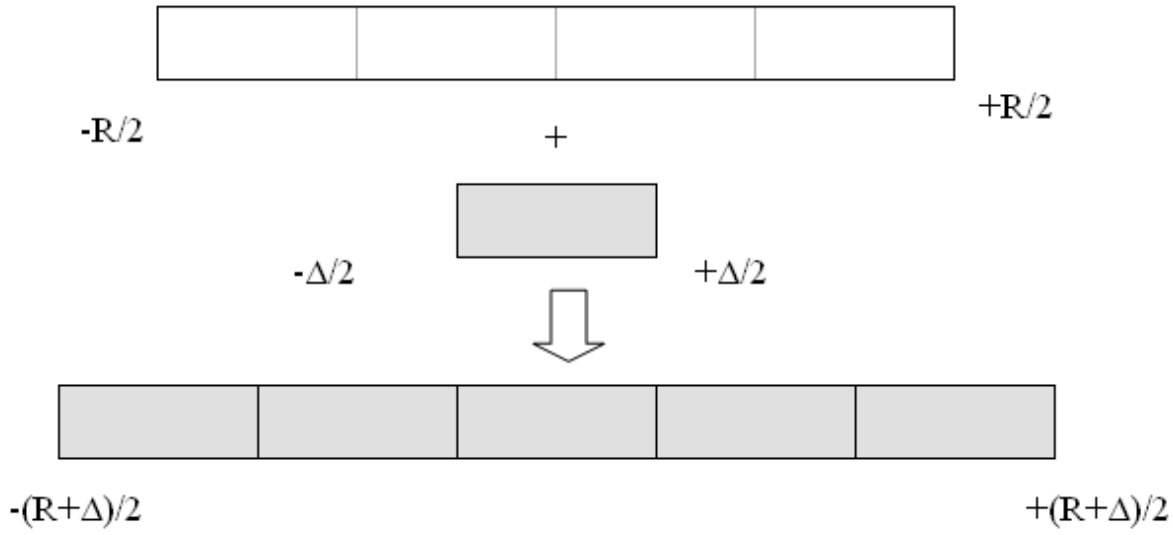

Figure 3.1: Illustration of how the number of distinguishable levels is obtained when an uncertainty is added to a signal

measurement device

$$\rho = \varrho \frac{\sigma_s^2}{\sigma_m^2}, \tag{3.3}$$

where ϱ is a positive constant. With an abuse of notation we use the term "number of distinguishable levels" also for ρ.

3.4.2 Cost Function

Let the number of distinguishable levels of a measurement device be ρ. Let the cost of using this device for one measurement be given by a function $C(\rho)$. Since we assume range of a measurement device can be scaled freely according to need, cost of using a device doesn't depend on the range it is adjusted to.

Observations on the plausible cost function

Before defining a cost function, we investigate the properties the plausible cost function should have.

I. Each measurement should have a nonnegative cost:

$$C(\cdot) \geq 0. \tag{3.4}$$

II. Cost function should be an increasing function of ρ.

III. Cost of using a measurement device with $\rho = 1$, i.e. measuring with one level should be 0.

IV. The plausible cost function of using a measurement device C should be a function of the number of distinguishable levels $C(\rho)$ such that

$$m \times C(\rho) \geq C(\rho^m), \tag{3.5}$$

where m is the number of usages of the measurement device [53]. If this inequality is not satisfied, there will be no point in having measurement devices with large number of distinguishable levels. This inequality guarantees that using a measurement device with large number of distinguishable levels is at least as economical as using a measurement device with a smaller number of distinguishable levels repeatedly to effectively measure the same number of distinguishable levels.

V. Since doing a measurement with a noise with infinite variance doesn't provide us any information, it should have zero cost.

VI. Doing a measurement without noise or with noise that only introduces a bias term (i.e. noise with zero variance) should have an infinite cost. This is because the original value of the output vector can be recovered perfectly in this case.

VII. Measuring a deterministic signal, i.e. a signal with zero variance, should have zero cost.

The fact that the logarithm function takes products to sums and satisfies equation 3.5 motivates the usage of a logarithm function in the definition of our cost function. Therefore the following function is proposed for the cost function:

$$C(\rho) = K \log(K_1 \rho). \tag{3.6}$$

Here K and K_1 are positive constants. K_1 is chosen to be positive so that the logarithm function is defined. K is chosen to be positive to be consistent with item I.

This cost function satisfies equation 3.5 for all $K_1 \geq 1$. The only K_1 satisfying strict equality is 1. To satisfy item III, K_1 is chosen as 1. Since we assume that the range of the measurement devices can be adjusted, it is possible to choose the ranges a measurement device is adjusted in a clever way to obtain more accurate measurements. That is, a measurement device with ρ levels can be used to distinguish between more than ρ levels by using the device more than once at the same measurement. Suppose we first use the measurement device adjusted to the range R_1 to determine the most significant figure. Then we can change the range to $R_2 = R_1/\rho$ and do the same measurement to determine the second significant figure. Then the cost of these measurements are $K \log(\rho) + K \log(\rho) = 2K \log(\rho)$. The same measurement could also be done with this accuracy by using a measurement device with ρ^2 levels only once. This measurement has the cost of $K \log(\rho^2)$ which is the same as the cost of using the the previous method.

The constant ϱ in the definition of number of distinguishable levels can be changed with a different value without violating the inequality stated in equation (3.5). These constraints do not force K and ϱ to have specific values. This observation implies an arbitrariness in the constants in the definitions of number of distinguishable levels and the cost function as far as our observations on the plausible cost function is considered. As a result, we arbitrarily choose K and ϱ as 1. Hence the cost associated using a measurement device with number of

distinguishable levels equal to ρ is defined as

$$C(\rho) = \log(\rho), \tag{3.7}$$

where the base of the logarithm determines the unit of cost. With this definition item V is consistent with choosing the definition of number of distinguishable levels as in equation 3.3 instead of $\rho = \varrho\frac{\sigma_g^2}{\sigma_m^2}$. While we can not strictly claim this definition is unique, it is fully consistent with our observations, and seems the most plausible choice.

3.5 Connections to Information Theory

This section exploits the links between our definition of cost function and some information-theoretic concepts.

3.5.1 Channel Capacity

Modelling the measurements as a process that causes uncertainty in the measured values by introducing an additive noise implies a channel capacity interpretation for our problem. A measurement is seen as a noisy channel whose input is a sample of the signal to be observed and whose output is the observation. All definitions of standard information-theoretic concepts in this section are adopted from [54].

For a channel the most natural parameter to consider is its capacity. Channel capacity can be informally defined as the maximum rate such that the message at one side of the channel can be transmitted over the channel with so that the original message is reconstructed on the other side of the channel with high

probability. Capacity of a channel with input x and output y is given by

$$R_c = \max_{p(x)} I(x;y) \tag{3.8}$$

where $I(x;y)$ is the mutual information between x and y. More information on mutual information is given in section 3.5.2.

A channel with Gaussian additive noise is called a Gaussian channel. The channel capacity of a Gaussian channel with noise variance σ_z^2 and power constraint p on the input is given by

$$R_c = \max_{E[(x)^2] \leq p} I(x;y) = 0.5 \log(1 + p/\sigma_z^2). \tag{3.9}$$

This capacity is achieved when x is Gaussian distributed with zero-mean and variance p.

Our choice of the cost function—with a scaling—is the same as the channel capacity for a channel which has the input g with $E[(g)^2] \leq \sigma_g^2$ and output s:

$$0.5 \times C = 0.5 \log \left(1 + \frac{\sigma_g^2}{\sigma_m^2} \right) = R_c. \tag{3.10}$$

If the input is assumed to be zero-mean Gaussian distributed with variance σ_g^2, this capacity is achieved with the input distribution we have. For any other distribution, $I(g;s)$ will be smaller. Hence the Gaussian distribution is the distribution which gives the maximum cost among the distributions satisfying the power constraint $E[(g)^2] \leq \sigma_g^2$.

In a channel, the sender side sends an input signal, the channel distorts this signal, and the receiver side tries to decide what was the original signal sent. If this decision is successful, then the channel can be said to transmit some information. Hence channel capacity can interpreted as a measure of the

maximum number of distinguishable inputs. For a random variable, the number of distinguishable inputs means different amplitude levels. Hence number of distinguishable inputs for a channel, i.e. the capacity of a channel and the number of distinguishable levels at the output are closely related. This close connection may be interpreted as a guideline for the definition of the plausible cost function. Hence the 0.5 scaling factor in front of R_c may be implying that it is more natural to use the ratio of standard deviations instead of variances as a measure of number of distinguishable levels.

3.5.2 Mutual Information

Mutual information between two random vectors x and y is denoted by $I(x;y)$ and is given by

$$I(x;y) = h(x) - h(x/y) = h(y) - h(y/x), \tag{3.11}$$

where $h(x)$ and $h(x/y)$ are entropy and conditional entropy as defined in [54].

We have seen that when g and m is Gaussian

$$0.5 \times C = I(g;s) = h(s) - h(s/g), \tag{3.12}$$

where $I(g;s)$ is the mutual information between g and s. Mutual information is the measure of the average reduction in the uncertainty of one random variable due to knowledge of another. With this concept, it is possible to talk about the amount of information one random variable contains about another. The fact that the cost of a measurement is the same as the mutual information between observed random variable and result of measurement—up to a scaling factor—implies an inherent connection between the cost of doing a measurement and information provided by it.

Chapter 4

Metrical Information

This section presents one of the possible problem formulations which can be used to exploit the relationship between the metrical information content of the input signal and the output signal. To isolate the metrical information problem from the structural information problem, we assume the structural sampling strategy is fixed and formulate our problem in a discrete framework. We discuss the best measurement strategy while measurements are done with measurement devices with different number of distinguishable levels.

4.1 System Model and Notation

This section presents the system model and notation for the case where the problem is investigated in a discrete framework. In this framework we assume the structural sampling strategies are chosen appropriately and samples of the input signal can be used as a good representation of the input signal.

The actual values of the vector that will be measured is formed according to the linear system of the vector form

$$g = Hf + n, \tag{4.1}$$

where f is the unknown vector we want to obtain information about, n is the inherent system noise and g is the vector that we attempt to observe.

These values will be measured by some measurement devices. We consider the model

$$\begin{aligned} s &= g + m & (4.2)\\ &= Hf + n + m, & (4.3) \end{aligned}$$

where m models the uncertainty introduced by the measurement devices.

In this system f,g,n,m are column vectors with N,M,M,M elements respectively. H is an $M \times N$ matrix. We assume f,n,m are Gaussian distributed with zero mean and covariance matrices of f and n are known. We also assume f,n and m are independent. The covariance matrices of n and m are diagonal. (Since the measurement devices may be calibrated if they are biased, it is reasonable to assume that m is a zero-mean random vector.)

The mean of a random vector is given by $\bar{x} = E[x]$. The covariance matrix for a variable x is given by $K_x = E[(x - \bar{x})(x - \bar{x})^\dagger]$ where $\dagger$ denotes the transpose.

4.2 Preliminaries

In this vector model, we see the measurement process as a sum of M parallel independent measurement channels as illustrated in Figure 4.1.

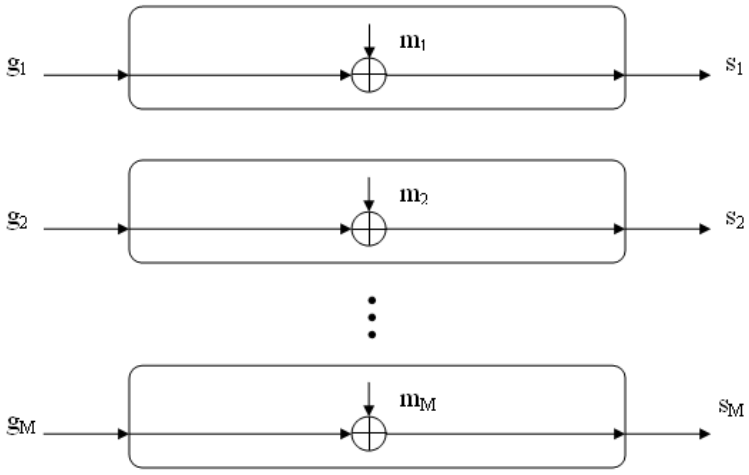

Figure 4.1: Measurement process as M parallel independent channels

The total measurement cost is given as the sum of the cost of each measurement

$$C = \sum_{i=1}^{M} C_i. \tag{4.4}$$

As the figure and the definition of the cost function suggests the problem at hand can be interpreted as an estimation problem under a communication cost constraint.

4.3 Problem Formulation—MMSE Estimation

As a criteria for information supplied by a measurement, the mean-square error when f is estimated from s by MMSE estimation method is used. Since f and s are jointly Gaussian, the MMSE estimate of f is equal to the MAP estimate of f given s. We would like to learn the trade-off between the cost of the measurements and the information gained. This problem can be formulated as a vector optimization problem as

$$\min_{\sigma_m} \text{ (with respect to } R_+^2) \; X(\sigma_m), \tag{4.5}$$

where

$$\sigma_m = [\sigma_{m_1} \ ... \ \sigma_{m_i} \ ... \ \sigma_{m_M}]^\dagger, \tag{4.6}$$

$$X(\sigma_m) = [\mathrm{tr}\,(\mathrm{K}_\varepsilon)\,, \sum_{\mathrm{i}=1}^{\mathrm{M}} \mathrm{C_i}], \tag{4.7}$$

$$\sigma_{m_i}^2 \geq 0, \quad i = 1, ..., M \tag{4.8}$$

$$\varepsilon = f - \hat{f}, \tag{4.9}$$

$$\hat{f} = E[f \mid s] = K_f H^\dagger K_s^{-1} s, \tag{4.10}$$

$$K_\varepsilon = E[K_{f/s}] = K_f - K_f H^\dagger K_s^{-1} H K_f, \tag{4.11}$$

$$K_s = H K_f H^\dagger + K_n + K_m, \tag{4.12}$$

$$K_m = \mathrm{diag}(\sigma_{\mathrm{m_i}}^2), \tag{4.13}$$

$$C_i = \log\left(\frac{\sigma_{s_i}^2}{\sigma_{m_i}^2}\right). \tag{4.14}$$

In this formulation R_+^2 denotes the non-negative orthant. This is an optimization problem with two objectives to be minimized. One is the MMSE and the other is the cost of measurements. We would like to minimize both, but there is no unique way to convert the benefit of low MMSE and low cost to each other. Hence what we investigate is the Pareto optimal points. A point is Pareto optimal if there is no other solution that performs at least as well on both criteria and strictly better on at least one criterion.

By applying scalarization, which is a standard technique for finding Pareto optimal points of a vector optimization problem we arrive at the scalar problem

$$\min_{\sigma_m} \lambda^\dagger X(\sigma_m), \tag{4.15}$$

where the variables are as defined below. Here $\lambda \succ 0$, where $\succ$ denotes componentwise strict inequality. For different values of λ, different Pareto optimal solutions of the vector optimization problem 4.5 is found.

A closely related problem is the problem of minimizing the MMSE for a given cost

$$\min_{\sigma_m} \operatorname{tr}(\mathrm{K}_\varepsilon) \tag{4.16}$$

such that

$$\sum_{i=1}^{M} C_i \leq C_{\max}. \tag{4.17}$$

The Lagrangian of this problem is the same as the scalarization of the vector optimization problem. The problem of minimizing the cost for a given MMSE also has the same Lagrangian. This problem is similar to the source coding problem in which a random vector is to be represented with the minimum finite number of bits under a distortion criterion.

The following are some observations on the range of our analysis:

- Our approach can handle the case where some of the measurements are not done at all. Hence we can determine which of the available measurements should be done in order to have a good estimate of the original vector. We know that any measurement with infinite noise variance will be effectively of no use and is not to have been done in the first place. With our definition, these types of measurements have zero cost.Thus measurements which are not worth doing will appear as measurements with infinite noise variance at the outcome of optimization procedure.

- Our approach can handle the case where there are repeated observations. Whenever a specific measurement is repeated with different measurement noises yielding a particular MMSE, the equivalent noise that will yield the same MMSE with only one measurement has a lower cost. That is, doing

one measurement always has a lower cost than repeating the measurements to achieve a particular MMSE. Hence if an optimum noise is found for a particular measurement, it is guaranteed that there is no better solution which takes into account the possibility that observations can be repeated. Details are given in Appendix B.

The fact that the optimal solutions for the noisy source representation problem and the channel capacity problem comes from transforming the vectors into appropriate domains motivates an approach which focuses on finding a suitable transformation for our problem. We expect that this suitable transformation will concentrate the information in independent coordinates and the useful information will be easily distinguished. One disadvantage of this type of approach comes from the fact in our scheme it is not possible to change the original vector before measurement, whereas while quantizing a vector or sending a message through a channel, it is possible to alter them before they face information losing effect of being quantized or being sent over a channel. Secondly, it may not be possible to transform back to the original domain in a meaningful way even if the solution is found in the transformed domain. For instance, if the optimal solution requires us not to measure some of the components of the transformed vector, this solution would not be expressible in the original domain for most of the cases, because the number of the components of the original vector that the component which were decided not to be measured in the diagonal domain transforms back to will be more than one. In the light of these observations, one may consider using a linear map before measurement to transform the vector to be observed into an appropriate domain. It is expected that if the cost of using a linear map is assumed to be zero, this approach will yield better performance for a given cost constraint compared to the one that directly observes the original vector.

Another plausible measurement scenario is the case where the set of available measurement devices are explicitly stated. This corresponds to the case where there are only certain devices with certain accuracies available to us. Here the noise variances of available measurement devices will be given and we will try to find the best assignment of measurement devices to measurements. The problem formulation will be the same except that the cost limit will be changed to $\sigma_{m_i} \in \mathcal{V}$ where the set $\mathcal{V}$ denotes the set of available measurement devices.

The problems presented in this section may be also interpreted in an experiment design framework. In this framework the goal of the problem will be to choose the measurements in a way that an error and/or cost criteria is satisfied.

4.3.1 Illustrative Examples

In this section some simple examples which illustrate the solution of the problem presented in section 4.3 for some special cases are presented.

Throughout this section k_{xij} denotes the ith row jth column of K_x matrix. We use the notation $\bar{k_x}$ for the vector $[k_{x11} \ \ldots \ k_{xii} \ \ldots \ k_{xMM}]^\dagger$. For convenience we assume the notation log represents the natural logarithm throughout this section.

1-Dimensional case

When a single random variable ($N = 1$) is to be measured once ($M = 1$), we refer to this case as 1-Dimensional case. Since there is only one measurement, the problem of finding the best measurement strategy such that the best error and cost is obtained is not meaningful. However there is still a curve which shows the trade-off between the cost and the error. In this case the matrices $H,\ K_f,\ K_n,\ K_m$ become the one-dimensional scaling factor h and the variances

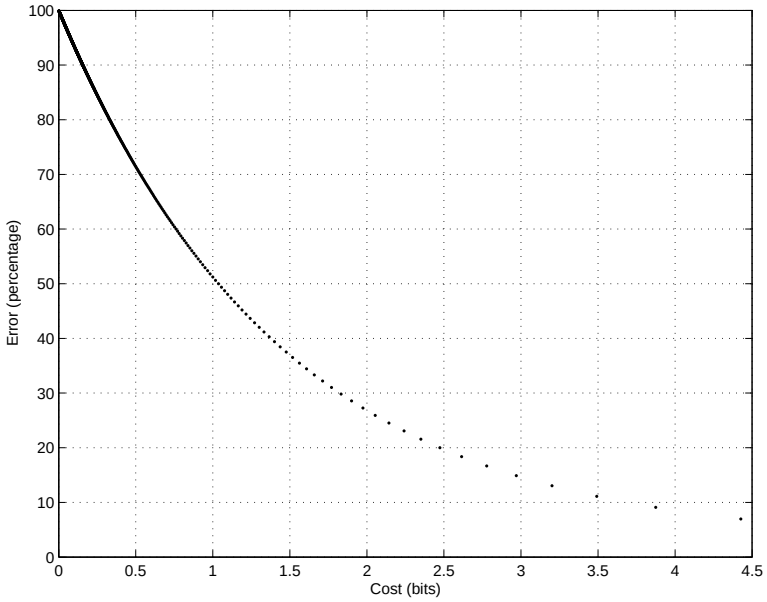

Figure 4.2: Error versus cost curve for the 1-dimensional case

$k_f,\ k_n,\ k_m$. For this case the error d is given as

$$d = kf - \frac{h^2 k_f^2}{h^2 k_f + k_n + k_m}. \tag{4.18}$$

The cost is given as

$$C = \log(1 + \frac{h^2 k_f + k_n}{k_m}). \tag{4.19}$$

The trade-off curve for $h = 2,\ k_f = 1,\ k_n = 0.1$ is given in Figure 4.2 as an illustrative example. Here percentage error is calculated as $\frac{d}{k_f}$. The only cost-error pairs which are achievable are the ones on the curve.

Diagonal Case

When the matrices $H,\ K_f,\ K_n$ are diagonal, we refer to this case as the diagonal case. For this case we look at the problem of the scalarization of the vector optimization problem:

$$\min_{\bar{k}_m} \sum_{i=1}^{M} \log\left(1 + \frac{k_{g_{ii}}}{k_{m_{ii}}},\right) + \nu \sum_{i=1}^{M} \left(k_{f_{ii}} - \frac{h_{ii}^2 {k_f}_{ii}^2}{h_{ii}^2 k_{f_{ii}} + k_{n_{ii}} + k_{m_{ii}}}\right), \tag{4.20}$$

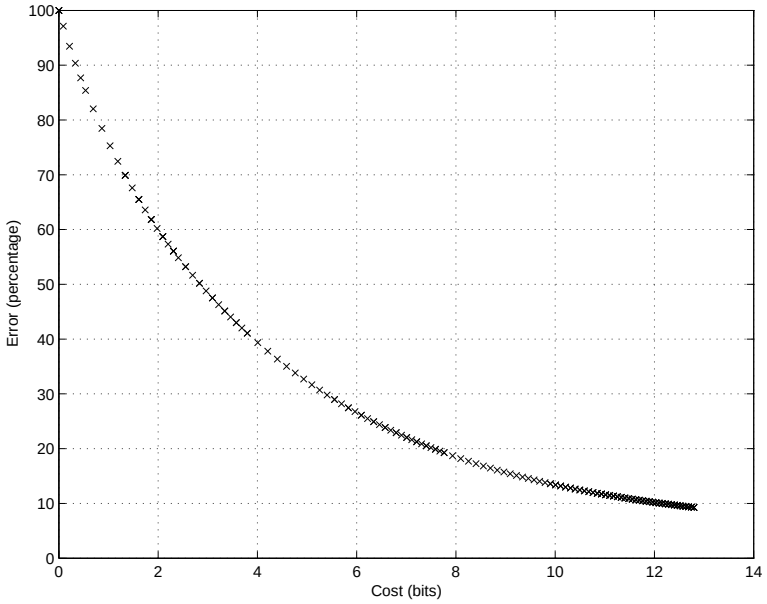

Figure 4.3: Error versus cost curve for the diagonal case

where ν is an nonnegative parameter. Taking derivative with respect to k_{mii} and equating it to zero gives the following optimal values for k_{mii}

$$k_{mii} = \begin{cases} \frac{{k_{g_{ii}}}^2}{\nu h_{ii}^2 k_{f_{ii}}^2 - k_{g_{ii}}} & \text{if } \nu h_{ii}^2 {k_{f_{ii}}}^2 - k_{g_{ii}} > 0 \\ \infty & \text{if } \nu h_{ii}^2 k_{f_{ii}}^2 - k_{g_{ii}} \leq 0. \end{cases} \tag{4.21}$$

The trade-off curve for the diagonal case can be obtained by varying the parameter ν. For high values of error, ν will be small. Here we see that for some cases it would be better not to do some of the measurements. This case occurs for high values of error, where it becomes unnecessary to measure every component of the vector.

The trade-off curve for the diagonal case is illustrated in Figure 4.3. While generating this trade-off curve H,K_f and K_m are taken to be diag($[2.4782,\ 1.6749,\ 2.5185]^\dagger$), diag($[0.98038,\ 0.716,\ 0.36592]^\dagger$) and diag($[0.075862,\ 0.19519, 0.14654]^\dagger$) respectively. We look at some of the cost-error pairs on the graph. For example, to obtain an error of 69.90%, the optimal method is to measure the first two components of the vector with noise variances

6.5112 and 7.2217, and totally discard the third component. To obtain an error of 54.71%, the optimal allocation is to measure the three components of the vector with noise variances 3.4, 2.5 and 1561.1. To obtain an error of 8.35%, the optimal method is given by the following noise variances: 0.0372, 0.0884, and 0.0467. These examples show how the third component becomes important as a better performance in terms of error is desired.

4.4 Numerical Results

This section presents some numerical results illustrating the solution of the problem presented in section 4.3. These results show the trade-off between cost and error when we estimate f from s.

The system matrix H is taken to be the 8 by 8 real equivalent of the 4 by 4 complex FRT matrix of order $a = 0.4$. Since the present formulation has not yet been generalized to complex matrices, we have used the real equivalent of the complex matrix. The real equivalent of a complex matrix is formed as follows [55]:

We consider the complex matrix $\bar{A}$, and the linear system model

$$\bar{y} = \bar{A}\bar{x}, \tag{4.22}$$

where $\bar{x}$ and $\bar{y}$ are complex vectors and $\bar{A}$ is a complex matrix. Here $\bar{x} = x_r + jx_i$ with $j = \sqrt{-1}$, and x_r and x_i are real random vectors. We form the composite real vector corresponding to a complex vector $\bar{x}$ as $x = [x_r \; x_i]^\dagger$. We investigate whether a real valued matrix A satisfying the equivalent real model

$$y = Ax \tag{4.23}$$

exists. Such an A matrix exists and is given by

$$A = \begin{pmatrix} \bar{A}_r & -\bar{A}_i \\ \bar{A}_i & \bar{A}_r \end{pmatrix}, \tag{4.24}$$

where $\bar{A}_r = \Re\{\bar{A}\}$ and $\bar{A}_i = \Im\{\bar{A}\}$.

For the generation of FRT matrix, an implementation of algorithm presented in [56], [57] and in chapter 6 of [32] is used. This implementation is available at [58].

Our purpose is to obtain Pareto optimal curve between cost and error. Therefore, ideally we would find the minimum error for a given cost or the minimum cost for a given error. However since an efficient method for this could not be obtained, a brute-force method was employed. For this purpose, a maximum noise variance $\sigma^2_{\max}$ is selected. Then 8 different variances are formed uniformly on the range from 0 to $\sigma^2_{\max}$. These correspond to different numbers of distinguishable levels. One permutation of these 8 noise variances determine the noise of 8 measurement devices and constitute one measurement scenario. There are 8^8 different scenarios. For each scenario cost and error is determined and marked on the cost-error plane. Points formed with this procedure cover a particular range of cost. To be able to cover the whole cost-error plane, this procedure is repeated for four different values of $\sigma^2_{\max}$: $a \max(\sigma_g^2)$ where $a = 0.1, 1, 4$ and 100. An arbitrary covariance matrix is used for K_f. The results are shown in Figure 4.4. The percentage error is obtained by normalizing the error as follows:

$$100 \frac{\varepsilon}{\sum_{i=1}^{M} E[f^2]}. \tag{4.25}$$

To be able to display the huge amount of points (4×8^8), some of the points from each data set have been omitted on the plot. The points that are shown on the figure have been chosen so that the ones near the boundary are kept.

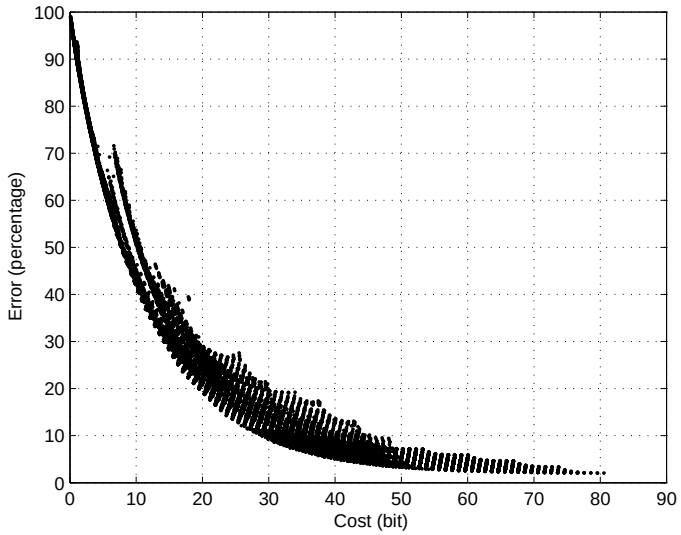

Figure 4.4: Illustration of the achievable region in the cost-error plane (8×8)

It is seen on the plot that the decrease in the percentage error is almost linear for the low values of cost. This linear relationship is so steep that it is relatively economical to decrease the error for the low values of cost. For instance halving the error from 100% to 50% costs less than 10 bits. With increasing cost the rate of error decrease becomes smaller. Reducing the error from 50% to 10% takes about 20 bits. After reaching the error of 10% with a cost about 30 bits, the rate of error decrease becomes much more smaller. The error that would be achieved if the measurements were perfect (corresponding to infinite cost) is 1.97%. The error for 60 bits is roughly near this level, and it takes 30 bits to reach that level from the error of 10%. Since after the cost of 60 bits the error decrease is very small, it would not be meaningful to increase the cost more than 60 bits for most of the cases.

For greater numbers of measurements, scanning the cost-error plane with this method becomes computationally infeasible because of memory requirements. For instance, for 16 measurements 16^{16} data points will be needed. Hence a different approach has been adopted for greater number of measurements. In this method again a maximum noise variance $\sigma^2_{\max}$ is determined. Then M

different variances are formed randomly on the range from 0 to $\sigma^2_{\max}$ according to a uniform distribution. Corresponding error and cost is found and marked on cost-error plane. This is done $(M/2)^{M/2}$ times. This randomly generated measurement noise variances cover a particular range of cost. To be able to cover the whole cost-error plane, this procedure is repeated for four different values of $\sigma^2_{\max}$: $a \max(\sigma^2_g)$ where $a = 0.1, 0.5, 2$ and 4. The 16×16 case is illustrated in Figure 4.5. Since this method is based on randomly generated noise variances, the trade-off curve obtained is less reliable than the previous method where the space of possible noises is scanned with a particular resolution. Nevertheless, even in the worst case, the curve obtained by this method provides an upper boundary for the actual trade-off curve.

This plot is similar to the plot of 8×8 case. Similar conclusions regarding to the rate of decrease of error with increasing cost can be drawn. In this case the error that would be achieved if the measurements were perfect (corresponding to infinite cost) is 13.09%. The error is roughly at this level for a cost of 100 bits. Hence for most of the applications, it would be unnecessary to pay more than 100 bits. Comparing the plots in Figure 4.4 and Figure 4.5 reveals that the cost required to reach the minimum error possible (100 bits) is roughly the twice of the 8×8 case (60 bits). This loose relationship is not surprising because the linear system represents the same transform (real equivalent of the FRT of order 0.4) for vectors of different length and the number of measurements for the current case (16) is twice the number of measurements for the previous case (8). Still, it should be emphasized that the trade-off curve heavily depends on the input probability distribution, and conclusions regarding to information transfer capability of a system cannot be drawn without reference to the input distribution.

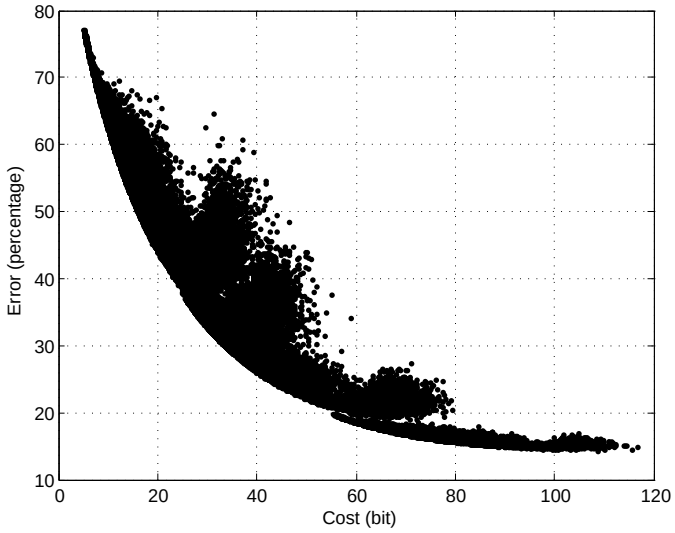

Figure 4.5: Illustration of the achievable region in the cost-error plane (16×16)

We underline that only the lower left boundary of these plots is meaningful and represents the trade-off between cost and error. The remaining points are merely outcomes of the brute-force method.

4.5 Precise Measurements Case

When the uncertainty introduced by the measurements are small with respect to the range of g, we name this case as the precise measurements case. This is the case where K_s is near K_g. Hence we use the first order approximation of the inverse of a positive definite symmetric matrix K [59]

$$(K + \Delta K)^{-1} \approx K^{-1} - K^{-1}\Delta K K^{-1}. \tag{4.26}$$

Then

$$K_s^{-1} \approx K_g^{-1} - K_g^{-1} K_m K_g^{-1} \tag{4.27}$$

and, using the linearity of the trace operator, the MMSE becomes

$$\mathrm{tr}\left(K_f - K_f H^\dagger K_g^{-1} H K_f\right) + \mathrm{tr}\left(K_f H^\dagger K_g^{-1} K_m K_g^{-1} H K_f\right). \tag{4.28}$$

The first part is the error that would be present even if the measurements were perfect. This error is unavoidable, since its origin is the information transfer capability of the physical phenomena itself. The second additive error component is due to the imperfect measurements. We denote these errors as D_s and D_m respectively.

Throughout this section k_{xij} denotes the ith row jth column of K_x matrix. We use the notation $\bar{k_x}$ for the vector $[k_{x11} \ \ldots \ k_{xii} \ \ldots \ k_{xMM}]^\dagger$. For convenience we assume the notation log represents the natural logarithm throughout this section.

D_m can be expressed as

$$\begin{aligned} \mathrm{tr}\left(K_f H^\dagger K_g^{-1} K_m K_g^{-1} H K_f\right) &= \mathrm{tr}\left(K_m K_g^{-1} H K_f K_f H^\dagger K_g^{-1}\right) && (4.29)\\ &= \mathrm{tr}\left(K_m P\right) && (4.30)\\ &= \sum_{i=1}^{M} k_{mii} p_{ii}, && (4.31) \end{aligned}$$

where p_{ii} is the diagonal elements of P. Here the last equality follows from the fact K_m is diagonal.

Then the cost minimization problem for a given error constraint can be formulated as follows:

$$\min_{\bar{k}_m} \sum_{i=1}^{M} \log\left(1 + \frac{k_{g_{ii}}}{k_{mii}}\right) \tag{4.32}$$

such that

$$\sum_{i=1}^{M} k_{mii} p_{ii} \leq D \tag{4.33}$$

$$\bar{k}_m \succeq 0. \tag{4.34}$$

$$\bar{k}_m = [k_{m11} \ \ldots \ k_{mij} \ \ldots \ k_{mMM}]^\dagger \tag{4.35}$$

$$k_{m_i} \geq 0 \quad i = 1, ..., M. \tag{4.36}$$

Here $D = D_t - D_s$ and $D \in \mathbb{R}_+$. When the total allowed error D_t is smaller than the error introduced by the process, the optimization problem has no solution. Here $\bar{k}_g \succeq 0$, since $k_{g_{ii}}$ are the variances of components of a random vector.

Following observations can be made on this optimization problem:

- This is a convex optimization problem. The constraint function on error is affine, hence a convex function of optimization variable $\bar{k}_m$. The non-negativeness constraint on components of $\bar{k}_m$ is linear, hence convex.

  The convexity of the objective function can be seen by restricting its domain to a line. We define $x(t) = x(z+tv)$ and consider $\bar{k}_m = z+tv$ where $z \in \mathbb{R}^m$ and $z + tv \geq 0$ We have

  $$x(t) = \sum_{i=1}^{M} \log\left(1 + \frac{k_{g_{ii}}}{z_i + tv_i}\right) \tag{4.37}$$

  Second derivative of $x(t)$ with respect to t is

  $$x^{''}(t) = \sum_{i=1}^{M} k_{g_{ii}} v_i^2 \frac{2z_i + 2tv_i + k_{g_{ii}}}{[(z_i + tv_i)^2 + k_{g_{ii}}(z_i + tv_i)]^2}. \tag{4.38}$$

  Since $x^{''}(t) \geq 0$ for $z + tv \geq 0$, the problem is convex.

- Since a $\bar{k}_m \in R_{++}$ can be found which satisfies the inequality constraints with strict inequalities, Slater's condition holds. Since the problem is convex, this implies strong duality.

- Since this is a convex problem with differentiable objective and constraint functions with strong duality, KKT conditions are necessary and sufficient

for optimality of both the dual and primal problem. Hence optimal value of $\bar{k}_m$ can be found by KKT conditions.

We introduce the Lagrange multipliers $\lambda \in \mathbb{R}^M$ for the inequality constraints $\bar{k}_m \succeq 0$ and $\nu \in \mathbb{R}$ for the error constraint. Then we have the KKT conditions

$$\bar{k}_m \succeq 0, \tag{4.39}$$

$$\lambda \succeq 0, \tag{4.40}$$

$$\nu \geq 0, \tag{4.41}$$

$$\lambda_i k_{mii} = 0, \quad i = 1, \ \ldots \ M \tag{4.42}$$

$$\nu \left(\sum_{i=1}^{M} k_{mii} p_{ii} - D \right) = 0, \tag{4.43}$$

$$\nabla \left(\sum_{i=1}^{M} \log \left(1 + \frac{k_{gii}}{k_{mii}} \right) \right) + \nu \nabla \left(\sum_{i=1}^{M} k_{mii} p_{ii} - D \right) - \lambda_i \nabla \left(k_{m_{ii}} \right) = 0. \tag{4.44}$$

Solution of these equations result in

$$k_{mii} = \frac{-k_{gii} + \sqrt{k_{gii}^{\,2} + \frac{4k_{gii}}{\nu p_{ii}}}}{2}, \tag{4.45}$$

where the parameter $\nu > 0$ and is chosen such that the total error is D, that is

$$\left(\sum_{i=1}^{M} k_{mii} p_{ii} - D \right) = 0 \tag{4.46}$$

is satisfied.

Although equation 4.45 gives the optimum k_m values for a particular level of error, it is possible to obtain the trade-off curve between cost and error with this equation. By changing the parameter ν, it is possible to satisfy the equation 4.46 for different D's and find the corresponding cost C and D_{t}. Then C and D_t will be a particular point on the curve. It is possible to find the other points on

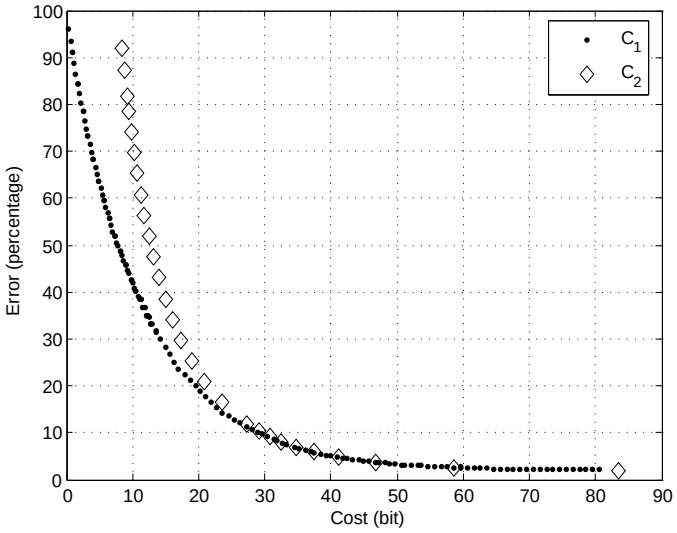

Figure 4.6: Comparison of cost-error curves (C_1: trade-off curve derived from the data presented in Figure 4.4. C_2: trade-off curve obtained for precise measurements case.)

the curve by varying ν. As ν goes to 0, k_{mii} goes to infinity and the curve for low values of cost, high values of error will be obtained. As ν goes to infinity, k_{mii} goes to 0 and the curve for high values of cost and low values of error will be obtained.

For a particular value of error, it is possible to solve the problem of precise measurements by finding the parameter ν. When ν is determined, optimum noise for each measurement device can be found by equation 4.45, and the corresponding cost can be found. By repeating this procedure for different values of error, it is possible to obtain the trade-off curve between cost and error. Figure 4.6 shows this curve for 8×8 case mentioned in section 4.4. To find the parameter ν, a trial and error approach has been adopted, positive-real numbers are tried with a step-size of 0.001. Then with equation 4.45 noise levels of measurement devices are determined. The corresponding error is compared with allowed error. When equation 4.46 is satisfied, the procedure is stopped. Error and cost are marked on cost-error plane. The points found by this method for different values of error are shown as C_1 on the plot. C_2 is the lower-left boundary derived from

the data presented in Figure 4.4. As expected, the solution from the precise measurements case approaches to the general solution for high values of cost.

4.6 Problem Formulation—Mutual Information

This section presents one of the possible problem formulations to understand how much information is transferred from the unknown vector to the observations when the measurements are done with finite precision. As the measure of information, we use the mutual information between the unknown vector f and the measured vector s. We want to find the best measurement strategy so that $I(f;s)$ is maximized when the total cost is limited.

This problem has an interesting channel capacity interpretation. The measurement process is modelled by parallel Gaussian channels. The total capacity of these channels are limited, which is given by the allowed cost. Hence this formulation aims to maximize the mutual information between the input and the output of a process when a part of it has a limited capacity.

This problem can be formulated as

$$\max_{\sigma_m} I\left(f;s\right) \tag{4.47}$$

such that

$$C = \sum_{i=1}^{M} C_i \leq C_{max}, \tag{4.48}$$

where

$$\sigma_m = [\sigma_{m_1} \ \ldots \ \sigma_{m_i} \ \ldots \ \sigma_{m_M}]^{\dagger}, \tag{4.49}$$

$$K_m = \mathrm{diag}(\sigma^2_{\mathrm{m_i}}), \tag{4.50}$$

$$C_i = \log\left(\frac{\sigma_{s_i}^2}{\sigma_{m_i}^2}\right), \tag{4.51}$$

$$I\left(f;s\right) = 0.5(\log\left|\det(\mathrm{K_s})\right| - \log\left|\det(\mathrm{K_n} + \mathrm{K_m})\right|). \tag{4.52}$$

A vector optimization problem related to this problem can be also formulated as in Section 4.3.

4.7 Remarks

The definitions of information-theoretic concepts in this section are adopted from [54].

Conditional mutual information: The conditional mutual information of random variables x and y given z is defined by

$$I(x;y/z) = h(x/z) - h(x/y,z). \tag{4.53}$$

Chain Rule for mutual information: Mutual information between two vectors can be written as the sum of conditional mutual information expressions.

$$\begin{aligned}
I(x;y) &= I(x;y_1,\ \ldots\ ,y_m) && (4.54)\\
&= \sum_{i=1}^{M} I(x;y_i/y_1,\ \ldots\ ,y_{i-1}) && (4.55)\\
&= I(x;y_1) + \sum_{i=2}^{M} I(x;y_i/y_1,\ \ldots\ ,y_{i-1}) && (4.56)\\
&= I(x;y_1) + I(x;y_2/y_1) + \sum_{i=3}^{M} I(x;y_i/y_1,\ \ldots\ ,y_{i-1}) && (4.57)\\
&= I(x;y_1) + I(x;y_2/y_1) +,\ \ldots\ ,+I(x;y_M/y_1,\ \ldots\ ,y_{M-1}) && (4.58)
\end{aligned}$$

Hence mutual information between the unknown vector f and the observation vector s can be written in terms mutual information between f and components of s. Hence the total information given by an observation about an unknown vector can be written as the sum of the information given by the first component plus the information given by the second component when the first component is given plus the information given by the third component when the first and the second components are known and so on up to the information given by the last component when all the other components are given.

4.8 Numerical Results

This section presents some numerical results illustrating the solution of the problem presented in section 4.6. These results show the trade-off between the cost of the measurements and the mutual information between the observed vector and the unknown vector.

The approach adopted to obtain the numerical results and H, K_f and K_n are the same with section 4.4. Throughout this section we use the term mutual information for the expression in equation 4.52 without the 0.5 scaling factor.

Figure 4.7 gives the results for the 8×8 case. The curve shows an almost linear relationship with a slope of one for low values of cost. This observation can be explained as follows: Using Hadamard's inequality on equation 4.52, the following relationship is obtained between $I(f;s)$ and total cost

$$|\det(\mathrm{K_s}(\mathrm{K_n}+\mathrm{K_m})^{-1})| \leq \prod_{\mathrm{i}=1}^{\mathrm{M}} \left(1+\frac{\sigma_{\mathrm{g_i}}^2-\sigma_{\mathrm{n_i}}^2}{\sigma_{\mathrm{n_i}}^2+\sigma_{\mathrm{m_i}}^2}\right) \leq \prod_{\mathrm{i}=1}^{\mathrm{M}} \left(1+\frac{\sigma_{\mathrm{g_i}}^2}{\sigma_{\mathrm{m_i}}^2}\right), \qquad (4.59)$$

where the left inequality holds with equality if only if the matrix $K_s(K_n+K_m)^{-1}$ is diagonal. Here the total cost constraint provides an upper limit for $I(f;s)$. For the cases of low cost, $\sigma_{m_i}^2$ is high compared to $\sigma_{g_i}^2$, and $K_s(K_n+K_m)^{-1}$ will

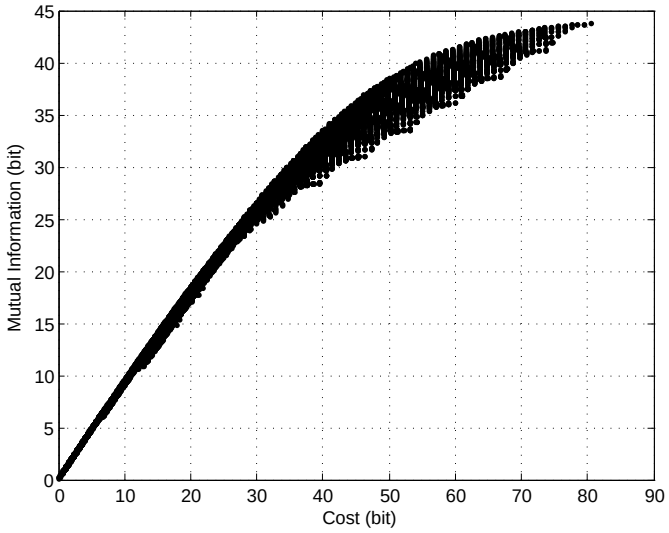

Figure 4.7: Illustration of the achievable region in the mutual information-cost plane (8×8)

be characterized by the diagonal matrix $K_m + K_n$. Hence mutual information will be close to the upper limit. For higher values of cost, $\sigma^2_{m_i}$ is smaller, mutual information moves away from the upper limit. As the allowed cost is increased, the rate of increase in mutual information decreases. For high values of cost, the curve approaches to 44.25 bits, which is the mutual information that would be achieved if the measurements were perfect (corresponding to infinite cost).

Figure 4.8 gives the results for the 16×16 case. For the 16×16 case, the procedure is repeated for three different values of $\sigma^2_{\max}$: $a \max(\sigma^2_g)$ where $a = 0.1, 0.5, 4$. This time the linear relationship for low values of cost does not show itself as clear as the previous case. Yet the curve stays below this linear bound and moves away from it as the cost increases. In this case the mutual information that would be achieved if the measurements were perfect (corresponding to infinite cost) is 39.75 bits.

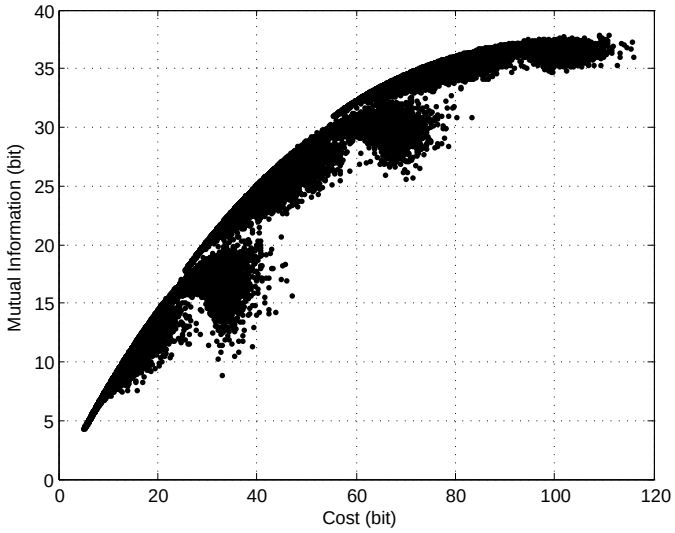

Figure 4.8: Illustration of the achievable region in the mutual information-cost plane (16×16)

Chapter 5

Structural Information

In this chapter we present a framework within which the relationship between the structural information content of the input signal and the output signal can be exploited. Throughout this section, we assume that the process is completely known.

5.1 Introduction

Although the information we want to extract about the input signal f may be different for different applications, we will assume that the basic aim is to recover f. Then the measure of success of our measurement strategy will be the quality of the recovered signal $\hat{f}$. Quality of the samples of the input signal or the output signal will be important to the extent they contribute to this recovery process. Otherwise they are only intermediate variables.

Structural information is related to the organization of the measurements in space. That is the spatial scheme of representing a signal in space. By organization of the samples, we mean the relative positions of the intended sampling

points. Sampling structure may be uniform or irregular. It may cover a part of the signal or all of it.

Although each measurement device is used to obtain the signal value at a particular point, in practice this may be impossible due to the imperfectness of measurement devices. The mechanism for putting sensors at given locations may not be precise enough which may result in an uncertainty on true sampling locations. Furthermore, the probes of the measurement device may be insensitive such that the measured value is the result of some averaging procedure around the intended measurement position. A review of these sampling issues are given in [60].

For a large class of averaging functions, a bandlimited signal is uniquely determined by the local averages by taking more dense samples required by the Nyquist rate. In this case, the required rate is determined by the band of the signal and the extension of averaging functions in the space [61]. Hence while associating a cost with a measurement made with such a device, the intimate relationship with cost of taking more dense samples should be preserved.

5.2 A Simplified Problem Formulation

This section presents a simplified formulation which still captures some of the basic properties of doing measurements in a continuous space.

We assume that measurements are perfect with respect to their features related to metrical information and structural information. That is, every measurement is done at the intended point of measurement without any averaging effect or jitter or any uncertainty on the measured value.

We assume a uniform sampling strategy. There are three important parameters related to a uniform sampling strategy: the extension of the space where the

samples are in, the sampling interval and the number of samples. Two of these parameters are independent, whereas one of them is a function of the other two. Since the number of samples is associated with the degrees of freedom of a set of signals, it is natural to choose it as one of the principal parameters.

We assume that the cost of a measurement strategy depends on the number of samples taken. We want to sample the output signal with a minimum cost as much as possible and obtain the best performance in terms of recovered continuous signal.

5.3 Stochastic Framework

We assume that the random process $f(x)$ is input to the linear system. The output is denoted by $L\{f(x)\}$. $n(x)$ is the random process denoting the process noise.

$$g(x) = L\{f(x)\} + n(x) \tag{5.1}$$

We assume that f and n are zero mean random vectors with known correlation functions. Since the processes are zero mean, correlation functions are the same with the covariance functions.

Throughout this section, we assume that any random process we consider is zero-mean. This assumption is not restrictive, since if a random process $y(x)$ under consideration is not zero-mean, it is possible to define the zero-mean random process $y(x) - \eta_y(x)$, where $\eta_y(x)$ is the mean function of the random process, and treat the mean and the centered random process separately.

5.3.1 Extension of a Random Process

We assume that the extension of a random process f is determined by its autocorrelation function.

We assume that the extension in space domain is determined by the variance function of the random process. We assume that there exists a Δu such that $R_f(x, x)$ is negligible for the values of satisfying $|x| > \Delta u/2$. For the values of $|x| > \Delta u/2$, $f(x)$ will not vary too much and hence can be treated as if it is deterministic and known. We take Δu as the extension of f in space domain.

The extension in frequency domain is determined by the 2-dimensional Fourier transform of the autocorrelation function of the random process:

$$S_f(v, v') = \int \int R_f(x, x') exp(-2\pi j(vx - v'x'))dxdx'. \tag{5.2}$$

If f is wide-sense stationary, $S_f(v, v')$ reduces to $S_f(v)\delta_f(v - v')$ where δ is the Dirac delta function and $S_f(v)$ is the usual power spectral density of f. For a random process, stationary or not, we look at $S_f(v, v')$ at $v = v'$. We assume that there exists a $\Delta\mu$ such that $S_f(v, v')$ is negligible for the values of satisfying $|v| > \Delta\mu/2$. We take $\Delta\mu$ as the extension in frequency domain. This definition is consistent with the definition of frequency extent for bandlimited wide-sense stationary (WSS) random processes.

5.3.2 Problem Formulation

We observe M samples of $g(x)$. Given $\bar{g} = [g(y_1) \ ... \ g(y_i) \ ... \ g(y_M)]^T = [g_1 \ ... \ g_i \ ... \ g_M]^T$ we want to recover $f(x)$ for a particular x. We denote the value to be estimated as f_x and the estimate $\hat{f}(x)$ or as $\hat{f}_x$. We restrict ourselves

to the affine estimators:

$$\hat{f}(x) = \sum_{i=1}^{M} a_{x,i} g_i + b_x \tag{5.3}$$

$$= a^T \bar{g} + b_x \tag{5.4}$$

By the orthogonality of the error $f(x) - \hat{f}(x)$ to the observation vector $\bar{g}$, the parameters of the estimator that minimizes the MMSE error is given as [62]

$$a_x^T = k_{f_x \bar{g}} K_{\bar{g}\bar{g}}^{-1} \tag{5.5}$$

$$b_x = 0. \tag{5.6}$$

Then the estimation error is given by

$$\varepsilon_x = k_{f_x f_x} - k_{f_x \bar{g}} K_{\bar{g}\bar{g}}^{-1} k_{\bar{g} f_x} \tag{5.7}$$

Since we are dealing with zero-mean random processes, correlations are the same with covariances. Hence we are justified to investigate the correlation relationships in our model.

The cross correlation between any two random processes $y(x)$ and $z(x)$ is defined by [63]

$$R_{yz}(x_1, x_2) = E[y(x_1) z^*(x_2)]. \tag{5.8}$$

Then the following hold for our model

$$R_{fg}(x_1, x_2) = L_2^* \{R_{ff}(x_1, x_2)\}, \tag{5.9}$$

$$R_{gg}(x_1, x_2) = L_1 \{R_{fg}(x_1, x_2)\} + R_{nn}(x_1, x_2), \tag{5.10}$$

where L_i means the space variable of the operator L is x_i and L^* denotes the linear operator with kernel $h^*(x, x')$.

The estimate and the estimation error can be rewritten in terms of these correlation functions. $k_{f_x\bar{g}}$ can be formed by looking at $R_{fg}(x_1, x_2)$ for a constant $x_1 = x$ and sampling at a finite number of points $x_2 = y_1 \; ... \; y_i \; ... \; y_M$

The total error in estimating f for the duration of f we are interested can be written as

$$\varepsilon = \int_{-\Delta u/2}^{\Delta u/2} \varepsilon_x dx. \tag{5.11}$$

5.4 Numerical Results

This section provides the numerical results showing the relationship between the number of samples, the location of samples and the error. For this purpose, the problem described in section 5.2 is addressed.

In this formulation, one of the principal parameters were assumed to be the number of samples. To choose the other parameter, we focus on the extension of a signal in space and frequency domain. We consider two different scenarios. In the first one, we fix the length of the interval that the samples are taken in. This length is taken to be the extension of the random process defined in section 5.3.1. We assume that samples are taken in the interval $-\Delta u/2 < x < \Delta u/2$ uniformly. Then we observe the error as the number of samples increases. In this case the sampling interval vary as the number of samples vary. In the second scenario, we fix the sampling interval and take it to be the $1/\Delta\mu$. This choice is in accordance with the mean-square sampling theorem for bandlimited WSS random processes. We observe the error as the number of samples is increased. In this case the range the samples are taken vary with the number of samples.

We assume that the input random process $f(x)$ is real. As the linear operator, the real part of fractional Fourier transform has been used. For computation of

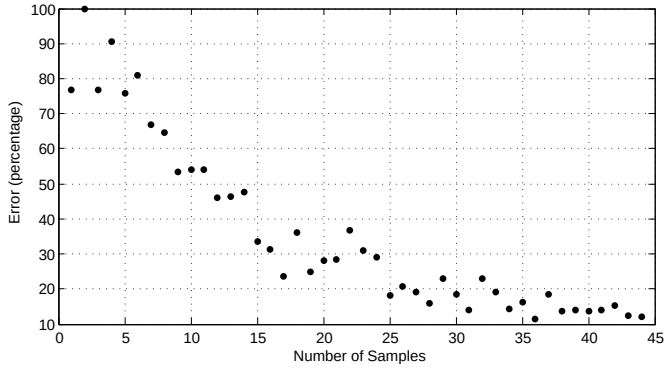

Figure 5.1: Error versus number of samples

the FRT, the algorithm described in [64] and on page 298 of chapter 6 of [32] is used. This implementation is available at [65].

The autocorrelation function of f is taken to be

$$R_f(x_1, x_2) = a(x_1)exp(-0.5(x_1 - x_2')^2)a(x_2'), \tag{5.12}$$

where

$$a(x) = \frac{1}{1 + 0.05x^2}. \tag{5.13}$$

This non-stationary covariance function is formed by the vertical scaling method described in [66] for forming a non-stationary covariance function from a stationary covariance function. The process noise variance is taken to be $0.05 \times \sigma_g^2$. The percent error is calculated according to the following formula

$$100\frac{\varepsilon}{\int_{-\Delta u/2}^{\Delta u/2} E\|f(x)\|^2 dx}. \tag{5.14}$$

The integrals other than the ones containing the FRT are calculated numerically with 2^{11} points with trapezoid rule. Extension in space is determined by finding the points that the signal drops to $\exp(-3)$ of its peak value. Extension in frequency domain is determined with the same criteria by taking into account the rapid fluctuations in this domain.

The results for the first scenario are shown in Figure 5.1. These results show that the general trend of error is to decrease, but the error does not necessarily decrease as the number of samples increases. In producing this graph we increased the number of samples in a uniform manner rather than adding new samples to the previously existing ones. Therefore the sampling points are rearranged for each number of samples. This rearrangement produces effects which cause secondary fluctuations from the general downward trend.

The deviations from the general trend may be explained by the relative importance of the sampling points. When two sampling strategies with different number of sampling points are compared, the position of the samples are also important. If some of the important sampling points are not covered by a sampling strategy, error may be more even if the number of samples increased. An illustrative case can be seen in Figure 5.1 by looking at the first three points corresponding to $1, 2, 3$ samples respectively (For the case with one sampling point, the sample is assumed to be taken at the middle of Δu). The case with one sampling point results in less error than the case with two sampling points. Since the samples of the two samples case correspond to the tails of the correlation function, these samples do not convey much information about the other points. The sample at the middle conveys more information. The information revealed by the three samples case is roughly the same with the one sample case, because contributions of the two samples at the tails to the estimation process is very small.

These results are obtained without searching the optimal locations. For some cases, it may be possible to fine-tune the sampling locations so that the error strictly decreases as the number of samples increases.

Numerical results for the second scenario are shown in Figure 5.2. Comparison of the first and the second scenario for low number samples show that the error for the second scenario is lower. This may be due to the fact that in this

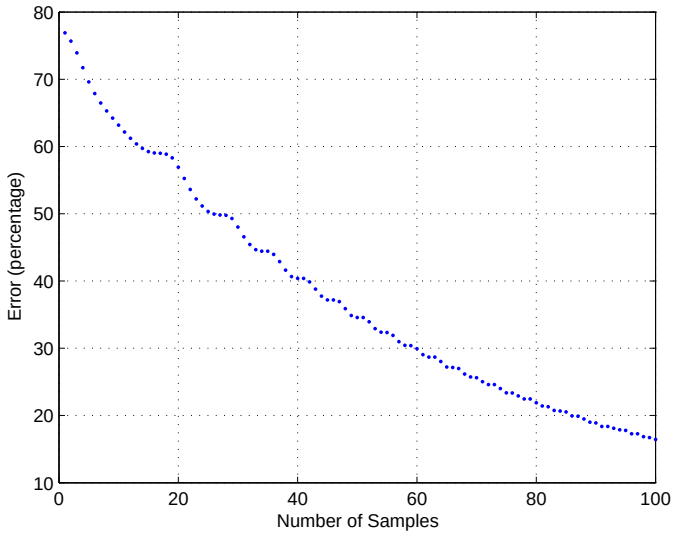

Figure 5.2: Error versus number of samples

case the low number of samples are concentrated around the origin where the correlations are high, where in the first scenario these samples are distributed around a large range, on whose tails the correlations are low. Although the error is less for low number of samples, for high number of samples this approach cannot reach the error level of the previous method. This may be due to the fact that in the current scenario the range covered by the samples is not large enough. This result may be indicating that when a sufficient number of samples are taken from the middle of the distribution, it is better to take samples from the tails instead of taking more samples around the middle.

Chapter 6

Structural and Metrical Information

In this chapter we present a unified framework within which the relationship between both the structural and metrical information of input and output signals can be exploited. Therefore this chapter combines and generalizes to some degree the developments of chapters 4 and 5. The framework of this chapter is the same with chapter 5 unless otherwise stated.

We consider a set of measurements as an experiment strategy which is characterized by its features related to both metrical information and structural information. To model the nature of doing an experiment in an abstract manner, without losing its roots in the underlying physical phenomena, we associate a cost with each measurement in a measurement scheme.

Metrical information were investigated in chapter 3.1. For convenience, we repeat some of the crucial ideas here. In our framework a measurement device is characterized by the number of distinguishable levels associated with it. Each measurement has a cost. Cost of a measurement is defined as cost of using the particular measurement device in this measurement. Cost of using measurement

device solely depends on the device's dynamic range. A measurement device can be adjusted to different ranges and be used with the same cost regardless of the range it is adjusted.

Structural information is related to resolution in space, i.e. the scheme of grasping the details of a signal in space.

6.1 A Simplified Problem Formulation

This section presents a simplified formulation which still captures some of the basic properties of doing measurements in a continuous space.

We assume that measurements have metrical inaccuracy, but they are perfect with respect to structural aspect. That is, every measurement is done at the intended point of measurement without any averaging effect or jitter, but there is uncertainty on the measured value. Although it is possible to allow different number of distinguishable levels for each measurement, we assume every measurement is done with the same number of distinguishable levels.

We assume a uniform sampling strategy. We assume that the cost of a measurement strategy depends on the number of samples taken and the cost of each measurement associated. We want to sample the output signal with a minimum cost as much as possible and obtain the best performance in terms of recovered continuous signal.

6.2 Stochastic Framework

We assume the random process $f(x)$ is input to the linear system. The output is denoted by $L\{f(x)\}$. $n(x)$ and $m(x)$ are also the random processes denoting

the process and measurement noise respectively.

$$s(x) = L\{f(x)\} + n(x) + m(x) \tag{6.1}$$

$$= g(x) + m(x). \tag{6.2}$$

We assume f,n, and m are zero-mean with known correlation functions. Since the processes are zero-mean, correlation functions are the same with the covariance functions.

We observe M samples of $s(x)$. Given $\bar{s} = [s(y_1)\ ...\ s(y_i)\ ...\ s(y_M)]^T = [s_1\ ...\ s_i\ ...\ s_M]^T$, we want to recover $f(x)$ for a particular x. We denote the value to be estimated as f_x and the estimate $\hat{f}(x)$ or as $\hat{f}_x$. We restrict ourselves to affine estimators.

With the analysis performed in section 5.3.2, the estimation error is found to be

$$\varepsilon_x = k_{f_x f_x} - k_{f_x \bar{s}} K_{\bar{s}\bar{s}}^{-1} k_{\bar{s} f_x}. \tag{6.3}$$

The following hold for our model

$$R_{fs}(x_1, x_2) = L_2^*\{R_{ff}(x_1, x_2)\}, \tag{6.4}$$

$$R_{ss}(x_1, x_2) = L_1\{R_{fs}(x_1, x_2)\} + R_{nn}(x_1, x_2) + R_{mm}(x_1, x_2). \tag{6.5}$$

The estimate and the estimation error can be rewritten in terms of these correlation functions.$k_{f_x \bar{s}}$ can be formed by looking at $R_{fs}(x_1, x_2)$ for a constant $x_1 = x$ and sampling at a finite number of points $x_2 = y_1\ ...\ y_i\ ...\ y_M$ The total error in estimating f for the duration of f we are interested can be written as

$$\varepsilon_T = \int_{-\Delta u/2}^{\Delta u/2} \varepsilon_x dx. \tag{6.6}$$

6.3 Numerical Results

This section provides the numerical results illustrating the relationship between the error and measurement strategy. The framework is based on the problem formulation presented in section 6.1. The approach adopted to obtain the numerical results is the same with section 5.4 unless otherwise stated.

In chapter 5, the measurements were assumed to be done perfectly (corresponding to infinite cost). In this chapter we allow uncertainty on the measurements. Although in the most general case, each measurement in a certain measurement strategy would be done with a different number of distinguishable levels, we assume that the number of distinguishable levels is constant for all sampling points.

We consider two different scenarios as in section 5.4, In the first case, it is assumed that the interval samples are taken in is constant. In the second case, it is assumed that the sampling interval is constant. These parameters are chosen as described in section 5.4. For both of the cases, a measurement scenario has two parameters: the number of samples and the number of distinguishable levels. Using the number of samples, we determine the location of samples. Then we calculate the variance of measurement noise at each sample point so that the number of distinguishable levels is constant in each sampling point. Then we calculate the error for this measurement strategy. Then we vary the number of samples and the number of distinguishable levels of measurement devices and observe the change in the error.

The results for the first case are presented in Figure 6.1. For number of distinguishable levels varying from 2 to 64 and number of samples varying from 1 to 64, the error and cost are found and marked on cost-error plane. This plot indicates the way in which the number and location of samples, and the accuracy of detectors should be chosen. It shows similar properties to the trade-off curves

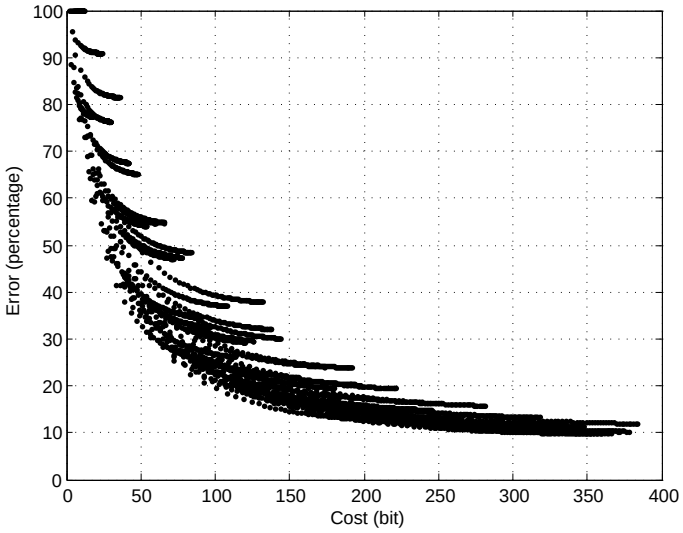

Figure 6.1: Error versus cost

in chapter 4, and similar conclusions can be drawn. For example, the error decrease for low values of cost is very steep. For halving the error from 100% to 50%, the cost needed is less than 50 bits. Then the rate of error decrease becomes smaller. For instance for halving the error from 20% to 10%, more than 150 bits are required. For most of the applications, a budget larger than 250 bits will be unnecessary, since the bits paid do not result in significant error decrease after this cost value.

The results for the second case are presented in Figure 6.2. For number of distinguishable levels ρ from 2 to 80 and number of samples from 1 to 80, error and cost is found and marked on cost-error plane. On the plot different lines showing a downward trend can be distinguished. Each of these lines corresponds to a different constant number of distinguishable levels. We see that for a given number of distinguishable levels, error is non-increasing with increasing the number of samples. The lower left one of these which corresponds to the case of $\rho = 2$ determines the lower-left boundary of the achievable cost-error region for low values of error. Hence to obtain low values of error, it is better to increase the number of samples rather than using more precise devices. This observation may be indicating that when samples are taken at a rate that is sufficient to grasp

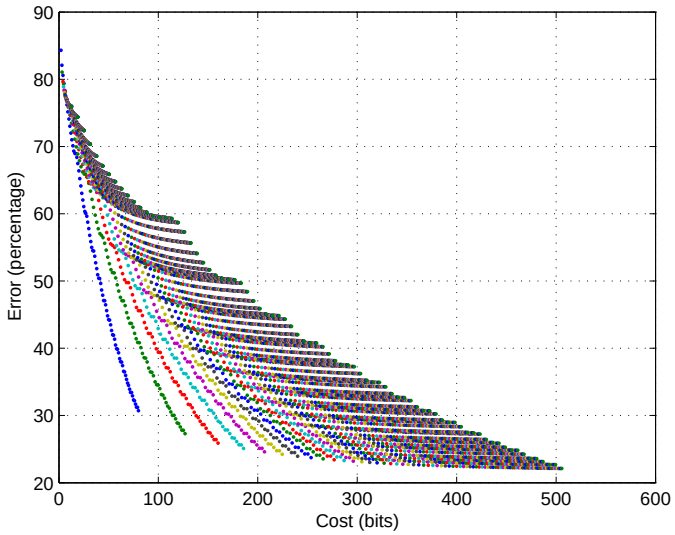

Figure 6.2: Error versus cost

the details of a signal, the burden of more precise devices is too much compared to the error decrease they offer.

The change in the error for high values of error as the number of samples and the number of distinguishable levels vary can be seen more clearly in Figure 6.3. This figure gives the error-cost plot for 2 and 3 samples with ρ varying from 2 to 80. It is seen that for relatively low values of cost (1-8 bits), it is better to measure with 2 samples. For relatively high values of cost, it becomes better to measure with $M = 3$. As the number of distinguishable levels increase, cost increases, but the decrease in the error is not as fast as the $M = 3$ case.

Comparison of Figure 6.1 and Figure 6.2 reveals that the approach based on the extension of the signal in space yields better results. For instance we see that for reducing the error from 100% to 30%, 50 bits are required in the first scenario, whereas 100 bits are required in the second one. In the first scenario, 250 bits are sufficient to reach the error level of 10%, whereas error with even 500 bits is about 20% for the second one.

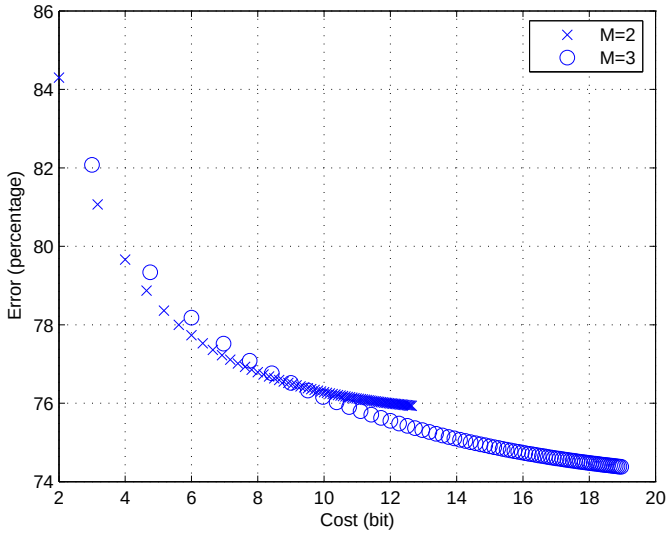

Figure 6.3: Error versus cost for $M = 2$ and $M = 3$

Chapter 7

Conclusions

In this thesis, we have presented a novel framework to understand the information-theoretic relationships in linear systems. Our basic goal was to understand what happens to the information contained in a signal after it is passes through a linear system. For this purpose, we reconsidered an interpretation of information which distinguishes *structural* and *metrical* information as two fundamental concepts.

We have provided an understanding of a measurement and proposed a mathematical model. We have also proposed a measure of cost for doing a measurement, which is a new approach in understanding a measurement. We have exploited this cost function's relationship with the concept of number of distinguishable levels and information theory.

We have investigated the problem separately with special emphasis to metrical information and structural information. For the case of metrical information we have focused on the accuracy of measurements, namely the resolution in amplitude. Two different trade-offs have been considered: the trade-off between the error and the cost, and the trade-off between the mutual information and the cost. These trade-offs illustrate the kinds of problems which can be formulated in

our framework. We have provided an analytic solution for the error-cost problem for the precise measurements case. We have provided numerical results which show how the achievable region forms on the cost-error plane. The boundary of this achievable region indicates the optimal way in which the accuracy of detectors should be chosen. Numerical results showing similar relationships between the cost and the mutual information are also presented.

We have investigated the problem related to structural information, where the emphasis is on the organization of the samples in space. The trade-off between the error and the cost is discussed. To understand the nature of resolving power in space, two different sampling scenarios with uniform sampling strategy are considered. First of these scenarios assumes that the optimal interval that samples are taken in is the extension of the signal in space and takes it as one of the given parameters of the sampling strategy. In the second one, it is assumed that the sampling interval should be one of the principal parameters of a sampling strategy and the optimal sampling interval is derived from the extension of the signal in frequency domain. We have provided numerical results which show how the error changes as number of samples change for these two different sampling scenarios. It is seen that the first strategy performs better for most of the cases.

Finally, we have unified the two frameworks and presented a framework in which the relationship between the error and the cost by means of both structural and metrical information can be investigated. To understand the relationship between the resolution in amplitude and resolution in space, numerical results are presented. These results indicate the optimal way in which the number and location of samples and the accuracy of detectors should be chosen. This final unification, while somewhat simplified, captures the fundamental relationship and trade-off between structural and metrical information and allows us to optimize with respect to the cost of the detection process as well as the costs of the

sampling strategy. For instance it becomes possible to answer questions in the nature of whether it is preferable to use a high number of less accurate or lower number of more accurate detectors.

7.1 Possible Extensions

This section provides some generalizations and extensions.

- The analytic solution of the problems posed in this thesis may be investigated. It may be also beneficial to focus on numerical approaches to obtain the Pareto optimal points.
- Since optical fields, which represent an important application area, have complex values, generalization to complex distributions and complex system kernels is important.
- It may be interesting to solve the problem for distributions other than the Gaussian case. Some cases which may have priority above others can be the cases of correlated noise and uniform noise. The ambiguity in the definition of the cost function for other distributions can be resolved if our definition of cost is interpreted as the capacity of measurement channel. Then the problem can be stated as finding the best measurement strategy when the sum of capacities of measurement channels is limited.
- Since our aim is to understand how much information knowing s gives about the field in general, instead of looking at the error in estimating a particular signal, we can investigate the error in estimation of the set of the signals that are important for us and related to g. This problem can be posed as the problem of estimating Af where the matrix A is a random matrix obeying a probability distribution.

- We may want to model the measurement process such that buying measurement devices have large costs, but when a measurement device is bought it can be used repeatedly with a little cost. This approach focuses on the long term budget considerations from an alternative practical view-point. It will be useful when the main concern is setting up a framework for doing experiments with a limited budget rather than understanding the information theoretic relationships.

- Set of available measurement devices may be characterized by various ways. We may have upper and lower limit constraints on the capacity of some of the measurements in addition to upper limit on total cost. Lower limits on capacity of a particular measurement device may indicate our desire to learn the value of a particular measurement with an accuracy greater than a predefined value. An upper limit may be the result of available measurement devices (i.e., the available set of measurement devices maybe such that it is not possible to do any of the measurements with an accuracy greater than a particular value). In our formulation these limits will appear as limits for noise variances.

- We may investigate best measurement strategy with a linear map before measurement. The best measurement strategy may require us to first map the output of the linear system to a suitable eigenvalue domain before measurement. In this domain, important information will be concentrated in specific locations. So we can distinguish the important information from the rest, and measure it with greater accuracy as required.

- A unified framework which relates measurement device selection problem and the coding problem can be posed. The noise present on the outputs of a sensor essentially maps signal to a finite number of distinguishable levels. The encoding of observations also represents the signals with a finite number of bits. Then there is a connection between the number of

distinguishable levels in a sensor output and the number of bits used to encode a quantity. A quantity that is preferred to be encoded with a greater number of bits will also be preferred to be measured with a device with a greater number of bits. The loss of information as a results of performing a measurement may be investigated in a unified framework with effects of compressing a measurement data for transmission and storage purposes. For instance, in sensor networks the links to the decision center generally have limited capacity. Then it will be meaningless to pay to obtain the values of these parameters with great accuracy. Hence the problem of determining a measurement strategy and the problem of compressing measurement values for transmission should be investigated together.

APPENDIX A

Quantization of a Gaussian Random Variable

This section provides figures related to uniform quantization of a Gaussian random variable. These figures are referred to in section 3.4.

We consider quantization of a random variable and find the best quantization interval Δ for a given number of intervals N_q. The error criteria is the expected square-error between the quantized variable and the original variable. Curves of N_q versus Δ for different σ^2 values are given in Figure A.1. Here N_q is the number of sampling intervals and Δ is the length of best interval. Here $\sigma_1^2 = 2$, $\sigma_2^2 = 3$ and $\sigma_3^2 = 4$.

Figure A.2 shows that for a given number of distinguishable levels N_q the ratio of σ^2 to Δ^2 is roughly constant and increases with increasing N_q. Although the graph is plotted for three different σ^2 values given above, they are indistinguishable in the resolution limits of the graph, since the ratio $\frac{\sigma^2}{N_q}$ is roughly independent of σ^2.

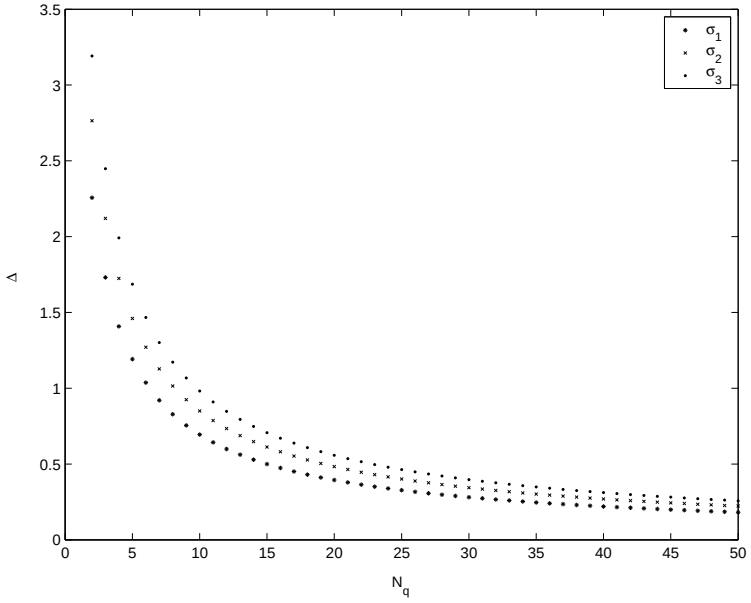

Figure A.1: The optimal value of Δ as a function of N_{q}

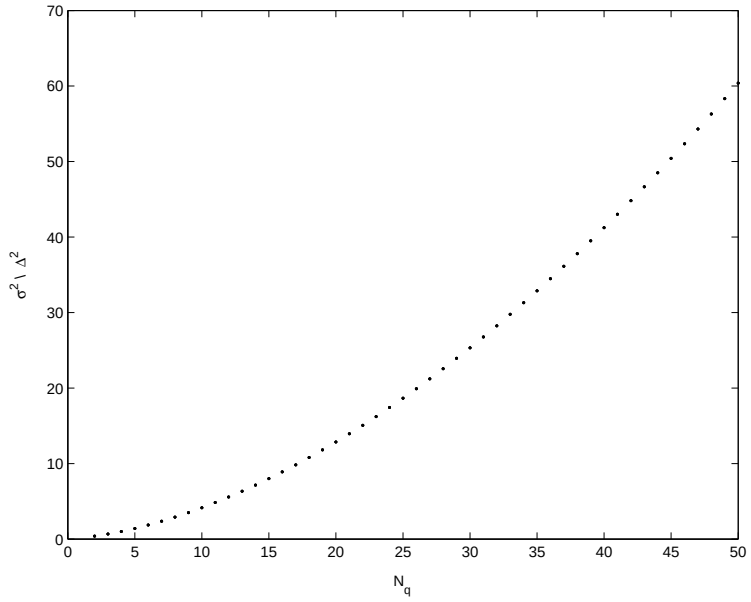

Figure A.2: Number of levels versus ratio of variances

APPENDIX B

Repeated Observations

This section provides the analysis that proves our approach is general enough so that it can handle the case where there are repeated observations. We prove whenever a specific measurement is repeated with different measurement noises yielding a particular MMSE, the equivalent noise that will yield the same MMSE with only one measurement has a lower cost.

B.1 Measurement of a Random Variable with Repeated Observations

Let the observation of a Gaussian random variable only once be modelled as

$$s_i = g_i + m_i. \tag{B.1}$$

Here $g_i \in R$ is a zero-mean random variable with variance $\sigma^2_{g_i}$. m_i is a zero-mean Gaussian random variable with variance $\sigma^2_{m_i}$.

The MSE of MMSE estimate of g_i from s_i is given by

$$\frac{\sigma_{g_i}^2 \sigma_{m_i}^2}{\sigma_{g_i}^2 + \sigma_{m_i}^2}. \tag{B.2}$$

Let the observation of a Gaussian random variable p times be modelled as

$$s = eg_i + m, \tag{B.3}$$

where $m \in \mathbb{R}^p$ is Gaussian with zero-mean and models the measurement noise. e is a vector of ones in $\mathbb{R}^p$. This model describes the case in which each time same random variable g_i is observed with possibly different noise levels. These different noise levels are modelled as noises with different variances.

The MSE of MMSE estimate of g_i from s is given by

$$\frac{\sigma_{g_i}^2 \sigma_{\mathrm{m-eqv}}^2}{\sigma_{g_i}^2 + \sigma_{\mathrm{m-eqv}}^2}, \tag{B.4}$$

where $\sigma_{\mathrm{m-eqv}}^2 = (\sum_{i=1}^{p} \frac{1}{\sigma_{m_i}^2})^{-1}$.

Looking at equations (B.2) and (B.4) the equivalent noise m_{MMSE} that gives the MMSE obtained with repeated measurements with only one measurement is the noise with variance $\sigma_{m_{\mathrm{MMSE}}}^2 = (\sum_{i=1}^{p} \frac{1}{\sigma_{m_i}^2})^{-1}$.

Cost of MMSE equivalent noise for observing a random variable once is smaller than cost of repeated measurements:

$$C(m_{\mathrm{MMSE}}) < C(m). \tag{B.5}$$

B.2 Measurement of a Random Vector with Repeated Observations

We will consider the vector model

$$s = \bar{E}g + m, \tag{B.6}$$

where $\bar{E}$ is a matrix of rank M with M columns. Rows of this matrix is selected from rows of the M by M identity matrix with selection more than once is allowed. This matrix shows which of the observations are repeated.

Let one of the measurements be repeated and let the index of the repeated measurement be 1. Since the indexing of measurements can be freely changed, this assumption is general enough.

$$s_r = \bar{E}_r g + m_r, \tag{B.7}$$

where $\bar{E}_r$ is a $P + M - 1$ by M matrix. $\bar{E}_r$ is given as follows:

$$\bar{E}_r = \begin{pmatrix} \bar{E}_{r1} \\ \bar{E}_{r2} \end{pmatrix}, \tag{B.8}$$

where $\bar{E}_{r1}$ is a matrix with P identical rows e_1^T. e_1^T is the first row of a M by M identity matrix: $e_1^T = \begin{pmatrix} 1 & 0 & ... & 0 \end{pmatrix}^T$. $\bar{E}_{r2}$ is the $M - 1$ by M sub-matrix of the M by M identity matrix obtained by omitting the first row.

m_{r} is a vector of $P + M - 1$ components representing the measurement noise with covariance matrix $K_{m_{\mathrm{r}}} = \mathrm{diag}(\sigma^2_{\mathrm{m_{ri}}})$:

$$m_r = \begin{pmatrix} m_{r1} \\ m_{r2} \end{pmatrix}, \tag{B.9}$$

where m_{r1} is a vector of P components giving the noise added to different instances of the repeated observation and m_{r2} is a vector of $M-1$ components representing the measurement noise added to the other observations. The covariance matrix is given by $K_{m_r} = \mathrm{diag}(\sigma^2_{\mathrm{m_{ri}}}) = \mathrm{diag}(\mathrm{diag}(\sigma^2_{\mathrm{m_{r1i}}}), \mathrm{diag}(\sigma^2_{\mathrm{m_{r2i}}}))$.

We would like to find the equivalent measurement that will substitute the repeated measurements. The model for finding this equivalent noise is the following:

$$s_o = \bar{E}_o g + m_o, \tag{B.10}$$

where $\bar{E}_o$ is the identity matrix. m_o is a vector of M components:

$$m_o = \begin{pmatrix} m_{o1} \\ \\ m_{o2} \end{pmatrix}. \tag{B.11}$$

Here m_{o1} is the equivalent noise with variance $\sigma^2_{m_{o1}}$. m_{o2} and m_{r2} comes from the same distribution. That is $K_{m_{o2}} = K_{m_{r2}}$.

RESULT: Under the models given above, the MMSE error obtained from an experiment with one measurement from each component of the original vector g and the MMSE error obtained from repeated measurements case is equal if and only if $\sigma^2_{m_{o1}} = (\sum_{i=1}^{p} \frac{1}{\sigma^2_{m_{r1i}}})^{-1}$

$$\varepsilon_{s_r} = \varepsilon_{s_o} \Leftrightarrow \sigma^2_{m_{o1}} = (\sum_{i=1}^{p} \frac{1}{\sigma^2_{m_{r1i}}})^{-1}. \tag{B.12}$$

Here the MSE of the MMSE estimate of g obtained from s_r is denoted by ε_{s_r}, and the MSE of the MMSE estimate of g obtained from s_o is denoted by ε_{s_o}.

PROOF: For a system $y = Hx + n$, MMSE is given as

$$\varepsilon = tr(K_x - K_x H^\dagger (H K_x H^\dagger + K_n)^{-1} H K_x^\dagger). \tag{B.13}$$

Instead of MMSE, we look at the matrices

$$H^{\dagger}(HK_xH^{\dagger}+K_n)^{-1}H. \tag{B.14}$$

These are the parts that are different for the two measurement scenarios we consider. Throughout the proof, we make use of the identity

$$\begin{bmatrix} A & B \\ C & D \end{bmatrix}^{-1} = \mathrm{diag}(0,\mathrm{D}^{-1})+\begin{bmatrix} I \\ -D^{-1}C \end{bmatrix}(\mathrm{A}-\mathrm{BD}^{-1}\mathrm{C})^{-1}\begin{bmatrix} I & -BD^{-1} \end{bmatrix}. \tag{B.15}$$

For repeated measurements case, we use the identity (B.15) on $(HK_gH^{\dagger}+K_{m_r})^{-1}$ with the first P by P sub-block as A matrix (all noise covariance are assumed to be positive-definite.) We then use the Woodbury identity on $(A-BD^{-1}C)^{-1}$ part.

For the case where each measurement is done once, we use the identity B.15 on $(K_g+K_{m_o})$. The first 1 by 1 subblock is used as A subblock.

Equating the error generating matrices for repeated case and reduced case reveals that the equivalent noise should be $\sigma^2_{m_{o1}}=(\sum_{i=1}^{p}\frac{1}{\sigma^2_{m_{r1i}}})^{-1}$. Since inverse of power of noise component comes as multiplicative component for a part of the trace that gives the MMSE, solution is unique. This equivalent noise power is same as the one obtained for repeated measurement of a random variable case.

For this value of noise power, MMSE's are equal. Hence converse is also true.

From the inequality B.5, for a given repeated measurement scenario, the equivalent noise that gives the same MMSE has always a less cost. Hence for a given MMSE, it is not possible to obtain a better cost with repeated measurements.

B.3 Finding the MMSE with Cost Equivalent Noise

The noise variance that gives the same cost with the repeated observations is found. The relationship between the cost equivalent noise and MMSE equivalent noise is as follows:

$$\sigma^2_{\text{cost}-\text{eqv}} < \sigma^2_{\text{MMSE}-\text{eqv}}. \tag{B.16}$$

Hence

$$\varepsilon_{s_{m_{\text{cost}-\text{eqv}}}} < \varepsilon_{s_{m_{\text{r}}}}. \tag{B.17}$$

Therefore, for every repeated measurement scenario, there is a measurement scenario where each measurement is done only once and better MMSE is achieved with the same cost.

B.4 Relationship between $\hat{f}$ and $\hat{g}$

LEMMA: If we find the equivalent noise such that g is estimated with the same error as in the repeated measurement case, then we guarantee that f is also estimated with its original error.

PROOF: we know the equivalent noise m_o such that

$$\bar{E}^\dagger(\bar{E}K_g\bar{E}^\dagger + K_{m_r})^{-1})\bar{E} = (K_g + K_{m_o})^{-1} \tag{B.18}$$

holds. For f, the error generating matrix for repeated observations case is

$$K_f H^\dagger \bar{E}^\dagger(\bar{E}K_g\bar{E}^\dagger + K_{m_r})^{-1})\bar{E}HK_f. \tag{B.19}$$

For f, the error generating matrix for reduced observations case is

$$K_f H^\dagger (K_g + K_{m_o})^{-1} H K_f. \tag{B.20}$$

These two are equivalent whenever equation (B.18) holds.

APPENDIX C

Covariance Matrices

This section provides the covariance matrices of the input used for the simulations presented in sections 4.4 and 4.8. Expression C.1 gives the K_f matrix for the 8×8 case. Expressions C.2 and C.3 give the first and second 16×8 sub-blocks of the 16 by 16 K_f matrix for the 16×16 case. The numerical values are shown with 5 digits.

$$\begin{pmatrix} 1.0709 & 0.0404 & -0.1565 & 0.1793 & -0.1602 & 0.1898 & 0.1674 & -0.2036 \\ 0.0404 & 0.6139 & -0.2293 & 0.1629 & 0.1610 & -0.0552 & -0.1143 & -0.0320 \\ -0.1565 & -0.2293 & 0.9382 & -0.2917 & -0.0988 & -0.0232 & 0.0283 & 0.0943 \\ 0.1793 & 0.1629 & -0.2917 & 1.2866 & 0.1175 & -0.0367 & -0.0795 & -0.1363 \\ -0.1602 & 0.1610 & -0.0988 & 0.1175 & 0.8029 & -0.1824 & -0.2060 & 0.0914 \\ 0.1898 & -0.0552 & -0.0232 & -0.0367 & -0.1824 & 1.2477 & 0.2185 & -0.0464 \\ 0.1674 & -0.1143 & 0.0283 & -0.0795 & -0.2060 & 0.2185 & 1.1280 & -0.0785 \\ -0.2036 & -0.0320 & 0.0943 & -0.1363 & 0.0914 & -0.0464 & -0.0785 & 1.2176 \end{pmatrix} \quad \text{(C.1)}$$

$$
\begin{pmatrix}
0.5764 & 0.0463 & 0.0221 & -0.2332 & 0.2549 & 0.0056 & 0.1708 & 0.0077 \\
0.0463 & 0.3272 & 0.1151 & -0.0316 & 0.1163 & -0.0768 & -0.1094 & -0.0233 \\
0.0221 & 0.1151 & 1.0220 & -0.2191 & 0.1860 & -0.2010 & -0.0581 & -0.1033 \\
-0.2332 & -0.0316 & -0.2191 & 0.5522 & -0.2182 & 0.2082 & -0.0703 & 0.0203 \\
0.2549 & 0.1163 & 0.1860 & -0.2182 & 0.7841 & -0.0383 & 0.1899 & 0.0092 \\
0.0056 & -0.0768 & -0.2010 & 0.2082 & -0.0383 & 0.9867 & -0.0541 & -0.0099 \\
0.1708 & -0.1094 & -0.0581 & -0.0703 & 0.1899 & -0.0541 & 0.4659 & -0.1851 \\
0.0077 & -0.0233 & -0.1033 & 0.0203 & 0.0092 & -0.0099 & -0.1851 & 0.9207 \\
0.0885 & -0.1293 & 0.0848 & -0.1329 & 0.1153 & -0.0446 & -0.0446 & -0.0323 \\
-0.0791 & -0.0350 & 0.0391 & 0.0373 & 0.2489 & -0.0569 & -0.0616 & 0.2194 \\
-0.0056 & -0.0570 & -0.0118 & 0.0158 & 0.0169 & -0.2081 & 0.1955 & -0.1817 \\
-0.0290 & 0.0464 & 0.0225 & -0.1849 & 0.0537 & -0.1168 & 0.1514 & 0.1170 \\
-0.1419 & -0.0378 & -0.0765 & 0.1209 & -0.0918 & 0.0351 & -0.0002 & -0.1231 \\
-0.0797 & 0.0535 & -0.0199 & 0.2146 & 0.1093 & -0.2332 & 0.0972 & 0.0643 \\
0.1897 & -0.0861 & -0.0139 & 0.0513 & 0.2265 & 0.1601 & 0.1526 & -0.1905 \\
0.0603 & 0.2859 & 0.1353 & 0.0536 & -0.0794 & -0.0181 & -0.2632 & 0.0495
\end{pmatrix}
\tag{C.2}
$$

$$
\begin{pmatrix}
0.0885 & -0.0791 & -0.0056 & -0.0290 & -0.1419 & -0.0797 & 0.1897 & 0.0603 \\
-0.1293 & -0.0350 & -0.0570 & 0.0464 & -0.0378 & 0.0535 & -0.0861 & 0.2859 \\
0.0848 & 0.0391 & -0.0118 & 0.0225 & -0.0765 & -0.0199 & -0.0139 & 0.1353 \\
-0.1329 & 0.0373 & 0.0158 & -0.1849 & 0.1209 & 0.2146 & 0.0513 & 0.0536 \\
0.1153 & 0.2489 & 0.0169 & 0.0537 & -0.0918 & 0.1093 & 0.2265 & -0.0794 \\
-0.0446 & -0.0569 & -0.2081 & -0.1168 & 0.0351 & -0.2332 & 0.1601 & -0.0181 \\
-0.0323 & -0.0616 & 0.1955 & 0.1514 & -0.0002 & 0.0972 & 0.1526 & -0.2632 \\
0.1354 & 0.2194 & -0.1817 & 0.1170 & -0.1231 & 0.0643 & -0.1905 & 0.0495 \\
0.7926 & 0.2576 & -0.1418 & 0.0429 & 0.1767 & 0.0161 & 0.1011 & -0.0673 \\
0.2576 & 0.8550 & -0.0939 & -0.0090 & 0.0551 & 0.1988 & 0.0776 & -0.1219 \\
-0.1418 & -0.0939 & 0.9649 & -0.0754 & -0.1074 & 0.1559 & 0.1550 & -0.0752 \\
0.0429 & -0.0090 & -0.0754 & 0.7521 & -0.0604 & 0.0051 & -0.2081 & -0.2012 \\
0.1767 & 0.0551 & -0.1074 & -0.0604 & 0.5350 & 0.1177 & 0.0010 & -0.0055 \\
0.0161 & 0.1988 & 0.1559 & 0.0051 & 0.1177 & 1.1707 & -0.0677 & 0.0993 \\
0.1011 & 0.0776 & 0.1550 & -0.2081 & 0.0010 & -0.0677 & 0.6868 & 0.0282 \\
-0.0673 & -0.1219 & -0.0752 & -0.2012 & -0.0055 & 0.0993 & 0.0282 & 0.8340
\end{pmatrix}
\tag{C.3}
$$

Bibliography

[1] D. MacKay, "Quantal aspects of scientific information," *IEEE Transactions on Information Theory*, vol. 1, pp. 60–80, February 1953.

[2] P. Billingsley, *Probability and Measure*. New York: Wiley, 1995.

[3] W. Lukozs, "Optical systems with resolving powers exceeding the classical limit," *Journal of Optical Society of America*, vol. 56, pp. 1463–1472, November 1966.

[4] E. N. Leith, "The evaluation of information optics," *IEEE Journal of Selected Topics in Quantum Electronics*, vol. 6, pp. 1297–1304, November/December 2000.

[5] E. N. Leith, "Some highlights in the history of information optics," *Information Sciences*, vol. 149, pp. 271–275, November/December 2003.

[6] W. Lukozs, "Optical systems with resolving powers exceeding the classical limit II," *Journal of Optical Society of America*, vol. 57, pp. 932–941, July 1967.

[7] G. Toraldo Di Francia, "Resolving power and information," *Journal of Optical Society of America*, vol. 45, pp. 497–501, July 1955.

[8] G. Toraldo Di Francia, "Degrees of freedom of an image," *Journal of Optical Society of America*, vol. 59, pp. 799–804, July 1969.

[9] F. Gori and G. Guattari, "Effects of coherance on the degrees of freedom of an image," *Journal of the Optical Society of America*, vol. 61, pp. 36–39, January 1971.

[10] F. Gori and G. Guattari, "Shannon number and degrees of freedom of an image," *Optics Communications*, vol. 7, pp. 163–165, February 1973.

[11] F. Gori and G. Guattari, "Degrees of freedom of images from point-like-element pupils," *Journal of Optical Society of America*, vol. 64, pp. 453–458, April 1974.

[12] F. Gori and L. Ronchi, "Degrees of freedom for scatterers with circular cross section," *Journal of Optical Society of America*, vol. 71, pp. 250–258, March 1981.

[13] L. Ronchi and F. Gori, "Degrees of freedom for spherical scatterers," *Optics Letters*, vol. 6, pp. 478–480, October 1981.

[14] R. Piestun and D. A. Miller, "Electromagnetic degrees of freedom of an optical system," *Journal of Optical Society of America A*, vol. 17, pp. 892–902, May 2000.

[15] F.Gori, *Advanced Topics in Shannon Sampling and Interpolation Theory*, ch. 2, pp. 37–83. Springer-Verlag.

[16] F. Gori, S.Paolucci, and L. Ronchi, "Degrees of freedom of an optical image in coherent illumination, in the presence of aberrations," *Journal of Optical Society of America*, vol. 65, pp. 495–501, May 1975.

[17] J. T. Winthrop, "Propagation of structural information in optical wave fields," *Journal of Optical Society of America*, vol. 61, pp. 15–30, January 1971.

[18] D. Gabor, "Light and information," in *Progress In Optics* (E. Wolf, ed.), vol. I, ch. 4, pp. 109–153, Elsevier, 1961.

[19] F. Gori and M. Santarsiero, "Coherence and the spatial distribution of intensity," *Journal of Optical Society of America A*, vol. 10, pp. 673–679, April 1993.

[20] R. Barakat, "Degrees of freedom of an image," *Optics Communications*, vol. 156, pp. 235–239, December 1998.

[21] A. Thaning, P. Martinsson, M. Karelin, and A. T. Friverg, "Limits of diffractive optics by communication modes," *Journal of Optics A-Pure and Applied Optics*, vol. 5, pp. 153–158, May 2003.

[22] P. Réfrégier, F. Goudail, P. Chavel, and A. Friberg, "Entropy of partially polarized light and application to statistical processing techniques," *Journal of Optical Society of America A*, vol. 21, pp. 2124–2134, November 2004.

[23] A. Stern and B. Javidi, "Shannon number and information capacity of three-dimensional integral imaging," *Journal of Optical Society of America A*, vol. 21, pp. 1602–1612, September 2004.

[24] M. A. Porras and R. Medina, "Entropy-based definition of laser beam spot size," *Applied Optics*, vol. 34, pp. 8247–8251, December 1995.

[25] J. Rubinstein and G. Wolansky, "A variational principle in optics," *Journal of Optical Society of America A*, vol. 21, pp. 2164–2172, November 2004.

[26] L. Onural, "Sampling of the diffraction field," *Applied Optics*, vol. 39, pp. 5929–5935, November 2000.

[27] A. Stern and B. Javidi, "Analysis of practical sampling and reconstruction from fresnel fields," *Optical Engineering*, vol. 43, pp. 239–250, January 2004.

[28] H. Stark, ed., *Image Recovery —Theory and Application.* Academic Press, 1987.

[29] D. Youla and H. Webb, "Image restoration by the method of convex projections: Part I-theory," *IEEE Transactions on Medical imaging*, vol. MI-1, pp. 81–94, October 1982.

[30] A. Çetin, H. Özaktaş, and H. M. Ozaktas, "Resolution enhancement of low resolution wavefields with POCS algorithm," *Electronics Letters*, vol. 39, pp. 1808–1810, December 2003.

[31] G. B. Esmer, V. Uzunov, L. Onural, H. M. Ozaktas, and A. Gotchev, "Diffraction field computation from arbitrarily distributed data points in space." to appear in *Signal Processing: Image Communication*.

[32] H. M. Ozaktas, Z. Zalevsky, and M. A. Kutay, *The Fractional Fourier Transform with Applications in Optics and Signal Processing*. New York: Wiley, 2001.

[33] F. T. Yu, *Optics and Information Theory*. New York: Wiley , 1976.

[34] F. T. Yu, *Entopy and Information Optics*. New York: Marcel Dekker, 2000.

[35] A. Papoulis, "Generalized sampling expansion," *IEEE Transactions on Circuits and Systems*, vol. CAS-24, pp. 652–654, November 1977.

[36] A. K. Dey, C. F. Martin, and F. H. Ruymgaart, "Input recovery from noisy output data, using regularized inversion of the Laplace transform," *IEEE Transactions on Information Theory*, vol. 44, pp. 1125–1130, May 1998.

[37] R. Venkataramani and Y. Bresler, "Multiple-input multiple-output sampling: Necessary density conditions," *IEEE Transactions on Information Theory*, vol. 50, pp. 1754–1768, August 2004.

[38] S. Farkash and S. Raz, "Linear systems in Gabor time-frequency space," *IEEE Transactions on Signal Processing*, vol. 42, pp. 611–617, March 1994.

[39] E. Palmer and G.Ledwich, "Optimal placement of angle transducers in power systems," *IEEE Transactions on Power Systems*, vol. 11, pp. 7–12, May 1996.

[40] C. Madtharad, S. Premrudeepreechacharn, N. R.Watson, and R. Saeng-Udom, "An optimal measurement placement method for power system harmonic state estimation," *IEEE Transactions on Power Delivery*, vol. 20, pp. 1514–1521, April 2005.

[41] E. Trucco, M. Umasuthan, A. M. Wallace, and V. Roberto, "Model-based planning of optimal sensor placements for inspection," *IEEE Transactions on Robotics and Automation*, vol. 13, pp. 182–194, April 1997.

[42] P. Bruzzone, A. Formisano, and R. Martone, "Optimal magnetic probes location for current reconstruction in multistrands superconducting cables," *IEEE Transactions on Magnetics*, vol. 38, pp. 1057–1060, May 2002.

[43] S. J. Reeves and L. P. Heck, "Selection of observations in signal reconstruction," *IEEE Transactions on Signal Processing*, vol. 43, pp. 788–791, March 1995.

[44] S. Kumar and J. H. Seinfeld, "Optimal location of measurements for distributed parameter estimation," *IEEE Transactions on Automatic Control*, vol. AC-23, pp. 690–698, August 1978.

[45] R. K. Mehra, "Optimization of measurement schedules and sensor designs for linear dynamic systems," *IEEE Transactions on Automatic Control*, vol. AC-21, pp. 55–64, April 1976.

[46] D. Slepian and J. K. Wolf, "Noiseless coding of correlated information sources," *IEEE Transactions on Information Theory*, vol. IT-19, pp. 471–480, July 1973.

[47] A. H. Kaspi and T. Berger, "Rate-distortion for correlated sources with partially separated encoders," *IEEE Transactions on Information Theory*, vol. IT-28, p. 828, November 1982.

[48] A. Gamal and T. M. Cover, "Achievable rates for multiple descriptions," *IEEE Transactions on Information Theory*, vol. IT-28, pp. 851–857, November 1982.

[49] T. J. Flynn and R. M. Gray, "Encoding of correlated observations," *IEEE Transactions on Information Theory*, vol. IT-33, pp. 773–787, November 1987.

[50] T. B. Jun Chen, Xin Zhang and S. B. Wicker, "An upper bound on the sum-rate distortion function and its corresponding rate allocation schemes for the CEO problem," *IEEE Journal on Selected Areas In Communications*, vol. 22, pp. 977–987, August 2004.

[51] R. Ahlswede and I. Csiszar, "Hypothesis testing with communication constraints," *IEEE Transactions on Information Theory*, vol. 32, pp. 533–542, July 1986.

[52] J. Max, "Quantizing for minimum distortion," *IRE Transaction on Information Theory*, vol. 6, pp. 7–12, March 1960.

[53] E. Arıkan, 2005. Private Communication.

[54] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. Wiley.

[55] I. E. Telatar, "Capacity of multi-antenna Gaussian channels," *European Transactions on Telecommunications*, vol. 10, pp. 585–595, November 1999. invited paper.

[56] Çağatay Candan, M. A. Kutay, and H. M. Ozaktas, "The discrete fractional Fourier transform," *IEEE Transactions on Signal Processing*, vol. 48, pp. 1329–1337, May 2000.

[57] Çağatay Candan, M. A. Kutay, and H. M. Ozaktas, "The discrete fractional Fourier transform," in *Proceedings of IEEE International Conference on Acoustics,Speech and Signal Processing*, vol. 3, pp. 1713–1716, March 1999.

[58] Çağatay Candan, "Discrete fractional Fourier transform matrix generator." http://www.ee.bilkent.edu.tr/∼haldun/dFRT.m, July 1998.

[59] S. Boyd and L. Vandenberghe, *Convex Optimization*. New York: Cambridge University Press, 2004.

[60] A. J. Jerri, "The Shannon sampling theorem —its various extensions and applications: A tutorial review," *Proceedings of the IEEE*, vol. 65, pp. 1565–1596, November 1977.

[61] W. Sun and X. Zhou, "Reconstruction of band-limited signals from local averages," *IEEE Transactions on Information Theory*, vol. 48, pp. 2955–2963, November 2002.

[62] I. B. Rhodes, "A tutorial introduction to estimation and filtering," *IEEE Transactions on Automatic Control*, vol. AC-16, pp. 688–706, December 1971.

[63] H. Stark and J. W. Woods, *Probability and Random Processes with Applications to Signal Processing*. Upper Saddle River, New Jersey: Prentice Hall, 2002.

[64] H. M. Ozaktas, O. Arıkan, M. A. Kutay, and G. Bozdağı, "Digital computation of the fractional Fourier transform," *IEEE Transactions on Signal Processing*, vol. 44, pp. 2141–2150, September 1996.

[65] M. A. Kutay, "Fast computation of the fractional Fourier transform." http://www.ee.bilkent.edu.tr/∼haldun/fracF.m, September 1996.

[66] D. MacKay, "Introduction to Gaussian proceses," *NATO ASI Series F Computer and Systems Sciences*, vol. 168, pp. 133–166, 1998.