Set Theoretic Estimation in Digital Signal Processing

Patrick L. Combettes

Center for Communications and Signal Processing
Department of Electrical and Computer Engineering
North Carolina State University

TR-89/12
September 1989
ABSTRACT

COMBETTES, PATRICK L. Set Theoretic Estimation in Digital Signal Processing.
(Under the direction of Professor H. Joel Trussell).

An abstract treatment of set theoretic estimation, along with new theoretical developments, are presented and new perspectives for applications in digital signal processing are opened. Set theoretic estimation is formally defined as a technique in which consistency with all the a priori knowledge pertaining to the true object serves as an estimation criterion. All the pieces of a priori knowledge are represented by sets in some abstract solution space and, hence, producing a set theoretic estimate amounts to finding a common point of sets. Two new methods for the synthesis of set theoretic estimates are developed.

In the first one, the method of successive projections for closed and convex subsets of Hilbert spaces is generalized to approximately compact subsets of metric spaces. The second strategy seeks a solution by random search in a restricted region of the solution space. With the introduction of these two methods, set theoretic estimation no longer is restricted to problems where all the sets are closed and convex and the solution space hilbertian. Thereby, greater flexibility with regard to the incorporation of the a priori knowledge is achieved. The use of various probabilistic noise properties is investigated in a general set theoretic framework. Sets based on pieces of a priori knowledge relative to the range, moments, absolute moments, second and higher order properties of the noise process are constructed and analyzed. Specific digital signal processing applications are considered and simulations are provided to illustrate the theoretical developments.
Set Theoretic Estimation in Digital Signal Processing

by

Patrick L. Combettes

A thesis submitted to the Graduate Faculty of
North Carolina State University
in partial fulfillment of the
requirements for the Degree of
Doctor of Philosophy

Department of Electrical and Computer Engineering

Raleigh

June 1989

Approved By:

S. Thomas Alexander

Arne A. Nilsson

Jack W. Silverstein

Oscar Wesler

H. Joel Trussell
Chairman of Advisory Committee
Patrick L. Combettes was born in Salon de Provence, France, on May 24, 1962. He received the "Diplôme d'Ingénieur" from l'Institut National des Sciences Appliquées de Lyon, France, in June 1985 and the Master of Science from North Carolina State University in May 1987, both in Electrical Engineering.

He has been a teaching and a research assistant in the Department of Electrical and Computer Engineering at North Carolina State University since August 1985.
ACKNOWLEDGEMENTS

It has been a privilege to have been Professor H. Joel Trussell's student during the past few years and I wish to thank him for his significant contributions to my education as a researcher. I also want to express my gratitude to Dr. Jack W. Silverstein for his valuable comments with regard to several aspects of this work and to Professor Oscar Wesler for his remarkable lectures in probability theory.

Finally, I am grateful to Thierry Chaussalet for bringing to my attention this paper by Paul Erdős (1945) which contained the result on the almost everywhere uniqueness of projections I was trying to prove, and to Victoria Chi for the figures in Chapter II.
# TABLE OF CONTENTS

I. INTRODUCTION ........................................................................................................... 1  
1.1. Opening Remarks ................................................................................................. 1  
1.2. A Critique of Conventional Estimation Techniques .......................................... 2  
1.3. Set Theoretic Estimation ...................................................................................... 4  
  1.3.1. Consistency with the a Priori Knowledge as an Estimation Criterion .......... 4  
  1.3.2. The Set Theoretical Formalization ................................................................. 5  
  1.3.3. Basic Issues .................................................................................................... 6  
  1.3.4. The State of the Art ....................................................................................... 7  
1.4. Research Methodology ......................................................................................... 8  

II. SET THEORETIC ESTIMATION BY SUCCESSIVE PROJECTIONS METHODS .......... 10  
2.1. Introduction ......................................................................................................... 10  
2.2. Preliminary Results .............................................................................................. 12  
  2.2.1. Remarks on Closedness, Compactness, and Convexity ................................. 12  
  2.2.2. Complements on Banach Spaces ................................................................... 13  
  2.2.3. Topological Properties of Set-Valued Maps ................................................... 14  
2.3. Projections in Metric Spaces .............................................................................. 16  
  2.3.1. Distance to a Set ............................................................................................ 16  
  2.3.2. Bounded and Approximative Compactness .................................................... 17  
  2.3.3. Projection Maps ............................................................................................ 18  
  2.3.4. Existence of a Projection .............................................................................. 20  
  2.3.5. Uniqueness of a Projection .......................................................................... 22  
  2.3.6. Some Remarks on Chebyshev Sets ................................................................. 23  
  2.3.7. Continuity of Projections ............................................................................. 23  
  2.3.8. Characterization of a Projection ................................................................... 24  
  2.3.9. Projections onto Monotonic Sequences of Sets ............................................ 25  
  2.3.10. A Property of the Points with Multiple Projections ................................. 25  
  2.3.11. A Few Counterexamples ............................................................................ 25  
2.4. Projections in the Cartesian Space ..................................................................... 26  
  2.4.1. Generalities .................................................................................................... 26  
  2.4.2. Ordinary and Multifurcation Points ............................................................... 27  
  2.4.3. Existence, Uniqueness, and Continuity of Projections ................................. 28  
  2.4.4. The Reach, the Convex Deficiency, and the Skeleton of a Set .................... 29  
  2.4.5. Properties of Ordinary and Multifurcation Points ....................................... 30  
2.5. The Method of Successive Projections in Hilbert Spaces .................................. 31  
2.6. The Method of Successive Projections in Metric Spaces .................................. 33  
  2.6.1. The Cyclic Projection Map ........................................................................... 34
CHAPTER I

INTRODUCTION

1.1. Opening Remarks

At the very core of most digital signal processing concepts lies an estimation problem. Indeed, a typical digital signal processing problem can be abstracted into estimating an object (e.g. a single parameter, a collection of parameters, a function) from the data provided by observing some stochastic process. For instance, in a typical signal deconvolution problem such as signal restoration, one seeks to obtain an estimate of a signal from data which have been degraded by some operator and contaminated by noise; in computerized tomography the goal is to reconstruct a two-dimensional image from the limited data provided by its line integrals; the problem of spectral estimation is to determine the spectral distribution of a random process, i.e. the distribution in frequency of its power; in speech prediction the goal is to form an estimate of a given sample as a function of past data samples; in system identification, the problem is to determine the parameters of a model given observations of the physical system being modeled.

These problems are usually approached via conventional estimation techniques, i.e. techniques that generate a solution which is optimal with respect to some predefined criterion. The most widespread optimality criteria encountered in the signal processing literature, include maximum likelihood, maximum a posteriori, minimum mean square error, and maximum entropy. Despite the fact that they have yielded results which have been judged
satisfactory in a broad spectrum of applications, conventional estimation techniques are
however open to criticism in several respects.

1.2. A Critique of Conventional Estimation Techniques

We have defined conventional estimation techniques as those which produce an estimate
which is optimal in some sense. Hence, this critique will mainly be a critique of the
rationality of the notion of optimality. We shall center our discussion around Bayesian
techniques, for they represent the vast majority of the estimation procedures employed in
signal processing.

In the Bayesian approach [11], [63], the object \( h \) to be estimated is treated as the realiza­
tion of a random variable \( H \) to which a distribution function is assigned (in most situa­
tions, there is however nothing random about \( h \), which may for instance be a fixed but
unknown physical constant). Then a loss function \( L(H, a) \) is adopted, where \( a \) is the esti­
mate of \( h \). A solution to the problem is any \( a \) which minimizes the posterior expected
loss, i.e. the expected loss conditioned on the data. At all events, a Bayesian technique is
fundamentally subjective since the specification of the prior distribution of \( H \) and of the
loss function is left to the user. In the extreme, the user could reach any conclusion he
desires through biased choices of these specifications. Moreover, as discussed and clearly
illustrated in [11], Bayesian procedures are not robust in the sense that small changes in
the specifications, i.e. the prior distribution and the loss function, may cause significant
changes in the estimate.

Common sense would argue that some degree of objectivity could be attained by choosing
these specifications in the way which seems the most appropriate, on the basis of the
available knowledge of the situation and of experience. In practice, however, little or no
regard is paid to such an analysis and the choice of the prior distribution and of the loss function is generally motivated by the desire to simplify the computations and to obtain closed form solutions. To illustrate this point, let us consider a few examples. In Bayesian digital image restoration [3], a common assumption is to regard the original image as the realization of a random vector with a multivariate Gaussian density. This model raises concern in that it allows solutions with negative intensity values, which is a physical impossibility. This example exhibits an undesirable feature of conventional estimation techniques which is to give rise to solutions which may violate some of the basic known properties of the true object. Another source of concern resides in the choice of the loss function. For instance, the very popular Wiener filter is based on the squared estimation error. This loss function is used only on account of its mathematical simplicity and the properties of the human eye, which is known not to be an optimal least-squares detector, are ignored. Because it makes the optimization procedure easy, the squared estimation error loss function is employed in all branches of signal processing without further formal justification. It is in particular central in adaptive signal processing [2].

As a conclusion, conventional estimation techniques are attractive because they provide the user with the comforting belief that he has solved the problem in an optimal way. However, because of the arbitrariness of the criterion of optimality and the inherent bias associated with the working hypotheses, there is little rationality in the quest for a "best" solution. In the following, we shall introduce a general estimation method which does not provide an optimal solution but, rather, a class of solutions which we shall define as a collection of objects which are consistent with all the a priori knowledge.
1.3. Set Theoretic Estimation

1.3.1. Consistency with the a Priori Knowledge as an Estimation Criterion

Following the definition Kant gave in 1781 in his *Kritik der reinen Vernunft*, we may take *a priori* knowledge to be knowledge which is independent of experience. Thus, in epistemology, an instance of knowledge is *a priori* if its justification condition does not depend on evidence from sensory experience; in contrast, an instance of knowledge is *a posteriori* if its justification condition depends on evidence from sensory experience. In our estimation framework, *a priori* information is information about the true object that arises from sources other than the statistical investigation. A typical piece of *a priori* knowledge may consist of a mathematical property of the true object (e.g. the intensity values in a digital image are nonnegative quantities), or a proposition pertaining to the system which generated the observed data, such as a constraint imposed by physical laws (e.g. the system that generated speech samples is necessarily stable).

An estimation problem is always accompanied with some *a priori* knowledge, which we shall represent by a collection of propositions \( \{K_\alpha\}_{\alpha \in A} \), where \( A \) is a nonempty index set. The amount of *a priori* knowledge depends on our ingenuity and the extent of our theoretical and practical understanding of the physical system under study. Each piece of *a priori* information \( K_\alpha \) narrows our ignorance of the true object and is therefore valuable in increasing objectively the precision of the estimate. Hence, it is sensible to adopt consistency with all the pieces of *a priori* knowledge, rather than optimality with respect to some arbitrary standard, as an estimation criterion. The estimates produced according to this principle are constrained to satisfy only those properties which the true object is known to possess. This central concept can be neatly formalized via set theory.
1.3.2. The Set Theoretical Formalization

In the past, it was thought that mathematics was a motley of autonomous disciplines which were isolated in their goals as well as in their methods and their language. In the late 1930's, the most influential Bourbaki group proposed to unify mathematics by means of abstract structures and established that it was possible, logically speaking, to derive practically the whole of known mathematics from a single source, namely set theory [14]. Nowadays it is commonly agreed that, in pure and applied mathematics, ideas can often most simply and concisely be expressed in terms of set concepts and set notations. Set theory also provides a natural and convenient setting for describing the problem of constructing estimates which agree with all the a priori knowledge.

Let \( \{K_a\}_{a \in A} \) be the collection of propositions representing the a priori knowledge for an estimation problem and let \( \Xi \) be the solution space, i.e., the space where the true solution lies. A collection of sets \( \{S_a\}_{a \in A} \) can be constructed in a propositional manner as follows

\[
S_a = \left\{ x \in \Xi \mid K_a \text{ holds for } x \right\} \quad (1.1)
\]

Such sets will be called the property sets. A solution to the problem is one which is consistent with all the a priori knowledge and must therefore belong to the intersection set

\[
S = \bigcap_{a \in A} S_a = \left\{ x \in \Xi \mid (\forall a \in A) \ K_a \text{ holds for } x \right\} \quad (1.2)
\]

The set \( S \) will be interchangeably called the solution set or the feasible set. A point in \( S \) will be called a set theoretic estimate; it is an object which lies in all the sets modeling the a priori constraints in \( \Xi \). Clearly, the true object lies in \( S \). It will always be assumed that the a priori propositions have been formulated in a coherent manner, so that the feasible set is not empty.
1.3.3. Basic Issues

We shall discuss here some fundamental questions in connection with set theoretic estimation and see how it profoundly differs, in philosophy and methodology, from conventional estimation techniques.

Conventional estimation techniques provide an optimal, usually unique, answer which is the result of some optimization process. Set theoretic estimation techniques rule out single answers and, instead, provide a set of answers. From our standpoint, any point in the solution set $S$ is as valid an estimate as any other, as it is consistent with every piece of a priori information. Because the set theoretic estimates are not derived in terms of optimality, it is legitimate to ask how their quality can be assessed. The quality of a set theoretic estimation procedure could be quantified by some measure of dispersion of the elements in $S$ (e.g. the diameter of $S$ if $\Xi$ is a metric space, or the ratio of the Lebesgue measure of $S$ to that of the union of all the property sets if $\Xi$ is the Cartesian space $\mathbb{R}^k$). Any measure of dispersion is however subjective and, hence, it may give rise to debatable interpretation. We shall therefore take care not to adopt any such measure. Qualitatively speaking, increasing the number of sets involved in the description of the solution will be beneficial. Indeed, suppose that $S'$ is the feasible set based on $m-1$ property sets $\{S_1, ..., S_{m-1}\}$. Then, upon the incorporation of an additional property set $S_m$, the feasible set becomes

$$ S = \bigcap_{i=1}^{m} S_i \subseteq \bigcap_{i=1}^{m-1} S_i = S' $$

(1.3)

which is clearly smaller than $S'$. As a result, the dispersion of the solution points about the true object has decreased.
1.3.4. The State of the Art

In applied mathematics, there are numerous problems which have been formulated within the general format of set theoretic estimation. We can specifically mention the resolution of systems of linear equations [55], multi-constrained optimization [66], and band-limited extrapolation [64]. Additional references can be found in [46] and [74]. In recent years, set theoretic estimation has also become very popular in digital signal processing. The problems which have been thus approached include signal restoration [32], [90], [104], [113], signal reconstruction [30], [65], image coding [87], tomographic image reconstruction [91], filter design [1], [29], and processing of electron microscopy data [28].

A fundamental problem in set theoretic estimation is the synthesis of a solution, i.e. the problem of finding a point in the intersection of a given collection of sets. It is our opinion that the lack of mathematical methods for performing this task constitutes the major limitation of set theoretic estimation. In fact, at present, the only theoretically sound method which is available is the method of successive projections (MOSP). Unfortunately, the MOSP applies only to collections of closed and convex property sets in a Hilbert solution space, which is a severe restriction since, for instance, some important pieces of a priori information may not be representable by convex sets.

Another central issue in set theoretic estimation is the selection and the construction of property sets. In this regard, a significant contribution was made in [104], where it was demonstrated that, in a signal restoration environment, incorporating sets based on some properties of the noise could greatly improve the quality of the estimate. In [31], the problem of constructing sets in an environment where the a priori information is inaccurate or partially defined was addressed through fuzzy set theory. In that approach, pieces of
imprecise information are modeled with fuzzy sets, i.e. classes of objects with a continuum of grades of membership [114].

1.4. Research Methodology

The purpose of this dissertation is to deepen the theoretical foundations of set theoretic estimation and to enlarge the field of its applications, with special emphasis on digital signal processing.

Chapter II will address the problem of obtaining set theoretic estimates by successive projections methods. The main contribution of this chapter will be the generalization of the MOSP to collections of approximately compact sets in metric spaces and the subsequent derivation of results for the convergence of the method to a solution. In Chapter III, a new method for generating set theoretic estimates, called method of random search (MORS), will be presented. In that method, a solution is sought by random search in a restricted region of the solution space. The advantages and limitations of the MORS will be compared to that of the MOSP. Chapter IV will be devoted to a generalization of the use of noise properties in a very broad class of set theoretic estimation problems. Sets will be constructed by constraining the sample statistics of the estimation residual to agree with various probabilistic pieces of knowledge relative to the noise process (range, moments, absolute moments, second order properties, higher order properties). The theoretical results of Chapters II, III, and IV will be applied to digital signal restoration in Chapter V and harmonic retrieval in Chapter VI. Simulations will be provided to illustrate the abstract developments and allow the comparison between set theoretic and conventional estimation methods. Conclusions and suggestions for future research will appear in Chapter VII.
Some parts of this dissertation involve a mathematical machinery which may be unfamiliar to some of the readers. To make our treatment complete in that respect, we have added two appendices, one on set theory, topology, topological vector spaces, and functional analysis, the other on measure theory, integration, probability theory and stochastic processes. Most of the definitions, notations, and basic results used throughout this dissertation can be found there. The symbol $\Box$ will be used to signal the end of a proof.
CHAPTER II

SET THEORETIC ESTIMATION BY SUCCESSIVE PROJECTIONS METHODS

2.1. Introduction

As was noted in Chapter I, a central problem in set theoretic estimation is the synthesis of a solution. This problem can be simply stated as follows: given a collection of property sets \( \{S_1, \ldots, S_m\} \) in some abstract solution space \( \Xi \), find an object \( z \) in the intersection \( S \) of these sets, which is assumed nonempty. In this chapter, the general theory of projections in metric spaces will be employed to solve this fundamental problem. More specifically, the mathematical procedure under focus will be the method of successive projections (MOSP). This iterative method has been shown to yield a solution in the case where all the property sets are closed and convex in a Hilbert space. All the problems mentioned in Section 1.3.4 were solved in that framework.

There is a vast body of problems whose set theoretic formulation does not comply with the requirements stated above, which narrows considerably the scope of the MOSP. For instance, one of the sets may not be convex or the underlying space may not be hilbertian. In signal processing, nonconvex property sets are common: in linear prediction, the set of stable prediction-error filters of order greater than two is not convex; in Section 5.4 several nonconvex sets will be mentioned in the context in digital signal restoration; in Chapter VI nonconvex sets will be used for the harmonic retrieval problem; finally, many
matrix property sets encountered in signal processing are nonconvex, as noted in [26]. In that same paper, an attempt was made to extend projection methods to more general types of property sets, in a signal enhancement environment.

The limitations of the MOSP in its present form have motivated the main objective of this chapter, namely to establish a rigorous method, based on an extension of the MOSP to arbitrary metric spaces, for obtaining a point in the intersection of a finite collection of sets. To make the exposition complete, necessary mathematical concepts are introduced in Section 2.2. In Section 2.3, an extensive account of the theory of projections in metric spaces is provided. The elements introduced there will be indispensable to subsequent developments. Section 2.4 is devoted to the study of projections in the Cartesian space, which is the underlying solution space in most practical applications. The MOSP in Hilbert spaces is thoroughly reviewed in Section 2.5. Recent developments of the theory are included. The highlight of this chapter is Section 2.6, where the MOSP is generalized to collections of approximately compact sets in metric spaces and where conditions for the convergence of the method to a solution are set forth.

Most of the material presented in this chapter will be purely mathematical. Although no mention will be made of any particular signal processing application at this point, the reader should constantly bear in mind the geometrical interpretation of the results and should make the natural connection with potential applications. In Chapter V, we shall demonstrate a practical application of the MOSP to the problem of digital signal restoration. In the simulations presented there, both convex and nonconvex property sets will be employed to produce a solution.
2.2. Preliminary Results

2.2.1. Remarks on Closedness, Compactness, and Convexity

The notions of closedness, compactness, and convexity are central to projection methods. In this section, we shall therefore provide some useful criteria under which these fundamental properties can easily be identified. As to defining properties, the reader is referred to Appendix A.

Proposition 2.1.

A subset $S$ of a metric space $\Xi$ is closed if any of the following conditions holds

(i) $S$ is a finite union or an arbitrary intersection of closed sets.

(ii) There exists a real-valued continuous function $f$ defined on $\Xi$ and a closed subset $C$ of $\mathbb{R}$ such that $S = \{x \in \Xi \mid f(x) \in C\}$. In particular, $C$ is of the form $\{\alpha\}$, $]-\infty, \alpha]$, $[\alpha, +\infty[$, or $[\alpha, \beta[$.

Proof. (i) see Section A.2; (ii) because $S$ is the inverse image of the closed set $C$ under the continuous function $f$. □

Proposition 2.2. [89]

A subset $S$ of a metric space $\Xi$ is compact if any of the following conditions holds

(i) All the closed balls in $\Xi$ are compact and $S$ is closed and bounded.

(ii) $S$ is a closed subset of a compact set.

(iii) $S$ is a finite union or an arbitrary intersection of compact sets.

(iv) There exists a continuous function $f$ of a real variable mapping into $\Xi$ and a compact subset $K$ of $\mathbb{R}$ such that $S = f(K)$. 

Proposition 2.3. [17]

A subset $S$ of a vector space $\Xi$ is convex if any of the following conditions holds

(i) $S$ is an arbitrary intersection of convex sets.

(ii) $S$ is closed and $(\forall (x, y) \in S^2) \quad \frac{1}{2}(x + y) \in S.$

(iii) There exists a real-valued convex function $f$ defined on $\Xi$ and a real number $\alpha$ such that $S$ can be written either as $f^{-1}([\infty, \alpha])$ or $f^{-1}([-\infty, \alpha]).$

(iv) Let $\Xi'$ be another vector space and let $f$ be a linear mapping from $\Xi$ into $\Xi'$. $S$ is the image of a convex set of $\Xi$ under $f$ or the inverse image of a convex set of $\Xi'$ under $f$. A special case is $\Xi' = \mathbb{R}$, where it is known that the intervals are the only nonempty convex sets.

Finally, we shall often use the following theorem in connection with compactness.

Theorem 2.1. [89]

In a metric space $\Xi$ every compact set is closed and bounded. Conversely, every closed and bounded subset of $\Xi$ is compact if and only if the closed balls of $\Xi$ are compact (if $\Xi$ is a normed vector space, this holds if and only if its dimension is finite).

2.2.2. Complements on Banach Spaces

Let $(\Xi, \| \cdot \|)$ be a Banach space. The unit sphere in $\Xi$ is denoted by $U = \{ x \in \Xi \mid \|x\| = 1 \}$. $\Xi$ is said to be strictly convex if

$$ (\forall (x, y) \in U^2) \quad \| \frac{x + y}{2} \| = 1 \implies x = y $$

(2.1)

$\Xi$ is said to be uniformly convex if

$$ (\forall \eta \in \mathbb{R}_+^+) (\exists \delta_\eta \in \mathbb{R}_+)(\forall (x, y) \in U^2) \quad \|x - y\| \geq \eta \implies \| \frac{x + y}{2} \| \leq 1 - \delta_\eta $$

(2.2)
A uniformly convex Banach space $\Xi$ is strictly convex; if the dimension of $\Xi$ is finite the converse is also true. Hilbert spaces and $L^p(\mu)$ spaces (see Section B.2) with $1 < p < +\infty$, are uniformly convex [82]. $\Xi$ is said to be locally uniformly convex if [34]

$$(\forall \eta \in \mathbb{R}_+^\ast) (\forall x \in U) (\exists \delta_{\eta}, z \in \mathbb{R}_+^\ast) (\forall y \in U) \ ||x - y|| \geq \eta \implies \ ||\frac{x + y}{2}|| \leq 1 - \delta_{\eta}, z \quad (2.3)$$

Finally, a uniformly convex Banach space $\Xi$ is said to be smooth if its dual $\Xi'$ is strictly convex [34].

2.2.3. Topological Properties of Set-Valued Maps

We review here a few facts about set-valued maps. For a full account, see Kuratowski [60], [61] and Berge [10]. $\xi^\Xi$ denotes the class of all subsets of a space $\Xi$.

Let $\Xi_1$, $\Xi_2$, and $\Xi_3$ be spaces. A set-valued map from $\Xi_1$ into $\xi^\Xi_3$ is a map $T$ which assigns to each point of $\Xi_1$ a subset of $\Xi_2$. If $\Xi_1 = \Xi_2$, by a fixed point of $T$ we mean any point $x$ in $\Xi_1$ such that $x \in T(x)$. By the image of a subset $S$ of $\Xi_1$ we mean the subset of $\Xi_2$ defined by $TS = \bigcup_{x \in S} T(x)$. If $T_1$ is a set-valued map from $\Xi_1$ into $\xi^\Xi_3$ and $T_2$ a set-valued map from $\Xi_2$ into $\xi^\Xi_3$, the composition $T = T_2 \circ T_1$ is defined by

$$T: \Xi_1 \to \xi^\Xi_3 \quad \quad x \mapsto T_2 T_1(x) = \bigcup_{y \in T_1(x)} T_2(y) \quad (2.4)$$

Let $(\Xi_1, \tau_1)$ and $(\Xi_2, \tau_2)$ be two topological spaces and let $2^\Xi_3$ denote the class of all nonempty closed subsets of $\Xi_2$. Following Kuratowski [60], [61], a set-valued map $T$ from $\Xi_1$ into $2^\Xi_3$ is said to be upper semi-continuous (u.s.c.) at a point $x_0$ in $\Xi_1$ if, for every open neighborhood $V$ of $T(x_0)$, there exists an open neighborhood $U$ of $x_0$ such that
(\forall x \in U) \ T(x) \subseteq V. \ T \text{ is said to be u.s.c. if it is u.s.c. at every point in } \Xi_1 \text{ or, equivalently, if the set } \{x \in \Xi_1 \mid T(x) \subseteq Q\} \text{ is open in } \Xi_1 \text{ for every open set } Q \text{ in } \Xi_2, \text{ that is}

(\forall Q \in \tau_2) \ \{x \in \Xi_1 \mid T(x) \subseteq Q\} \in \tau_1 \quad (2.5)

If \ T \text{ is u.s.c. and if we further assume that } \Xi_2 \text{ is a metric space, then } \ T \text{ is closed in the sense that the set } \{(x,y) \in \Xi_1 \times \Xi_2 \mid y \in T(x)\} \text{ is closed in the topological product } \Xi_1 \times \Xi_2. \text{ Consequently, if } \ T \text{ is closed, the following property holds at every point } x \text{ in } \Xi_1: \text{ if } \{x_n\}_{n \geq 0} \text{ is a sequence of points in } \Xi_1 \text{ converging to } x \text{ and if } \{y_n\}_{n \geq 0} \text{ is a sequence of points in } \Xi_2 \text{ converging to } y \text{ such that } (\forall n \in \mathbb{N}) \ y_n \in T(x_n), \text{ then we have } y \in T(x). \text{ If the metric space } \Xi_2 \text{ is compact, } \ T \text{ is u.s.c. if and only if it is closed.}

In [10], Berge defines u.s.c. maps in a slightly different manner by imposing that, in addition to (2.5), the set } \ T(x) \text{ be compact in } \Xi_2 \text{ for every } x \text{ in } \Xi_1. \text{ Thereafter, we shall call such a map upper Berge semi-continuous (u.B.s.c.).}

**Theorem 2.2.** [10]

Let \( \Xi_1 \) and \( \Xi_2 \) be two topological spaces and let \( T \) be a u.B.s.c. map from \( \Xi_1 \) into \( 2^{\Xi_2} \). Then the image \( TK \) of a compact subset \( K \) of \( \Xi_1 \) is compact in \( \Xi_2 \).

**Theorem 2.3.**

Let \( \Xi_1, \Xi_2, \) and \( \Xi_3 \) be metric spaces. Let \( T_1 \) be a u.s.c. [respectively u.B.s.c.] set-valued map from \( \Xi_1 \) into \( 2^{\Xi_2} \) and let \( T_2 \) be a u.s.c. [respectively u.B.s.c.] set-valued map from \( \Xi_2 \) into \( 2^{\Xi_3} \). Then the composition product \( T = T_2 \circ T_1 \) is a u.s.c. [respectively u.B.s.c.] and closed set-valued map from \( \Xi_1 \) into \( 2^{\Xi_3} \).

**Proof.** Let \( Q \) be an arbitrary open set in \( \Xi_3 \). We need to show that the set \( A = \{x \in \Xi_1 \mid T(x) \subseteq Q\} \) is open in \( \Xi_1 \). We can write \( A = \{x \in \Xi_1 \mid T_1(x) \subseteq Q'\} \) where
$Q' = \{ y \in \Xi_2 \mid T_2(y) \subseteq Q \}$. But since $T_2$ is u.s.c. and $Q$ is open in $\Xi_3$, $Q'$ is open in $\Xi_2$.

Consequently, since $T_1$ is u.s.c., $A$ is open in $\Xi_1$. Hence $T$ is u.s.c. and, since $\Xi_3$ is a metric space, it is closed from $\Xi_1$ into $2^{\Xi_3}$. Now, let $x$ be an arbitrary point in $\Xi_1$. If $T_1$ is u.B.s.c. then $K = T_1(x)$ is compact in $\Xi_2$. If $T_2$ is u.B.s.c. then, by Theorem 2.2, the set $T(x) = T_2K$ is compact in $\Xi_3$, which proves that $T$ is u.B.s.c. from $\Xi_1$ into $2^{\Xi_3}$.  

The above theorem has been stated for the context in which it will be used. It is however noted that it would still be true if $\Xi_1$ and $\Xi_2$ were arbitrary topological spaces.

2.3. Projections in Metric Spaces

In this section, $\Xi$ is a metric space with distance $d$.

2.3.1. Distance to a Set

Let $S_1$ and $S_2$ be two nonempty subsets of $\Xi$. We shall call distance between $S_1$ and $S_2$ the number

$$d(S_1, S_2) = \inf \left\{ d(x_1, x_2) \mid x_1 \in S_1, x_2 \in S_2 \right\}$$

(2.6)

It is important to note that $d(., .)$ does not define a distance on $\xi^\Xi$ [56].

**Theorem 2.4.** [16]

Let $S_1$ and $S_2$ be respectively a nonempty compact and a nonempty closed subset of $\Xi$ such that $d(S_1, S_2) = 0$. Then $S_1 \cap S_2 \neq \emptyset$.

Let $S$ be a nonempty subset of $\Xi$. The distance from a point of $\Xi$ to $S$ is determined by the function
\[ \phi_S : \Xi \to \mathbb{R}_+ \]
\[ x \mapsto d(x, S) = d(\{x\}, S) = \inf \left\{ d(x, y) \mid y \in S \right\} \tag{2.7} \]

**Theorem 2.5.** [16]

Let \( S \) be a nonempty subset of a metric space \( \Xi \). Then

(i) \( \phi_S \) is continuous.

(ii) \( \phi_S \) is nonexpansive: \( (\forall (x, y) \in \Xi^2) \quad |\phi_S(x) - \phi_S(y)| \leq d(x, y) \).

(iii) \((\forall z \in \Xi)\quad \phi_S(z) = 0 \iff z \in \overline{S} \).

It is noted that, for every \( z \) in \( \Xi \), \( \phi_S(z) = \phi_S(x) \).

### 2.3.2. Bounded and Approximative Compactness

A set in a metric space is said to be boundedly compact if its intersection with an arbitrary closed ball is compact; a set in a normed vector space (n.v.s.) is said to be boundedly weakly compact if its intersection with an arbitrary closed ball is weakly compact [24], [107]. The closed ball with center \( x \) and radius \( r \) in \( \Xi \) is denoted by \( B[x, r] \).

**Lemma 2.1.**

Let \( S \) and \( S_1 \) be respectively a nonempty bounded and a nonempty boundedly compact set in a metric space and let \( \alpha \) be a nonnegative real number. Then the set \( \{ z \in S_1 \mid \phi_S(x) \leq \alpha \} \) is compact.

**Proof.** Let \( A = \{ z \in S_1 \mid \phi_S(x) \leq \alpha \} \). We can write \( A = S_1 \cap \phi_S^{-1}([0, \alpha]) \). Since \([0, \alpha]\) is closed and since, by Theorem 2.5, \( \phi_S \) is continuous, \( A \) is closed, as the intersection of two closed sets. Since \( S \) is bounded it is contained in a closed ball centered at some point \( z \) in \( S \), say \( B[z, r] \). Now let \( x \) be an arbitrary point in \( A \). Then there exists a point \( y \) in \( S \) such that
\[d(z,y) \leq \alpha + \frac{1}{2}.\] Hence \(d(z,x) \leq d(z,y) + d(y,x) \leq r + \alpha + \frac{1}{2} = r'.\) Therefore \(x\) belongs to the closed ball \(B\) of center \(z\) and radius \(r'\), and it follows that \(A \subseteq B \cap S_1\). Hence \(A\) is a closed subset of \(B \cap S_1\) which is compact since \(S_1\) is boundedly compact. Thus \(A\) is compact by (ii) in Proposition 2.2. \(\square\)

A set \(S\) in a metric space \(\Xi\) is called approximately compact if, for every \(x\) in \(\Xi\), every sequence \(\{y_n\}_{n \geq 0}\) of points in \(S\) such that \(\{d(x,y_n)\}_{n \geq 0}\) converges to \(\phi_S(x)\) possesses a subsequence converging to a point in \(S\). The notion of approximative compactness was introduced by Efimov and Stechkin for real Banach spaces [41] and was naturally extended to arbitrary metric spaces by Singer [94]. A set \(S\) in a n.v.s. \(\Xi\) is called approximately weakly compact if, for every \(x\) in \(\Xi\), every sequence \(\{y_n\}_{n \geq 0}\) of points in \(S\) such that \(\{d(x',y_n)\}_{n \geq 0}\) converges to \(\phi_S(x)\) possesses a subsequence converging weakly to a point in \(S\) [18], [107].

It is noted that in a n.v.s. a type of compactness implies its weak counterpart; in finite dimensional spaces, the distinction disappears. Conditions under which an approximately weakly compact set is approximately compact in infinite dimension can be found in [18]. Moreover, it follows from the definitions that in a metric space every compact set is boundedly compact, every boundedly compact set is approximately compact, and every approximately compact set is closed. In a n.v.s., the same is true if each type of compactness is replaced by its weak counterpart.

2.3.3. Projection Maps

For every point \(x\) in \(\Xi\), we shall call \(y\) a projection of \(x\) onto a nonempty subset \(S\) of \(\Xi\) if \(y\) belongs to \(S\) and if \(y\) assumes the greatest lower bound in (2.7), i.e. \(\phi_S(x) = d(x,y)\). In particular, by a projection of a vector \(x\) onto a nonempty subset \(S\) of a n.v.s. \((\Xi, ||.||)\), we
mean a vector \( y \) of \( S \) such that

\[
\| x - y \| = \inf \left\{ \| x - z \| \mid z \in \mathcal{S} \right\}
\] (2.8)

In the literature, a projection is also called an element of best approximation. We strongly stress the fact that a projection may not exist or may not be unique. The set-valued projection map \( \Pi_S \) which assigns to each point \( z \) in \( \Xi \) the set of its projections onto a subset \( S \) is defined by

\[
\Pi_S: \Xi \rightarrow \mathcal{C}^S
\]

\[
x \mapsto \left\{ y \in S \mid \phi_S(x) = d(x, y) \right\}
\] (2.9)

**Lemma 2.2.**

Let \( S \) be a nonempty subset of a metric space \( \Xi \). Then

\[
(\forall x \in \Xi) \quad \Pi_S(x) = B[x, \phi_S(x)] \cap S
\] (2.10)

**Proof.** Fix \( x \) in \( \Xi \) and let \( B = B[x, \phi_S(x)] \). Then \( \Pi_S(x) \subseteq S \cap B \) follows from (2.7) and (2.9). Conversely, let \( z \in S \cap B \). Then \( d(z, z) \leq \inf \{ d(x, y) \mid y \in S \} = \phi_S(x) \) since \( z \) belongs to \( S \). But, since \( z \) also belongs to \( B \), \( \phi_S(x) = d(z, z) \). Hence \( z \in \Pi_S(z) \). \( \square \)

**Theorem 2.6.**

Let \( S \) be a nonempty subset of a metric space \( \Xi \) and let \( \Pi_S \) denote the projection map onto \( S \). Then, for every point \( x \) in \( \Xi \)

(i) \( \Pi_S(x) \) is bounded.

(ii) \( \Pi_S(x) \) is closed if \( S \) is closed.

(iii) \( \Pi_S(x) \) is compact if \( S \) is boundedly compact.
Proof. (i) obvious by Lemma 2.2; (ii) for, by Lemma 2.2, it is the intersection of two closed sets; (iii) obvious from Lemma 2.2 and the definition of bounded compactness. □

2.3.4. Existence of a Projection

Let \( S \) be a nonempty subset of a metric space \( \Xi \). We say that \( S \) is proximinal (or is an existence set) if every point in \( \Xi \) has at least one projection onto \( S \) i.e., with our notations, \( (\forall x \in \Xi) \quad \Pi_S(x) \neq \emptyset \). It is noted that a proximinal set is necessarily closed since no point in \( \overline{S} - S \) has a projection, which forces \( \overline{S} = S \). Therefore, if \( S \) is proximinal, \( \Pi_S(x) \) will be nonempty, closed in \( S \), and bounded for every \( x \) in \( \Xi \) by Theorem 2.6. Thus, the map \( \Pi_S \) defined in (2.9) actually maps \( \Xi \) into \( 2^S \). If \( S \) is proximinal, we say that an operator \( \gamma \) from \( \Xi \) into \( S \) is a selection of \( \Pi_S \) if \( \gamma(x) \in \Pi_S(x) \), for all \( x \) in \( \Xi \).

Theorem 2.7.

Let \( S \) be a nonempty subset of a metric space \( \Xi \). Then \( S \) is proximinal if any of the following conditions holds

(i) \( S \) is approximately compact.

(ii) \( S \) is boundedly compact.

(iii) \( S \) is compact.

(iv) \( S \) is closed and all the closed balls are compact in \( \Xi \).

Proof. (i) see [41]; (ii) see [62] or regard it a special case of (i) [41]; (iii) see [16] or regard it as a special case of (ii); (iv) see [89] or regard it as a special case of (ii): if \( B \) is an arbitrary closed ball in \( \Xi \), then \( S \cap B \) is a closed subset of \( B \), and \( B \) is compact. Thus \( S \cap B \) is compact, which makes \( S \) boundedly compact. □
In the special case of n.v.s.'s, several additional existence criteria can be established.

**Theorem 2.8.**

Let $S$ be a nonempty subset of a n.v.s. $E$. Then $S$ is proximinal if any of the following conditions holds

(i) $S$ is approximately weakly compact.

(ii) $S$ is boundedly weakly compact.

(iii) $S$ is weakly compact.

(iv) $S$ is a finite dimensional vector subspace.

(v) $S$ is sequentially weakly closed and $E$ is a uniformly convex Banach space.

(vi) $S$ is closed and $E$ has finite dimension.

**Proof.** (i) see [18]; (ii) special case of (i); (iii) special case of (ii); (iv) see [62], [95]; (v) for $S$ is then approximately compact [18], [41]; (vi) special case of (iv) in Theorem 2.7 by Theorem 2.1 (hence, such an $S$ is boundedly compact). □

**Corollary 2.1.**

In a finite dimensional n.v.s., the class of nonempty approximately compact sets, the class of nonempty boundedly compact sets, the class of proximinal sets, and the class of nonempty closed sets coincide.

**Proof.** From Section 2.3.2, a boundedly compact set is approximately compact. From (i) in Theorem 2.7, a nonempty approximately compact set is proximinal. From a previous remark, proximinal sets are closed. Finally, from the proof of (vi) in Theorem 2.8, a closed subset of a finite dimensional n.v.s. is boundedly compact. □
2.3.5. Uniqueness of a Projection

A nonempty subset $S$ of a metric space $\mathfrak{X}$ is called a Chebyshev set if every point in $\mathfrak{X}$ has exactly one projection onto $S$. By the projection operator onto a Chebyshev subset $S$ of $\mathfrak{X}$, we mean the function $\pi_S$ from $\mathfrak{X}$ onto $S$ which maps every point $x$ into its unique projection onto $S$ ($\pi_S$ is the unique selection of $\Pi_S$).

**Theorem 2.9.** [45]

Let $S$ be a nonempty closed and convex set in a uniformly convex Banach space. Then

(i) $S$ is a Chebyshev set.

(ii) $\pi_S$ is continuous.

**Theorem 2.10.** [17], [101]

Let $S$ be a nonempty convex and complete subset of a pre-Hilbert space $(\mathfrak{X}, \langle \cdot, \cdot \rangle)$. Then

(i) $S$ is a Chebyshev set.

(ii) $\pi_S$ is continuous and nonexpansive: $(\forall (x, y) \in \mathfrak{X}^2) \ ||\pi_S(x) - \pi_S(y)|| \leq ||x - y||$.

(iii) Characterization of $\pi_S$: $(\forall x \in \mathfrak{X})(\forall y \in S) \ \text{Re}(\langle x - \pi_S(x), y - \pi_S(x) \rangle) \leq 0$.

In particular, this theorem applies when $S$ is closed and convex and $\mathfrak{X}$ is a Hilbert space; when $S$ is a Banach subspace and $\mathfrak{X}$ is a pre-Hilbert space; when $S$ is a closed vector subspace and $\mathfrak{X}$ is a Hilbert space (in particular $S$ is a finite dimensional vector subspace).

The properties of the operator $\pi_S$ in that latter case are the object of the next theorem.

**Theorem 2.11.** [17]

Let $S$ be a closed vector subspace of a Hilbert space $(\mathfrak{X}, \langle \cdot, \cdot \rangle)$. Then

(i) $(\forall x \in \mathfrak{X})(\exists! y \in S)(\exists! z \perp S) \ x = y + z$, furthermore $y = \pi_S(x)$.

(ii) $\pi_S$ is nonexpansive, linear, and idempotent: $\pi_S \circ \pi_S = \pi_S$. 
(iii) \( S \neq \{0\} \implies \|\pi_S\| = 1. \)

(iv) \( \pi_S \) is self-adjoint: \((\forall (x,y) \in \Xi^2)\) \( <\pi_S(x),y> = <x,\pi_S(y)> \).

Let \( \pi_1 \) and \( \pi_2 \) be respectively the projection operators onto the closed vector subspaces \( S_1 \) and \( S_2 \) of \( \Xi \) and let \( V = \{x + y \mid x \in S_1, y \in S_2\} \). Then

(v) If \( \pi_1 \circ \pi_2 = \pi_2 \circ \pi_1 \), then \( V \) is closed and we have \( \pi_V = \pi_1 + \pi_2 - \pi_1 \circ \pi_2 \).

(vi) \( \pi_1 \circ \pi_2 = 0 \iff S_1 \perp S_2 \). If so, \( V \) is closed and we have \( \pi_V = \pi_1 + \pi_2 \).

2.3.6. Some Remarks on Chebyshev Sets

Upon reading the previous section, it is legitimate to ask whether a Chebyshev set has to be convex. The answer is negative, for nonconvex Chebyshev sets can be constructed [54]. However, Motzkin has showed that in the Euclidean plane a nonempty set is a Chebyshev set if and only if it is closed and convex [71]. Motzkin's result was then generalized to the Euclidean space in [53]. It was then established that, in a finite dimensional Banach space \( \Xi \), the class of Chebyshev sets coincides with the class of nonempty closed and convex sets if and only if \( \Xi \) is strictly convex and smooth [40]. In infinite dimensional spaces, the discussion is more involved and the reader is referred to [6], [8], [58], [95], and [107].

It is also worth noting that a Chebyshev set need not be approximately compact. However, a Chebyshev set in a uniformly convex smooth Banach space is convex if and only if it is approximately compact [41].

2.3.7. Continuity of Projections

Let \( S \) be a proximinal set in a metric space \( \Xi \). In Section 2.3.4 it was noted that the range of the projection map onto \( S \) is in fact \( 2^S \). Hence, from this point on, by projection map onto \( S \), it is meant the set-valued projection map with domain \( \Xi \) and range \( 2^S \) of (2.9).

Moreover, when we say that the projection map onto \( S \) is u.s.c. [respectively u.B.s.c.] it is
understood that we mean u.s.c. [respectively u.B.s.c.] relative to the topology of $\Xi$ and the corresponding induced topology of $S$.

**Theorem 2.12.**

Let $S$ be an arbitrary nonempty set in a metric space. The projection map onto $S$ is denoted by $\Pi_S$ and, if $S$ is a Chebyshev set, the projection operator onto $S$ is denoted by $\pi_S$. Then

(i) $\Pi_S$ is closed if $S$ is proximinal.

(ii) $\Pi_S$ is u.s.c. if $S$ is approximately compact.

(iii) $\Pi_S$ is u.B.s.c. if $S$ is boundedly compact.

(iv) $\pi_S$ is continuous if $S$ is an approximately compact Chebyshev set.

**Proof.** (i), (ii), and (iv) see [94]; (iii) $\Pi_S$ is already u.s.c. by (ii) and $\Pi_S(x)$ is compact for all $x$ from (iii) in Theorem 2.6. $\square$

### 2.3.8. Characterization of a Projection

The following characterization is given in [24]. Let $S$ be a nonempty subset of a n.v.s. $(\Xi,||.||)$, $y_0$ a point in $S$, and $z$ a point in $S^c$. Then $y_0$ belongs to $\Pi_S(z)$ if and only if it is a fixed point of the map $\Gamma_z$ defined by

$$\Gamma_z: S \rightarrow \xi^S$$

$$y \mapsto S \cap B\left[ x, \frac{\phi_S(x) + ||x-y||}{2} \right]$$

Moreover, in this case, we have $\Gamma_z(y_0) \subset \Pi_S(z)$. The problem of characterizing a projection in a n.v.s. is also addressed in [62], where specific cases are studied.
2.3.9. Projections onto Monotonic Sequences of Sets

**Theorem 2.13.** [89]

Let \( \{S_n\}_{n \geq 0} \) be a sequence of nonempty closed and convex sets in a Hilbert space \( E \). Then

\[
\left( \forall x \in \Xi \right) \lim_{n \to +\infty} \phi_{S_n}(x) = \phi_S(x) \quad \text{and} \quad \lim_{n \to +\infty} \pi_{S_n}(x) = \pi_S(x) \tag{2.12}
\]

whenever either one of the following propositions holds

(i) \( \{S_n\}_{n \geq 0} \) is a decreasing sequence with \( S = \bigcap_{n \geq 0} S_n \neq \emptyset \).

(ii) \( \{S_n\}_{n \geq 0} \) is an increasing sequence with \( S = \bigcup_{n \geq 0} S_n \).

2.3.10. A Property of the Points with Multiple Projections

Let \( S \) be a proximinal set in a locally uniformly convex Banach space \( \Xi \) and let \( M \) denote the set of points which have more than one projection onto \( S \). Then \( M \) is a set of the first category, i.e. a countable union of nowhere dense sets [107]. Therefore, since \( \Xi \) is a complete metric space, it may be concluded that the set \( M^c \) of points with exactly one projection onto \( S \) is everywhere dense: the smallest closed set containing \( M^c \) is \( \Xi \) itself.

2.3.11. A Few Counterexamples

In this section, we have gathered a few cases which illustrate the necessity of the assumptions on which the theorems rest and the fact that, in some respects, the results cannot be improved.

- Two disjoint sets \( S_1 \) and \( S_2 \) in a metric space such that \( d(S_1,S_2)=0 \): take \( S_1 \) an hyperbola, and \( S_2 \) one of its asymptote in the Euclidean plane (compare with Theorem 2.4).

- A point \( x \) with no projection onto a bounded set \( S \) in a finite dimensional n.v.s.: take
the Euclidean real line, $S = [−1, 1]$, and $x = 2$. Then $ϕ_S(x) = 1$ but $x$ possesses no projection onto $S$.

- A point $x$ with infinitely many projections onto a closed and convex subset $S$ in a Banach space: take $\mathbb{R}^2$ with the distance $d_1$ defined in (2.14), $S = B[0, 1]$, and $x = (1, 1)$. Then $\Pi_S(x) = \{((1−\alpha) \mid 0 ≤ \alpha ≤ 1\}$ (compare with Theorem 2.9).

- A point with no projection onto a nonempty closed set: see [101].

- A point with no projection onto a nonempty closed and bounded set: see [89].

- A proximinal set in a real Hilbert space which is not approximately compact and a projection mapping onto it which is not u.s.c.: see [94].

- A Chebyshev set in a real pre-Hilbert space which is nonconvex and for which the projection operator is continuous: see [54].

- A projection operator $\pi_S$ which is not continuous, where $S$ is a Chebyshev vector subspace in a Banach space: see [50].

- A projection operator $\pi_S$ which is not nonexpansive, where $S$ is compact and convex in a finite dimensional n.v.s.: see [35].

2.4. Projections in the Cartesian Space

2.4.1. Generalities

Let $k$ be a positive integer. The Cartesian space $\mathbb{R}^k$ is the space of real $k$-tuples $(x_1, \ldots, x_k)$ and has the structure of a $k$-dimensional vector space over $\mathbb{R}$, where addition and scalar multiplication are defined in the usual manner. The open segment delimited by two distinct points $x$ and $y$ is denoted by $\text{seg}(x, y)$ and defined as

$$\text{seg}(x, y) = \{\alpha x + (1−\alpha)y \mid 0 < \alpha < 1\}$$

(2.13)

By applying (B.9) to the integral defined in (B.12) it follows that, for every positive
extended real number $p$, $(\mathbb{R}^k, d_p)$ possesses the structure of a complete metric space with

$$d_p(x, y) = \begin{cases} \sum_{i=1}^{k} |x_i - y_i|^p & \text{if } 0 < p < 1 \\ \left( \sum_{i=1}^{k} |x_i - y_i|^p \right)^{1/p} & \text{if } 1 \leq p < +\infty \\ \sup_{1 \leq i \leq k} \left( |x_i - y_i| \right) & \text{if } p = +\infty \end{cases}$$

(2.14)

Moreover, for $1 \leq p \leq +\infty$, the distance induces a norm and $(\mathbb{R}^k, d_p)$ is a Banach space. In particular, $(\mathbb{R}^k, d_p)$ is a uniformly (and therefore strictly) convex Banach space for $1 < p < +\infty$. For $p = 2$, it becomes a Hilbert space which is referred to as the $k$-dimensional Euclidean space and denoted by $E^k$. It is worth noting that a broader class of Hilbert spaces can be obtained on $\mathbb{R}^k$ with the following distance

$$(\forall (x, y) \in \mathbb{R}^k \times \mathbb{R}^k) \quad d(x, y) = \sum_{i=1}^{k} \sum_{j=1}^{k} a_{ij} x_i y_j$$

(2.15)

where $[a_{ij}]$ is a symmetric positive definite $k \times k$ matrix. Figure 2.1 represents the sphere $\{x \in \mathbb{R}^2 \mid d_p(0, x) = 1\}$ for various values of $p$. It is clear that the notion of proximity strongly depends on the metric. Figure 2.1 also provides an illustration of the notion of strict convexity: if $1 < p < +\infty$, every open segment in the unit ball is disjoint from the unit sphere.

2.4.2. Ordinary and Multifurcation Points

Historically, Bouligand has been the first to study the properties of points with multiple projections onto a nonempty closed subset of the Euclidean space $E^k$ [13]. Upon becoming aware of their importance in Euclidean geometry, he proceeded to categorize the points in the space into what he calls ordinary points, i.e. points with a unique projection, and mul-
Figure 2.1. Points at Distance 1 from the Origin for Various p's.

tifurcation points, i.e. points with more than one projection (in Section 2.3.10, the set of multifurcation points was denoted by $M$). A few years later, Motzkin gave a characterization of the closed and convex subsets of the Euclidean plane in terms of multifurcation points [71], a result which was generalized to the Euclidean space by Jessen [53]. Then Pauc gave a more detailed geometrical analysis for the case of the Euclidean plane [75] and subsequently for the Euclidean space [76]. The next major contribution is due to Erdős who approached the problem from a measure theoretic point of view [42].

2.4.3. Existence, Uniqueness, and Continuity of Projections

If $1 \leq p \leq +\infty$, Corollary 2.1 states that the class of nonempty closed subsets, the class of nonempty boundedly compact sets, and the class of proximinal sets coincide in $(\mathbb{R}^k, d_p)$. 
Thus, the projection map onto a nonempty closed subset is u.B.s.c. by (iii) in Theorem 2.12. For $1 < p < +\infty$, $(\mathbb{R}^k, d_p)$ is a uniformly convex Banach space and therefore, by Theorem 2.9, every point has a unique projection onto a nonempty closed and convex subset $S$ and the projection operator $\pi_S$ is continuous. But, in this case, the dual of $(\mathbb{R}^k, d_p)$ is $(\mathbb{R}^k, d_q)$, where $p^{-1} + q^{-1} = 1$, and it is a strictly convex Banach space as well. Hence, $(\mathbb{R}^k, d_p)$ is a strictly convex smooth finite dimensional Banach space and, by a result of Section 2.3.6, the only Chebyshev sets are the nonempty closed and convex sets. In addition, by (ii) in Theorem 2.10, $\pi_S$ is nonexpansive for $p = 2$.

2.4.4. The Reach, the Convex Deficiency, and the Skeleton of a Set

The concepts of reach, convex deficiency and skeleton will be introduced here because of their important connection with nonunique projections. In the following, $S$ is a nonempty subset of the Euclidean space $E^k$.

The notion of the reach of a set was introduced by Federer in [43]. The reach of $S$ is the largest $\eta$ in $[0, +\infty]$ such that, if $x$ is a point of the space such that $\phi_S(x) < \eta$, then $x$ is an ordinary point with respect to $S$. We have

(i) $\text{reach}(S) > 0 \implies S$ is closed

(ii) $\text{reach}(S) = +\infty \iff S$ is closed and convex

Loosely speaking, $x$ is within the reach of $S$ if it can be projected unambiguously onto $S$.

Now we suppose that $S$ is closed and we denote by $C$ its convex hull, i.e. the smallest convex set containing $S$. The convex deficiency of $S$ is defined as $D = C - S$ [27]. A point $x$ in $S^c$ is called a skeletal point of $S$ if $B[x, \phi_S(x)]$ is contained in no other $B[x', \phi_S(x')]$ [27]. The skeleton of $S$ is the set of all its skeletal points.
2.4.5. Properties of Ordinary and Multifurcation Points in the Euclidean Space

The Lebesgue measure in $E^k$ is denoted by $\lambda$.

**Proposition 2.4.**

Let $S$ be a nonempty closed subset of the Euclidean space $E^k$. In the following, the terms projection, ordinary and multifurcation points are relative to $S$, $K$ is the skeleton of $S$ and $M$ the set of its multifurcation points. Then

(i) If $x$ is an arbitrary point in $E^k$ and $y$ one of its projections, every point in $\text{seg}(x,y)$ is an ordinary point, i.e., $\text{seg}(x,y) \cap M = \emptyset$.

(ii) If $x_1$ and $x_2$ are two multifurcation points and if $y_1$ and $y_2$ are one of their respective projections, then $\text{seg}(x_1,y_1) \cap \text{seg}(x_2,y_2) = \emptyset$.

(iii) The multifurcation points $x$ such that there is no $(n-1)$-dimensional hyperplane containing all the segments $\{\text{seg}(x,y) \mid y \in \Pi_S(x)\}$ are at most countable.

(iv) $M = \emptyset$ $\iff$ $S$ is convex.

(v) $M$ is dense in $K$.

(vi) A point $y$ in $S$ is an interior point of $\{x \in \Xi \mid \phi_S(x) = d(x,y)\}$ if and only if it is isolated in $S$ (i.e., $\exists r \in \mathbb{R}^*_+$ s.t. $S \cap B(y,r) = \{y\}$).

(vii) For every $y$ in $S$, $\{x \in \Xi \mid \phi_S(x) = d(x,y)\}$ is closed and convex and contains $y$.

(viii) For every $y$ in $S$, $\{x \in \Xi \mid \phi_S(x) = d(x,y)\}$ is bounded if and only if $y$ is an interior point of the convex hull of $S$.

(ix) $\pi_S$ is a continuous projection operator on $M^c$.

(x) Every selection of $\Pi_S$ is $\lambda$-a.e. differentiable.

(xi) $\forall r \in \mathbb{R}^*_+$ $\lambda \left( \{x \in E^k \mid \phi_S(x) = r\} \right) = 0$.

(xii) $\lambda \left( \bigcup_{x \not\in S} \Pi_S(x) \right) = 0$. 
The last property is due to Erdös and is certainly the most remarkable: in the Euclidean space, the projections onto a nonempty closed set are Lebesgue almost everywhere unique. Loosely speaking, $M$ can be covered by a countable number of $k$-dimensional cubes whose total volume is arbitrarily small. We can also give a probabilistic interpretation of this result. Consider the probability space $(\Omega, \mathcal{F}, P)$ where $\Omega$ is $\mathbb{E}^k$, $\mathcal{F}$ the corresponding Borel $\sigma$-algebra, and $P$ any probability measure which is absolutely continuous with respect to the Lebesgue measure in $\mathbb{E}^k$. In this context, every point $x$ in $\mathbb{E}^k$ has almost surely a unique projection (the probability of the event "$x$ is a multifurcation point" is zero).

2.5. The Method of Successive Projections in Hilbert Spaces

The method of successive projections (MOSP) in Hilbert spaces rests to a large extent on Theorem 2.10.

**Definition 2.1.**

Let $\{S_1, \ldots, S_m\}$ be a collection of nonempty closed and convex subsets of a Hilbert space $\Xi$. For all $i$ in $\{1, \ldots, m\}$ we denote by $\pi_i$ the projection operator onto the (Chebyshev) set $S_i$ and by $\pi$ the composition $\pi_1 \circ \pi_2 \circ \ldots \circ \pi_m$. Given a point $x_0$ in $\Xi$, we shall call a sequence of successive projections (SOSP) the sequence $\{x_n\}_{n \geq 0}$ constructed according to the recursion

$$ (\forall n \in \mathbb{N}) \quad x_{n+1} = \pi(x_n) = \pi^{n+1}(x_0) $$

(2.16)
Proposition 2.5.

Suppose that, in addition to the hypotheses of Definition 2.1, the intersection $S$ of the sets is nonempty. Then, if the SOSP converges, it is to a point in $S$.

Proof. By (ii) in Theorem 2.10, each $\pi_i$ is continuous. Hence $\pi$ is continuous and, if the SOSP $\{x_n\}_{n \geq 0}$ converges to a point $x$, $\{\pi(x_n)\}_{n \geq 0}$ converges to $\pi(x)$. Consequently, by (2.16), $\{x_{n+1}\}_{n \geq 0}$ converges to $\pi(x)$. Hence, $x$ is necessarily a fixed point of $\pi$, i.e. $x = \pi(x)$. But since every fixed point of $\pi$ is in $S$ [113], it follows that $x$ is in $S$. □

The main convergence results of the MOSP in Hilbert spaces can now be stated.

Theorem 2.14.

Let $\{S_1, \ldots, S_m\}$ be a collection of closed and convex subsets of a Hilbert space $E$ whose intersection $S$ is not empty. Then, for every $x_0$ in $E$,

(i) The SOSP converges weakly to a point in $S$.

(ii) The SOSP converges strongly to a point in $S$ if one of the sets is boundedly compact.

(iii) The SOSP converges strongly to a point in $S$ if all the sets are linear varieties.

Proof. (i) is due to Brégman [19] and (ii) to Stiles [99]; (iii) was proved by Halperin [48] for vector subspaces but his proof can routinely be extended to linear varieties (i.e. images under translations of vector subspaces). □

For $m = 2$, Halperin's result is known as the Alternating Projection Theorem and is due to Von Neumann [108]. If the dimension of $E$ is finite, (i) and (ii) are equivalent and (iii) is a particular case of (i) because then a closed set is boundedly compact and the notions of weak and strong convergence coincide. It is worth noting that the operator $\pi$ is nonexpan-
sive since, by (ii) in Theorem 2.10, each \( \pi_i \) is nonexpansive. Thus, the above theorem fits into the general problem of the convergence of a sequence of successive approximations \( \{T^n(x_0)\}_{n \geq 0} \) to a fixed point of \( T \), where \( T \) is a nonexpansive mapping. Such convergence properties are discussed in [78].

For completeness, let us mention that, in order to improve the speed of convergence of the MOSP, relaxed projection operators were introduced in [46], which extend the projections beyond the boundary of the sets. It is also noted that in the MOSP presented above the sets are activated in cyclic order at each iteration. Other schemes have been considered in the literature, which aim at an optimal speed of convergence. In that respect, a unified framework for the study of a very broad class of projection algorithms, along with strong convergence results, were recently presented in [74]. Two salient features characterize the algorithms presented there: at each step a convex linear combination of the projections onto selected sets is formed and an optimal relaxation coefficient is computed. By updating these parameters at each step, the speed of convergence is optimized.

### 2.6. The Method of Successive Projections in Metric Spaces

As was seen in Section 2.1, the framework of the MOSP as described in the previous section is unsuitable for a wide class of problems. Our purpose here is to broaden the scope of the MOSP by placing ourselves in the general setting of a metric space where \( \Gamma = \{S_1, \ldots, S_m\} \) is a collection of proximinal sets. We shall first define a SOSP in such a context and then proceed to establish conditions under which a SOSP would converge to a solution, i.e. a point in the intersection of the \( S_i \)'s.
2.6.1. The Cyclic Projection Map

Definition 2.2.
Let $\Gamma = \{S_1, S_2, \ldots, S_m\}$ be an ordered collection of proximinal sets in a metric space $\Xi$. For every $i$ in $\{1, 2, \ldots, m\}$, we denote by $\Pi_i$ the projection map onto $S_i$, regarded as a set-valued map from $\Xi$ into $2^\Xi$. Then the composition map $\Pi = \Pi_1 \circ \Pi_2 \circ \ldots \circ \Pi_m$ will be called the cyclic projection map of $\Gamma$.

Theorem 2.15.
The cyclic projection map of an arbitrary ordered collection of nonempty approximately [respectively boundedly] compact sets $\{S_1, \ldots, S_m\}$ in a metric space $\Xi$ is a u.s.c. [respectively u.B.s.c.] map from $\Xi$ into $2^\Xi$.

Proof. Let $\Pi = \Pi_1 \circ \ldots \circ \Pi_m$ be the cyclic projection map of $\{S_1, \ldots, S_m\}$ and let $i$ be a fixed integer in $\{1, \ldots, m\}$. By Theorem 2.12, the projection map onto the approximately [respectively boundedly] compact set $S_i$ is u.s.c. [respectively u.B.s.c.] from $\Xi$ into $2^{S_i}$. From the definition of the relative topology, an open set in $S_i$ is the intersection of an open set in $\Xi$ with $S_i$. It follows at once that, if $S_i$ is approximately [respectively boundedly] compact, $\Pi_i$ is u.s.c. [respectively u.B.s.c.] from $\Xi$ into $2^\Xi$. By using Theorem 2.3 inductively, the finite composition product $\Pi = \Pi_1 \circ \ldots \circ \Pi_m$ is also is u.s.c. [respectively u.B.s.c.] from $\Xi$ into $2^\Xi$. \qed

2.6.2. The Sequence of Successive Projections

Definition 2.3.
Let $\Pi$ be the cyclic projection map of an ordered collection of proximinal sets $\Gamma = \{S_1, \ldots, S_m\}$ in a metric space $\Xi$. Then, given a point $x_0$ in $\Xi$, we shall call a SOSP
(relative to $\Gamma$ and $x_0$) any sequence $\{x_n\}_{n \geq 0}$ constructed according to the recursion

$$\forall n \in \mathbb{N} \quad x_{n+1} \in \Pi(x_n) \tag{2.17}$$

In words, one selects a projection $y_m$ of $x_0$ onto $S_m$, then a projection $y_{m-1}$ of $y_m$ onto $S_{m-1}$ and so on. The projection of $y_2$ onto $S_1$ which has been selected is $x_1$. A SOSP $\{x_n\}_{n \geq 0}$ is constructed by continuing this cyclic process \textit{ad infinitum}. It is noted that such a sequence exists since $(\forall x \in \Xi) \quad \Pi(x) \neq \emptyset$. Moreover, since the values of $\Pi$ are subsets of $S_1$, the sequence $\{x_n\}_{n \geq 1}$ lies in $S_1$.

For a given collection $\Gamma$ and a point $x_0$, the uniqueness of a SOSP depends on the properties of the sets in the region where the iterations are performed. Clearly, for every starting point $x_0$, the cyclic projection map of a collection of Chebyshev sets generates a unique SOSP. In connection with the uniqueness question, let us recall two important properties of a space which is frequently encountered in applications, namely the finite dimensional Euclidean space $E$. First, by a result of Section 2.3.6, the class of Chebyshev sets and the class of nonempty closed and convex sets coincide in $E$. Second, by (xiii) in Proposition 2.4, the points which admit more than one projection onto a nonempty closed subset of $E$ form a set of Lebesgue measure zero.

Proposition 2.6.

Let $\Pi$ be the cyclic projection map of an ordered finite collection of nonempty approximately compact sets $\Gamma = \{S_1, \ldots, S_m\}$ in a metric space $\Xi$. Then, if a SOSP converges, it is to a fixed point of $\Pi$.

\textbf{Proof.} Suppose that $\{x_n\}_{n \geq 0}$ is a SOSP that converges to a point $x$. Then the sequence $\{x_{n+1}\}_{n \geq 0}$ converges also to $x$. But, since $(\forall n \in \mathbb{N}) \quad x_{n+1} \in \Pi(x_n)$, and since $\Pi$ is u.s.c. by
Theorem 2.15 and therefore closed, we get that \( x \in \Pi(x) \). \( \square \)

Unlike Proposition 2.5, the above proposition does not guarantee that the limit of a convergent SOSP is a solution point. Indeed, \( \Pi \) may admit fixed points outside the solution set. In that respect, Figure 2.2 shows a system of two closed sets in the Euclidean plane where \( x \) is a fixed point of the cyclic projection map \( \Pi \) which lies outside the intersection \( S \). In that example, the set of fixed points of \( \Pi \) is \( F = \{x\} \cup S \).

![Figure 2.2. Fixed Points and Region of Attraction.](image)

Definition 2.4.

Let \( \Gamma = \{S_1, \ldots, S_m\} \) be an ordered collection of proximinal sets in a metric space \( \Xi \) whose intersection \( S \) is not empty. Let \( \Pi \) be the cyclic projection map of \( \Gamma \) and let

\[
Y = \left\{ x \in S_1 - S \mid \exists \ z' \in \Pi(x) \text{ such that } \phi_S(z') \geq \phi_S(x) \right\}
\]

(2.18)

We shall define the radius of attraction of \( \Gamma \) as
\[ p = \begin{cases} \inf \left\{ \phi_S(x) \mid x \in Y \right\} & \text{if } Y \neq \emptyset \\ +\infty & \text{otherwise} \end{cases} \] (2.19)

and the region of attraction of \( \Gamma \) as

\[ R = S \cup \left\{ x \in S_1 \mid \phi_S(x) < \rho \right\} \] (2.20)

Thus, the region of attraction is a subset of \( S_1 \) containing \( S \). It is important to note that, under our assumptions on the collection of sets \( \Gamma \), the radius of attraction \( \rho \) can assume all the values in \([0, +\infty]\). For instance, the case \( \rho = +\infty \) occurs when \( \Gamma \) consists of two intersecting straight lines in the Euclidean plane. From (2.20), it is seen that if \( \rho = +\infty \) then the region of attraction is \( S_1 \) itself. Figure 2.2 illustrates a case where \( 0 < \rho < +\infty \) for a collection \( \Gamma = \{S_1, S_2\} \) of closed subsets of the Euclidean plane. The shaded area represents the region of attraction \( R \) of \( \Gamma \). In the case where \( \rho = 0 \), the region of attraction reduces to the solution set \( S \). Since this case is less obvious, we shall now construct an example to show that it can actually occur.

**Example.**

Let \( \Xi \) be the Euclidean real line. For every \( n \) in \( \mathbb{N} \), we define

\[
\begin{aligned}
  a_n &= \frac{1}{n} \quad \text{and} \quad b_n = \frac{3n + 2}{3n(n + 1)} \quad \text{if} \quad n > 0 \\
  a_0 &= b_0 = 0
\end{aligned}
\] (2.21)

Consider the collection \( \Gamma = \{S_1, S_2\} \) in \( \Xi \) where

\[
S_1 = \bigcup_{n \geq 0} \{a_n\} \quad \text{and} \quad S_2 = \bigcup_{n \geq 0} \{b_n\} \] (2.22)

We claim that the radius of attraction of \( \Gamma \) is zero.
Proof. First of all, we check that $\Gamma$ satisfies the assumptions of Definition 2.4. It is readily seen that

$$ (\forall n \in \mathbb{N}) \quad n > 0 \quad \Rightarrow \quad a_{n+1} < b_n < a_n $$

Thus $S = \{0\} \neq \emptyset$. Moreover, the discrete sets $S_1$ and $S_2$ are closed in the finite dimensional n.v.s. $\Xi$ and are therefore proximinal by Corollary 2.1. Now, let $\epsilon$ be a fixed positive real number. Then, there exists an $n$ in $\mathbb{N}$ such that $1/(n+1) < \epsilon$. Whence, there exists an $x = a_{n+1}$ in $S_1 - S$ such that $\phi_S(x) < \epsilon$. Let $\Pi = \Pi_1 \circ \Pi_2$ and $\rho$ be respectively the cyclic projection map and the radius of attraction of $\Gamma$. From (2.23), a projection of $x$ onto $S_2$ is either $b_n$ or $b_{n+1}$. It is easily computed that $|a_{n+1} - b_{n+1}| < |a_{n+1} - b_n|$. Hence, $\Pi_2(x) = \{b_{n+1}\}$. But, from (2.23), a projection of $b_{n+1}$ onto $S_1$ is either $a_{n+1}$ or $a_{n+2}$. However, since $|b_{n+1} - a_{n+1}| < |b_{n+1} - a_{n+2}|$, we have $\Pi_1(b_{n+1}) = \{a_{n+1}\} = \{x\}$. We conclude that $\Pi(x) = \{x\}$ and, hence, $x$ lies in the set $Y$ of (2.18). Thus, necessarily, $\rho \leq \epsilon$. Since $\epsilon$ can be arbitrarily small, it follows that $\rho = 0$. \qed

Two direct consequences of Definition 2.4 are

$$ (\forall x \in R - S)(\forall x' \in \Pi(x)) \quad \phi_S(x') < \phi_S(x) $$

and

$$ (\forall x \in R) \quad \Pi(x) \subset R $$

An interpretation of this last equation is that $R$ is a trapping set.

Definition 2.5.

Let $\Gamma = \{S_1, \ldots, S_m\}$ be an ordered collection of proximinal sets in a metric space $\Xi$ whose intersection $S$ is nonempty. Let $R$ and $\Pi$ be respectively the region of attraction and the cyclic projection map of $\Gamma$. We shall say that a point $x_0$ in $\Xi$ is a point of attraction of $\Gamma$
if for every SOSP \( \{x_n\}_{n=0} \) there exists a nonnegative integer \( v \) such that \( x_v \) belongs to \( R \).

We shall call the smallest such \( v \) the index of attraction of a given SOSP.

It follows readily from Definition 2.5, (2.24), and (2.25) that, if \( x_0 \) is a point of attraction and \( \{x_n\}_{n=0} \) a SOSP with index of attraction \( v \), then \( \{\phi_S(x_n)\}_{n \geq v} \) is a nonincreasing sequence and

\[
(\forall n \in \mathbb{N}) \quad x_{v+n} \in \left\{ x \in S_1 \mid \phi_S(x) \leq \phi_S(x_v) \right\} \subset R
\]

(2.26)

In words, the tail of every SOSP starting at a point of attraction lies in the region of attraction. Moreover, (2.24) implies that all the fixed points of \( \Pi \) in \( R \) belong to the solution set. Thus, if all the sets are approximately compact in Definition 2.5, the limit of every convergent SOSP starting at a point of attraction is a solution point on account of Proposition 2.6.

2.6.3. A Convergence Result

We remind the reader that a topological space is said to be connected if it is not the union of two disjoint nonempty closed sets and that a compact connected Hausdorff topological space is said to be a continuum [61]. Moreover, we shall say that a continuum is nontrivial if it does not reduce to \( \emptyset \) or a singleton.

**Theorem 2.16.**

Let \( \{x_n\}_{n=0} \) be a sequence of points in a compact subset of a metric space \((\Xi, d)\) such that \( \{d(x_n, x_{n+1})\}_{n=0} \) converges to zero. Then either \( \{x_n\}_{n=0} \) converges or its set of cluster points is a nontrivial continuum.

**Proof.** A straightforward generalization of a theorem given in [73] for Euclidean spaces.
We now proceed to present our main result in connection with the convergence of a SOSP to a common point of sets in metric spaces.

**Theorem 2.17.**

Let \( \Gamma=\{S_1,\ldots,S_m\} \) be an ordered collection of approximately compact sets in a metric space \( \Xi \), whose intersection \( S \) is nonempty and bounded, and such that \( S_1 \) is boundedly compact. Let \( x_0 \) be a point of attraction of \( \Gamma \), let \( \{x_n\}_{n \geq 0} \) be an arbitrary SOSP, and let \( C \) be the set of its cluster points. Then either \( \{x_n\}_{n \geq 0} \) converges to a point in \( S \) or \( C \) is a nontrivial continuum in \( S \). A sufficient condition for the former is \( \sum_{n \geq 0} \phi_S(x_n) < +\infty \).

**Proof.** Let \( \Pi=\Pi_1\circ\ldots\circ\Pi_m \) be the cyclic projection map of \( \Gamma \). Since \( x_0 \) is a point of attraction of \( \Gamma \), the SOSP \( \{x_n\}_{n \geq 0} \) possesses an index of attraction \( \nu \). From (2.26), the sequence \( \{x_n\}_{n \geq \nu} \) lies in the set \( A=\{z \in S_1 \mid \phi_S(z) \leq \phi_S(x_\nu)\} \), which is compact by Lemma 2.1. Hence, \( \{x_n\}_{n \geq 0} \) admits a cluster point \( y \). By continuity of \( \phi_S \), \( \phi_S(y) \) is a cluster point of \( \{\phi_S(x_n)\}_{n \geq 0} \). But, as \( \{\phi_S(x_n)\}_{n \geq 1} \) is a nonincreasing sequence bounded from below, it must converge to \( \phi_S(y) \). Moreover \( (\forall n \in \mathbb{N}) \; \phi_S(x_{\nu+n}) \leq \phi_S(y) \). Now suppose \( y \notin S \). Then, since \( y \in A \subset R \), (2.24) yields \( (\forall y' \in \Pi(y)) \; \phi_S(y') < \phi_S(y) \). Let \( V=\{z \in \Xi \mid \phi_S(z) < \phi_S(y)\} \). Then \( V=\phi_S^{-1}([0,\phi_S(y)]) \) is open in \( \Xi \) as the inverse image under the continuous function \( \phi_S \) of the (relatively) open set \( [0,\phi_S(y)] \) of \( \mathbb{R}_+ \). Hence, \( V \) is an open neighborhood of \( \Pi(y) \) since \( \Pi(y) \subset V \). \( \Pi \) is u.s.c. at \( y \) by Theorem 2.15. It follows that there exists an open neighborhood \( U \) of \( y \) such that \( (\forall z \in U) \; \Pi(z) \subset V \). But \( y \) is a cluster point of \( \{x_n\}_{n \geq 1} \), and, hence, there exists a positive integer \( n \) such that \( x_{\nu+n-1} \in U \). The successor \( x_{\nu+n} \) of \( x_{\nu+n-1} \) in the SOSP belongs to \( \Pi(x_{\nu+n-1}) \) and therefore to \( V \). Consequently \( \phi_S(x_{\nu+n}) < \phi_S(y) \), which contradicts a previous inequality statement. Hence, \( y \in S \). Thus, \( \{\phi_S(x_n)\}_{n \geq 0} \) converges to zero and \( \emptyset \neq C \subset S \). Now, let \( n \) be fixed in \( \mathbb{N} \) and let \( y_0=x_n \). For every \( j \) in
\{0, \ldots, m-1\}, we denote by \( y_{j+1} \) the projection of \( y_j \) onto \( S_{m-j} \) which has been selected in the process of obtaining \( x_{n+1} \). We have \( x_{n+1} = y_m \). It is noted that \( S \) is boundedly compact as a closed subset of \( S_1 \). Hence, by (ii) in Theorem 2.7, there exists a point \( z \) in \( S \) which is a projection of \( y_0 \) onto \( S \). Since \( z \) belongs to \( S \), it belongs to all the \( S_{m-j} \)'s, and thus

\[
(\forall j \in \{0, \ldots, m-1\}) \quad d(y_j, y_{j+1}) = \inf \{d(y_j, y) \mid y \in S_{m-j} \} \leq d(y_j, z) \tag{2.27}
\]

Let \( d = d(y_0, z) \). Let us prove that

\[
(\forall j \in \{0, \ldots, m\}) \quad d(y_j, z) \leq 2^j d \tag{2.28}
\]

The statement is clearly true for \( j = 0 \). For any integer \( j \) in \( \{0, \ldots, m-1\} \) for which (2.28) holds, (2.27) yields

\[
d(y_{j+1}, z) \leq d(y_j, y_{j+1}) + d(y_j, z) \leq 2d(y_j, z) \leq 2^{j+1} d \tag{2.29}
\]

which completes the proof by induction. By using (2.27) and (2.28), we get

\[
d(x_n, x_{n+1}) \leq \sum_{j=0}^{m-1} d(y_j, y_{j+1}) \leq \sum_{j=0}^{m-1} d(y_j, z) \leq \sum_{j=0}^{m-1} 2^j d = (2^m - 1)d \tag{2.30}
\]

It is easy to see that \( d = \phi_S(x_n) \). Therefore

\[
(\forall n \in \mathbb{N}) \quad d(x_n, x_{n+1}) \leq (2^m - 1)\phi_S(x_n) \tag{2.31}
\]

Since \( \{\phi_S(x_n)\}_{n=0}^\infty \) converges to zero, so does \( \{d(x_n, x_{n+1})\}_{n=0}^\infty \). Thus, since \( C \subset S \), it follows from Theorem 2.16 that either \( \{x_n\}_{n=0}^\infty \) converges to a point in \( S \) (\( C \) is then a singleton) or \( C \) is a nontrivial continuum in \( S \). To prove the last assertion, let \( p \) and \( q \) be any two nonnegative integers such that \( p < q \). We have

\[
d(x_p, x_q) \leq \sum_{n=p}^{q-1} d(x_n, x_{n+1}) \leq \sum_{n=p}^{q-1} d(x_n, x_{n+1}) \leq (2^m - 1)\phi_S(x_n) \tag{2.32}
\]

Since the rightmost expression in (2.32) is nothing but the tail of a convergent series, it
must go to zero as \( p \) goes to infinity. Consequently, \( d(x_p, x_q) \) goes to zero as \( p \) and \( q \) go to infinity and \( \{x_n\}_{n \geq 0} \) is a Cauchy sequence. But, if a Cauchy sequence \( \{x_n\}_{n \geq 0} \) possesses a cluster point \( y \), it converges to \( y \) [15]. Since \( \emptyset \neq C \subseteq S \), we conclude that \( \{x_n\}_{n \geq 0} \) converges to a point in \( S \). \( \square \)

We now present an alternative proof of the fact that every cluster point of a SOSP \( \{x_n\}_{n \geq 0} \) is in the solution set.

**Proof.** Let \( \nu \) be the index of attraction of \( \{x_n\}_{n \geq 0} \) and let \( A = \{x \in S_1 \mid \phi_S(x) \leq \phi_S(x_\nu)\} \).

Since \( A \) is compact we can extract a subsequence \( \{x_{n_k}\}_{k \geq 0} \) from \( \{x_n\}_{n \geq \nu} \) converging to a point \( y \). Moreover, \( \phi_S(y) \) is the limit of the convergent sequence \( \{\phi_S(x_n)\}_{n \geq 0} \). For every \( k \) in \( \mathbb{N} \), let \( y_k \) denote the successor of \( x_{n_k} \) in \( \{x_n\}_{n \geq 0} \), that is \( y_k = x_{n_k+1} \). Then \( \{y_k\}_{k \geq 0} \) lies in \( A \) and it thereby possesses a subsequence \( \{y_{k_j}\}_{j \geq 0} \) converging to a point \( y' \). By construction, \( y' \) is also a cluster point of \( \{x_n\}_{n \geq 0} \) and, thus, \( \phi_S(y') \) is a cluster point of \( \{\phi_S(x_n)\}_{n \geq 0} \), which leads to \( \phi_S(y') = \phi_S(y) \). For every \( j \) in \( \mathbb{N} \), let \( z_j \) denote the predecessor of \( y_{k_j} \) in \( \{x_n\}_{n \geq 0} \). Then \( \{z_j\}_{j \geq 0} \) is a subsequence of \( \{x_{n_k}\}_{k \geq 0} \) and it must converge to \( y \).

But \( \{y_{k_j}\}_{j \geq 0} \) converges to \( y' \) and \( (\forall j \in \mathbb{N}) \ y_{k_j} \in \Pi(z_j) \). By Theorem 2.15, \( \Pi \) is u.s.c. and a fortiori closed, which forces \( y' \in \Pi(y) \). Now suppose \( y \notin S \). Since \( y \in A \subseteq R \), (2.24) yields \( \phi_S(y') < \phi_S(y) \), which contradicts a previous statement. Whence, \( y \in S \). \( \square \)

**Comments.** In the proof of Theorem 2.17 it is shown that, without any summability assumption on \( \{\phi_S(x_n)\}_{n \geq 0} \), the sequence \( \{d(x_n, x_{n+1})\}_{n \geq 0} \) converges to zero. It is noted that in an ultrametric space, i.e. a metric space \((\Xi, d)\) such that

\[
(\forall (x, y, z) \in \Xi^3) \quad d(x, z) \leq \text{Sup} \left\{ d(x, y), d(y, z) \right\}
\] (2.33)
this condition guarantees that \( \{x_n\}_{n \geq 0} \) is a Cauchy sequence [16], which would establish at once its convergence. Ultrametric spaces are however seldom encountered in applications (see Bourbaki [16], [17] and Dieudonné [37] for examples and properties). It is also worth noting that if the sequence \( \{x_n\}_{n \geq 0} \) was not a SOSP, not even the summability of \( \{\phi_S(x_n)\}_{n \geq 0} \) would guarantee its convergence to a point in \( S \), as illustrated by the following example. Let \( \Xi \) be the Euclidean plane and let \( S = \{(x_1, x_2) \in \mathbb{R}^2 : \text{Sup}\{|x_1|, |x_2|\} \leq 1\} \) be the unit square. Now, let \( (\forall n \in \mathbb{N}) \ x_n = (1 + 2^{-n}, (-1)^n) \). Then, the sequence \( \{x_n\}_{n \geq 0} \) admits two distinct cluster points (1,1) and (1,-1). Whence, it does not converge. Nonetheless

\[
\sum_{n \geq 0} \phi_S(x_n) = \sum_{n \geq 0} 2^{-n} = 2 < +\infty
\]  

(2.34)

**Corollary 2.2.**

Let \( \Gamma = \{S_1, \ldots, S_m\} \) be an ordered collection of closed sets in a finite dimensional normed vector space, whose intersection \( S \) is nonempty and bounded. Let \( x_0 \) be a point of attraction of \( \Gamma \), let \( \{x_n\}_{n \geq 0} \) be an arbitrary SOSP, and let \( C \) be the set of its cluster points. Then either \( \{x_n\}_{n \geq 0} \) converges to a point in \( S \) or \( C \) is a nontrivial continuum in \( S \). A sufficient condition for the former is \( \sum_{n \geq 0} \phi_S(x_n) < +\infty \).

**Proof.** By Corollary 2.1, Theorem 2.17 applies. \( \square \)

**2.6.4. Questions Relative to the Implementation of the MOSP**

In the previous sections, the MOSP has been developed as a conceptual mathematical procedure for obtaining a point in the intersection \( S \) of a collection \( \Gamma = \{S_1, \ldots, S_m\} \) of sets. We shall now address the main issues pertaining to its implementation. In the following discussion, it is assumed that \( \Xi = \mathbb{R}^n \), as is the case in most applied problems. Moreover,
because of its computational advantages, the Euclidean norm $\|\cdot\|$ is chosen to metricize $\mathbb{R}^n$. Therefore, the main result of interest here is Corollary 2.2.

2.6.4.1. Closedness and Boundedness Restrictions on the Sets

We have $\overline{S}_i = \{ z \in \mathbb{R} \mid d(z, S_i) = 0 \}$. Hence, the requirement that the sets be closed is not restrictive since we can always replace a property set by its closure. Moreover, the condition that the solution set be bounded should not cause concern for, in practice, there are always boundedness constraints on the components of a feasible solution. Thus, the conditions of closedness and boundedness placed on the sets in Corollary 2.2 are quite mild and can always be satisfied in applied problems.

2.6.4.2. Computation of the Projections

For a fixed $z$ in $\mathbb{R}^n$, let $\phi_z$ denote the functional $\phi_z : y \mapsto \|z - y\|$. Each elementary step of the MOSP involves the computation of a projection $y_p$ of a point $z$ onto a proximinal subset $S_i$ of $\mathbb{R}^n$, i.e. a global minimum of $\phi_z$ over $S_i$. As there exists no universal method to solve efficiently this quadratic minimization problem, a complete discussion is neither possible nor intended. In practice, the projection onto a given $S_i$ should be considered on a case-by-case basis.

If $S_i$ is convex, any local minimum of $\phi_z$ is a global one. Specific algorithms have been established for special cases such as polyhedrons [5] or polytopes (i.e. the convex hull of a finite number of points) [111]. Generally speaking, quadratic programming algorithms are of interest when $S_i$ is specified by functional inequalities. In applied problems, an equation is often available for the boundary of $S_i$ (e.g. $\partial S_i = \{ z \in \mathbb{R}^n \mid g_i(z) = 0 \}$). Since the projec-
tion of a point in $S_i^c$ onto $S_i$ belongs to $\partial S_i$, the problem can be approached via the method of Lagrange multipliers. If $S_i$ is characterized only by a contact function, the quadratic programming algorithm of [9] can be used. If $S_i$ is not convex, $\phi_z$ may admit several local minima and, thus, the problem of finding a projection of $x$ is one of global optimization. In some problems, there may not be criteria for deciding whether a local solution is global and global optimization methods must be employed. Both deterministic and stochastic procedures have been proposed in the literature to solve the global optimization problem and we refer the reader to [4] for a detailed survey. Recent developments of the stochastic approach can be found in [84] and in the references therein. Stochastic methods should be used when there is no certainty that deterministic algorithms will produce a global minimum to some acceptable degree of accuracy. They are regarded as very reliable tools which, under relatively mild conditions, offer an asymptotic guarantee of convergence (in some probabilistic sense) to a global solution.

2.6.4.3. Finding a Point of Attraction

In Corollary 2.2, it is stated that the iterations should be started at a point of attraction. Since our definition of a point of attraction is not constructive, such a point may be difficult to find. Indeed, because of the geometrical complexity of the system of sets, it is usually impractical to characterize a priori the region of attraction and to establish whether or not a point is a point of attraction. This potential limitation of the MOSP should however be mitigated by noting that in a practical application an approximate solution is often available. Since points of attraction are more likely to be found in the vicinity of the solution set, this approximate solution is a good candidate for a starting point of a SOSP. Then, the point $y$ produced by the algorithm is accepted as a solution if it belongs to all the property sets. If it does not, then none of the points in the
corresponding SOSP belong to the region of attraction. Hence, loosely speaking, a new starting point should be chosen outside the path followed by that unsuccessful SOSP. A heuristic way to do this is to start the iterations at the symmetric $x'_0$ of $x_0$ with respect to $y$, i.e. $x'_0 = 2y - x_0$. In general, this method does not ensure that a point of attraction will be found. Nonetheless, it has proven successful in most of our applications.

2.6.4.4. Stopping Rule

In practical problems, the MOSP presented above consists in generating a sequence of points $\{x_n\}_{n \geq 0}$, where $x_0$ is a point of attraction of $\Gamma$, according to the algorithm

\[
(\forall n \in \mathbb{N}) \quad \begin{cases} 
\text{Stop if } (\forall i \in \{1, \ldots, m\}) \ x_n \in S_i \\
\text{Let } x_{n+1} \text{ be any point in } \Pi(x_n) \text{ else}
\end{cases} \quad (2.35)
\]

Under the hypotheses of Corollary 2.2, if a solution is not obtained in a finite number of iterations, it is any cluster point or the limit of $\{x_n\}_{n \geq 0}$. Since we are interested in obtaining a solution in finite time, it remains to discuss a practical convergence criterion for the MOSP.

In the proof of Theorem 2.17, it was seen that, if $x_0$ is a point of attraction of $\Gamma$ and $\{x_n\}_{n \geq 0}$ a SOSP, then the sequence $\{||x_{n+1} - x_n||\}_{n \geq 0}$ converges to zero. Therefore, the following criterion can be used as a stopping rule

\[
||x_{n+1} - x_n|| \leq \epsilon 
\]

(2.36)

This criterion has notorious drawbacks. It may for instance stop the algorithm prematurely in case of slow convergence. Another problem is that the choice of the parameter $\epsilon$ is somewhat arbitrary. We shall however adopt it because it has proven useful in practice.

It is noted that the MOSP does not necessarily yield a solution point in a literal sense since, because of the truncation of the SOSP, it may produce a point which does not
exactly lie in all the $S_i$’s. From a practical standpoint, this should not raise concern because slight deviations are usually allowable in specifying the boundaries of the $S_i$’s.

2.7. Summary

A method for obtaining a common point of a finite collection of sets in a metric space has been presented. This method (MOSP) was developed by generalizing the method of successive projections for closed and convex sets in Hilbert spaces to approximately compact sets in metric spaces. Sequences of successive projections were constructed via the composition of the set-valued projection maps onto the individual sets. It was shown that, under certain hypotheses, a solution could be obtained either as a cluster point or as the limit of any such sequence. Potential applications of this result are found in any problem whose set theoretic format cannot be reduced to that of closed and convex sets in a Hilbert space, which is a common instance in many digital signal processing applications.

The MOSP will be applied to digital signal restoration in Chapter V.
3.1. Introduction

The question of synthesizing set theoretic estimates, i.e. the question of finding a common point of sets in an abstract space, can naturally be regarded as a search problem. In Chapter II, this search problem was approached via a deterministic procedure, namely the method of successive projections (MOSP). Although this method can be applied to a wide spectrum of set theoretic estimation problems, it bears inherent limitations, which leaves the generation of set theoretic estimates an open question in many cases.

In sciences, it is common to resort to random strategies when deterministic methods are unsuccessful. In this chapter, we shall adopt this approach and employ a random search method to generate set theoretic estimates. In that method, it is assumed that the Cartesian space $\mathbb{R}^k$ is the solution space. Points are generated uniformly at random over a restricted region $G$ of $\mathbb{R}^k$. Because $G$ is constructed so that it intersects with the feasible set $S$ in a set of positive Lebesgue measure, this procedure is bound to yield a solution with probability one.

The limitations of the MOSP are exhibited in Section 3.2. In Section 3.3, the literature on random search methods is briefly reviewed. The main contribution of this chapter is in Section 3.4, where we develop a method of random search (MORS) for synthesizing set
theoretic estimates. An analysis of the merits and shortcomings of the MORS is provided in Section 3.5. Alternative search strategies for the MORS are considered in Section 3.6. In Section 3.7, a random search technique for the fuzzy set theoretic estimation problem is outlined. Finally, recommendations for the use of the MORS in digital signal processing appear in Section 3.8.

3.2. The Limitations of the Method of Successive Projections

Even in $\mathbb{R}^k$, there are limitations to the use of the MOSP. The mathematical restrictions for the convergence of the MOSP in $\mathbb{R}^k$ were stated in Corollary 2.2 and discussed in Section 2.6.4. It was seen that the closedness and boundedness requirements on the property sets were of no practical consequence. It was also pointed out that, in cases where little geometric description of the property sets is available, finding a point of attraction to start the iterations could be a problem.

From the methodological standpoint, the chief limitation of the MOSP is that it operates via projections. Indeed, as was seen in Section 2.6.4.2, carrying out the projection onto a property set may, in some cases, amount to solving a very delicate global minimization problem. As a result, in such instances, the user may elect to give up a valuable piece of $a$ priori knowledge because of the complexity associated with the analytical derivation and the numerical implementation of the projection onto the corresponding set.

In light of these elements, it becomes apparent that set theoretic estimation cannot rely exclusively on the MOSP for the synthesis of solutions. It is therefore of utmost importance to introduce an alternative method.
3.3. A Brief Review of the Random Search Literature

Random search methods have been employed mainly in optimization theory, where they have become a prominent tool for solving the basic problem of finding a global minimum over a compact subset $G$ of a functional $\phi$ on $\mathbb{R}^k$ and have been reported to outperform their deterministic counterparts. The basic version of that method, which was proposed in [23], is often called the pure random search method. It consists in drawing independently $n$ points from a uniform distribution over $G$ and selecting the point which yields the smallest value of $\phi$ as a solution. In order to increase the speed of convergence of the method, improvements have been proposed, which introduce some degree of determinism and render the search adaptive. The early random search techniques based on this principle are surveyed in [96]. In [98], general convergence results for more sophisticated algorithms are given and convergence rates are examined. Further developments of the stochastic global optimization problem appear in [84]. It is worth noting that most convergence proofs rely on the assumption that the functional $\phi$ is continuous on $G$.

Recently, a new stochastic algorithm, called simulated annealing, was proposed in [57] for solving complex multiextremal optimization problems with discrete space. This algorithm generates a discrete Markov chain whose state space is the domain of the cost function to be minimized and whose transition probabilities are adaptively controlled by a so-called temperature parameter.

Outside the field of optimization, the only rigorous treatment of the general theory of random search seems to be that given by Rényi in [83].
3.4. The Method of Random Search in Set Theoretic Estimation

3.4.1. Preliminaries

Throughout this chapter, \( \lambda \) denotes the Lebesgue measure in \( \mathbb{R}^k \) and \( \Sigma \) denotes the Borel \( \sigma \)-algebra of \( \mathbb{R}^k \). Let \( A \) be a Borel subset of \( \mathbb{R}^k \) and \( f \) a multivariate density with support \( A \) (i.e. \( f = 0 \) on \( A^c \)). We shall often use the expression "drawing a point \( x \) from \( A \) according to \( f \)". It is understood that the mathematical interpretation of this statement is that \( x \) is an observation of the realization of a random vector \( X=(X_1,...,X_k) \) defined on \( (\mathbb{R}^k, \Sigma) \) which possesses density \( f \). The probability measure on \( (\mathbb{R}^k, \Sigma) \) will be denoted by \( P \) and defined as that induced by \( X \), that is

\[
(\forall B \in \Sigma) \quad P(B) = \int_B f \, d\lambda
\]  

(3.1)

In common words, \( P(B) \) will be the probability that a point drawn from \( A \) according to the density \( f \) lies in \( B \). In particular, suppose that \( 0 < \lambda(A) < +\infty \) and let \( f \) be the uniform density on \( A \), i.e. \( f = (\lambda(A))^{-1} 1_A \), where \( 1_A \) is the indicator function of \( A \). Then, (3.1) yields

\[
(\forall B \in \Sigma) \quad P(B) = \int_B \frac{1}{\lambda(A)} 1_A \, d\lambda = \frac{1}{\lambda(A)} \int_A 1_A \, d\lambda = \frac{\lambda(A \cap B)}{\lambda(A)}
\]  

(3.2)

3.4.2. Methodology

\( \Gamma = \{S_1,...,S_m\} \) is the collection of Borel property sets for a given set theoretic estimation problem in \( \mathbb{R}^k \). It is assumed that the problem is in a format such that the solution set \( S \), i.e. the intersection of all the property sets, has positive Lebesgue measure. Moreover, the sets in \( \Gamma \) are ordered by nondecreasing Lebesgue measure, i.e.
\[ \lambda(S_1) \leq \cdots \leq \lambda(S_m) \] (3.3)

and it is assumed that \( \lambda(S_1) < +\infty \).

**Definition 3.1.**

We shall call a search region for \( \Gamma \) a Borel subset \( G \) of \( \mathbb{R}^k \) of finite Lebesgue measure such that \( S \cap G \) has positive Lebesgue measure.

Practically speaking, the above definition stipulates that \( G \) should be a region where, by generating points uniformly, there is a nonzero probability of hitting \( S \). It is noted that we do not require that \( G \) be bounded. The task of generating a set theoretic estimate can simply be achieved by performing a pure random search in the following manner. Given a search region \( G \) for \( \Gamma \), points are drawn independently at random from a uniform distribution over \( G \). The procedure is stopped when a point in \( S \) is found. Henceforth, this method will be called the method of random search (MORS). The algorithmic description of the MORS is as follows.

1. Generate \( z \) from a uniform distribution over \( G \) and set \( i = 1 \).
2. If \( z \not\in S_i \), return to 1.
3. If \( i \neq m \), set \( i = i + 1 \) and return to 2.
4. Stop.

**3.4.3. Basic Probabilistic Analysis**

In this section, we describe some elementary properties of the MORS. \( G \) is a search region for \( \Gamma \).
Definition 3.2.

The acceptance rate of the MORS will be defined as the probability $\gamma$ that a point drawn from a uniform distribution over the search region is in the solution set. The factor $1-\gamma$ will be called the rejection rate of the MORS.

From (3.2), the acceptance rate is trivially given by

$$\gamma = \frac{\lambda(S \cap G)}{\lambda(G)} \quad (3.4)$$

Each trial of the MORS can be regarded as a Bernoulli experiment where the outcome is a success with probability $\gamma$. An ideal case where $G$ would have all its measure in $S$, e.g. $G \subset S$, would yield $\gamma = 1$. Since such a case is of no theoretical interest, it is ruled out. Hence, under our assumptions, the rejection rate lies in $]0,1[$.

Proposition 3.1.

The MORS produces a solution with probability one.

Proof. Let $A$ be the event "the MORS produces a solution eventually" and, for every positive integer $n$, let $B_n$ be the event "the MORS produces no solution in $n$ trials". All the trials are independent and the probability of producing no solution on a single trial is $1-\gamma$. Hence, $P(B_n) = (1-\gamma)^n$. But $\{B_n\}_{n\geq 1}$ is a decreasing sequence of events and, thus, it converges to

$$\lim_{n \to +\infty} B_n = \bigcap_{n \geq 1} B_n = A^c \quad (3.5)$$

By the continuity theorem [69], it follows that
\[ p = P\left( B_n^c \right) = 1 - (1 - \gamma)^n \]  

(3.7)

Whence, it is readily established that the number of trials required to find at least one solution point with probability \( p \) is

\[ n = \frac{\ln(1 - p)}{\ln(1 - \gamma)} \]  

(3.8)

Another way to describe the performance of the MORS is to consider the number of trials \( \eta \) required to obtain the first success. Clearly, \( \eta \) has a geometric distribution and takes positive value \( n \) with probability \( \gamma(1 - \gamma)^{n-1} \). Consequently, the expected value of \( \eta \) is

\[ \bar{\eta} = \frac{1}{\gamma} \]  

(3.9)

Therefore, in order to minimize the number of trials until a solution is found, the rejection rate should be minimized.

3.4.4. Considerations for the Determination of a Search Region

The search region \( G \) for \( \Gamma \) should be determined so as to minimize the search time, an objective which involves two main factors. First, the geometry of \( G \) should be simple so that performing a single trial, i.e. generating a point uniformly over \( G \), can be achieved at low computational cost. Second, \( G \) should yield a low rejection rate, which would in turn guarantee that the number of trials required to find a solution is small. Oftentimes, an antinomy arises between these two requirements and, therefore, a compromise is in order.
3.4.4.1. Uniform Generation of Points

Direct methods for efficiently generating points uniformly over a subset $G$ of the Cartesian space are limited to cases where $G$ is a simple region (hyperrectangle, hyperparallelogram, simplex, hypersphere, hyperellipsoid). We shall now review these methods for various regions, by increasing order of computational complexity.

Before we proceed, let us recall an important result. Let $T: \mathbb{R}^k \mapsto (T_1(x), \ldots, T_k(x))$ be a continuous transformation from $\mathbb{R}^k$ into $\mathbb{R}^k$ whose restriction to an open set $A$ admits an inverse $T^{-1}$ with continuous first order partial derivatives and nonzero Jacobian $J$. Let $X$ be a random vector with multivariate density $f$ supported by a Borel subset $G$ of $A$. Then $T(X)$ possesses a multivariate density $f'$ on $T(G)$ given by (e.g. [110])

$$f' = |J| f \circ T^{-1}$$

(3.10)

In particular, if $T$ is a nonsingular affine transformation (i.e. $T:x \mapsto Mx + z$, where $M$ is an invertible matrix and $z$ a vector), $|J| = |\det(M)|^{-1} = \lambda(G)/\lambda(T(G)) > 0$. Hence, if $X$ is uniformly distributed on $G$, (3.10) yields

$$f' = \frac{\lambda(G)}{\lambda(T(G))} \cdot \frac{1}{\lambda(G)} \cdot 1_G \circ T^{-1} = \frac{1}{\lambda(T(G))} 1_T(G)$$

(3.11)

This proves that $T(X)$ will be uniformly distributed on $T(G)$. In other words, uniformity is preserved under nonsingular affine transformations.

First, let us consider the most simple case, i.e. when $G$ is the hyperrectangle

$$G = [a_1, b_1] \times \cdots \times [a_k, b_k]$$

(3.12)

Then, the random vector $X = (X_1, \ldots, X_k)$, where the $X_i$'s are independent random variables (r.v.'s), each $X_i$ being uniformly distributed over $[a_i, b_i]$, is uniformly distributed over $G$. 
If $G$ is a right hyperparallelogram, it is the image under a rotation $M$ of a hyperrectangle $G'$. Thus, the random vector $MX$, where $X$ is uniformly distributed over $G'$, will be uniformly distributed over $G$.

Now suppose that $G$ is the elementary simplex

$$G = \left\{ x \in \mathbb{R}^k \mid (\forall i \in \{1, \ldots, k\}) \quad x_i \geq 0 \quad , \quad \sum_{i=1}^{k} x_i \leq 1 \right\}$$  \hspace{1cm} (3.13)

Let $\{Y_1, \ldots, Y_{k+1}\}$ be independent and identically distributed (i.i.d.) r.v.'s, with an exponential density of parameter 1. Then the random vector

$$X = \left( \frac{Y_1}{\sum_{i=1}^{k+1} Y_i}, \ldots, \frac{Y_k}{\sum_{i=1}^{k+1} Y_i} \right)$$  \hspace{1cm} (3.14)

is uniformly distributed over $G$ [86] (for the generation of exponential variates, the reader is referred to [36]). Now, let us consider an arbitrary simplex $G'$ in $\mathbb{R}^k$ defined by its $k+1$ vertices, say $z_i = (z_{i1}, \ldots, z_{ik})$, for $0 \leq i \leq k$. Let $M$ denote the matrix of general term

$$M_{ij} = z_{ij} - z_{0j} \quad 1 \leq i, j \leq k$$  \hspace{1cm} (3.15)

Then, if $X$ is uniformly over the elementary simplex of (3.13), $MX + z_0$ will be uniformly distributed over $G'$.

Let $G$ denote the closed Euclidean ball of center $z$ and radius $r$ and let $\{Y_1, \ldots, Y_k\}$ be i.i.d. normal r.v.'s, with mean 0 and variance 1 (for the generation of normal variates, the reader is referred to [36]). Then the random vector

$$X = \left( \frac{Y_1}{(\sum_{i=1}^{k} Y_i^2)^{1/2}}, \ldots, \frac{Y_k}{(\sum_{i=1}^{k} Y_i^2)^{1/2}} \right)$$  \hspace{1cm} (3.16)
is uniformly distributed on the unit Euclidean hypersphere \([86]\)

\[
E = \left\{ x \in \mathbb{R}^k \mid \sum_{i=1}^{k} x_i^2 = 1 \right\}
\] (3.17)

Consequently, the random vector \(rU^{1/k}X + z\), where the random vector \(X\) is uniformly distributed on \(E\) and where the r.v. \(U\) is uniformly distributed over \([0,1]\), is uniformly distributed over \(G\). An extension for the uniform generation of points in a hyperellipsoid is also given in \([86]\).

In order to consider more complex regions, we shall need the following proposition.

**Proposition 3.2.**

Let \(\{G_n\}_{n \geq 0}\) be a finite or countable collection of disjoint Borel subsets of \(\mathbb{R}^k\) of positive Lebesgue measure such that their union \(G\) has finite Lebesgue measure. For every \(n\) in \(\mathbb{N}\), let \(f_n\) denote the uniform density over \(G_n\). Then, the uniform density over \(G\) is given by

\[
f = \sum_{n \geq 0} \frac{\lambda(G_n)}{\lambda(G)} f_n
\] (3.18)

**Proof.** By definition

\[
(\forall n \in \mathbb{N}) \quad f_n = \frac{1}{\lambda(G_n)} 1_{G_n}
\] (3.19)

Hence, (3.18) gives

\[
f = \frac{1}{\lambda(G)} \sum_{n \geq 0} 1_{G_n}
\] (3.20)

But, since the \(G_n\)'s are disjoint

\[
\sum_{n \geq 0} 1_{G_n} = 1 \bigcup_{n \geq 0} G_n = 1_G
\] (3.21)
Thus

\[ f = \frac{1}{\lambda(G)} 1_G \]  

(3.22)

which completes the proof. □

This result provides an efficient means to generate uniformly over a complex region \( G \) which can be expressed as the union of disjoint simple regions. In practice, when used in that context, the partition of \( G \) will be finite. Theoretically, Proposition 3.2 also provides a means to generate points uniformly over an unbounded region \( G \) of \( \mathbb{R}^k \) whose Lebesgue measure is finite. In that context, the partition will be countable.

For sake of completeness, let us mention that a technique for generating points which, asymptotically, cover uniformly an arbitrary bounded Borel subset \( G \) of \( \mathbb{R}^k \) has been proposed in [97]. Although this technique places no geometric restrictions on \( G \) in principle, it has, from our viewpoint, a major shortcoming. Namely, if \( G \) is too complex (in particular nonconvex), the determination at each iteration of the line set described in [97] will be computationally costly.

### 3.4.4.2. Minimization of the Rejection Rate

As was seen above, the search region \( G \) should be chosen so that the rejection rate is as low as possible. Loosely speaking, (3.4) indicates that the efficiency of the MORS depends on how well \( G \) overlaps with \( S \). The difficulty of this problem is that \( S \) is unknown. If \( G \) is determined in a too conservative manner, the rejection rate will be high. On the other hand, if \( G \) is too small, it may not intersect with \( S \).

The solution set \( S \) is the intersection of all the property sets and, therefore, it is contained
in each of them. Hence, the best safe choice for the search region is one of the property sets with the smallest Lebesgue measure i.e., by (3.3), $S_1$.

3.4.4.3. Conclusions

In terms of rejection rate, the best choice for $G$ is $S_1$. However, efficient uniform generation of points over $G$ requires that $G$ have a simple geometry. There is no definite answer as to how the best compromise between these two conditions can be reached in a general set theoretic estimation problem. Nonetheless, we can propose general guidelines.

If $S_1$ is a simple region, then, clearly, it is the best candidate for a search region. In general, each property set $S_i$ should be inserted into a set $R_i$ determined as follows. First of all, if $S_i$ is convex, $R_i$ should be the smallest simple region which covers it. If $S_i$ is not convex, it will be inefficiently covered by a single simple region. Hence, to achieve a better covering, $R_i$ should take the form of a disjoint union of simple regions. Then, given a collection of covering sets $\{R_1,...,R_m\}$, $G$ should be chosen as the $R_i$ of smallest Lebesgue measure. If $G$ turns out to be a union of simple regions, the uniform generation should be carried out according to Proposition 3.2.

To keep the cost of the uniform generation as low as possible, in choosing the $R_i$’s, hyperrectangles should be considered first. If they do not provide a good boxing of the sets, hyperparallelograms should be considered next, then simplexes, and finally hyperspheres and ellipsoids. In selecting $G$, the main objective, namely the minimization of the overall search time, should be borne in mind. For example, in selecting a simplex rather than a hyperrectangle, it should be made sure that the improvement incurred by a better rejection rate will not be offset by the increase in the cost of the uniform generation.
3.5. Advantages and Limitations of the Method of Random Search

3.5.1. The Advantages

The most obvious advantages of the MORS are its simplicity and ease of implementation, as evidenced by the algorithmic description of Section 3.4.2. Most of all, it truly provides total flexibility with regard to the incorporation of the \textit{a priori} knowledge since no topological or geometrical restrictions are placed on the property sets. Moreover, in the MORS, using a property set $S_i$ reduces to testing points for membership in $S_i$, which is a much easier task than projecting onto $S_i$, as is done in the MOSP. Hence, there is no limitation on the analytical complexity of the pieces of \textit{a priori} knowledge that can be involved in the description of the solution. Computationally, another plus for the MORS is that it lends itself to parallel implementation. Indeed, all the trials of the MORS can be made simultaneously since the outcome of a trial is independent from the preceding ones. Thus, parallel processors can be used to perform these trials, which reduces the execution time of the search.

3.5.2. The Limitations

From the above exposition, it may seem that the MORS is an ideal method for synthesizing set theoretic estimates. Unfortunately, the MORS is limited to problems in which it is feasible to approximate the solution set $S$ by a region $G$ over which uniform generation of points can be performed efficiently and such that a good rejection rate is achievable. Without the latter restriction, the MORS would eventually generate a solution by Proposition 3.1, but it might require a prohibitive number of trials. As a general rule, it becomes increasingly difficult to achieve a low rejection rate as the dimension $k$ of the problem increases. To appreciate the extent of this problem, let us consider a basic
example. Suppose that the solution set is the diamond

$$ S = \left\{ x \in \mathbb{R}^k \mid \sum_{i=1}^{k} |x_i| \leq r \right\} \quad (3.23) $$

In addition, suppose that the search region which has been chosen to approximate $S$ is the circumscribed hypercube

$$ G = \left\{ x \in \mathbb{R}^k \mid \sup_{1 \leq i \leq k} \left| x_i \right| \leq r \right\} \quad (3.24) $$

Then

$$ \lambda(S \cap G) = \lambda(S) = \frac{(2r)^k}{k!} \quad \text{and} \quad \lambda(G) = (2r)^k \quad (3.25) $$

Consequently, the rejection rate is

$$ 1 - \gamma = 1 - \frac{1}{k!} \quad (3.26) $$

Thus, from (3.8), we obtain that the number of trials required to obtain a solution with probability $p$ is

$$ n = \frac{\ln(1 - p)}{\ln(1 - \frac{1}{k!})} \quad (3.27) $$

and, from (3.9), the average number of trials required to obtain a solution is

$$ \bar{\eta} = \frac{1}{\gamma} = k! \quad (3.28) $$

If we fix $p$ to 95%, then $n = 5$ and $\bar{\eta} = 2$ for $k = 2$, $n = 71$ and $\bar{\eta} = 24$ for $k = 4$, $n = 12 \times 10^4$ and $\bar{\eta} = 4 \times 10^4$ for $k = 8$, $n = 63 \times 10^{12}$ and $\bar{\eta} = 21 \times 10^{12}$ for $k = 16$. This explosive growth of the number of trials with respect to the number of parameters to be estimated renders the MORS inefficient as the dimension of this particular problem becomes large.
3.6. Alternative Search Strategies

In this section, we discuss the possibility of using alternative search strategies to improve the performance of the MORS.

3.6.1. Adaptive Searches

In the MORS, the search is purely random in that a given trial is not influenced by the outcome of any other trial. It is therefore natural to ask whether the search could be made adaptive and, thereby, more efficient. Unfortunately, the answer to that question seems to be negative. Indeed, in the presence of nonconvex property sets, the fact that a point belongs to the \( n-1 \)-st \( S_i \)'s \( (0 < n < m + 1) \) but not to the \( n \)-th does not, in general, provide much useful information as to the location or the proximity of the solution set. As a result, it is not clear how those trials which have failed to produce a feasible point could be used to construct a better guess for the next trial.

3.6.2. Nonuniform Searches

Given a search region \( G \), the MORS operates in an unbiased manner in that all the points belonging to \( G \) are regarded as equally likely solutions whereas points outside \( G \) are discarded. This principle gave rise to a uniform search procedure. In this section, we shall see that, under certain circumstances, the rejection rate can be improved if the original uniform distribution of the basic MORS is replaced by a nonuniform multivariate density. Let \( f \) be an arbitrary multivariate density supported by \( G \). The probability that a point drawn from \( G \) according to \( f \) is in \( S \) is given by (3.1) as

\[
\gamma_f = \frac{\int_{S} f d\lambda}{\int_{S \cap G} f d\lambda} = \frac{\int_{S} f d\lambda}{\int_{G} f d\lambda} \tag{3.29}
\]
In order for $f$ to be larger than its uniform counterpart of (3.4), $f$ should satisfy

$$
\int_{S \cap G} f d\lambda > \frac{\lambda(S \cap G)}{\lambda(G)}
$$

(3.30)

In words, this inequality indicates that $f$ should have more mass concentrated on $S$ than the uniform distribution, which is sensible. Practically speaking, in a given problem, sufficient information may be available to substantiate the assumption that the true solution is more likely to lie in some subset $G'$ of the search region $G$. Consequently, a nonuniform multivariate density $f$ should be used to intensify the search over $G'$. The higher the relative mass put over $G'$, the stronger is the user's belief that the true solution lies in $G'$. For instance, an approximate solution may be available, along with an error analysis for its maximum deviation from the true solution. This information could be used to determine $G'$ and construct $f$. Of course, in the absence of such information, there is no reason to bias the search and, therefore, a maximum uncertainty position should be adopted, i.e. $f$ should be the uniform density over $G$.

In principle, the only modification with respect to the basic algorithm of Section 3.4.2 is at step 1, where $x$ is no longer generated according to a uniform distribution over $G$. Techniques for the generation of points according to nonuniform multivariate distributions are discussed in [36]. These techniques, however, are computationally expensive and, in practice, it is usually sufficient to subdivide $G$ into $n$ disjoint hyperrectangles $\{G_1, \ldots, G_n\}$ of equal Lebesgue measure and to assign to each $G_i$ a weight $p_i$, where

$$\sum_{i=1}^{n} p_i = 1.$$ 

The sample points should then be generated according to the density

$$f = \sum_{i=1}^{n} p_i f_i$$

where $f_i$ is the uniform density over $G_i$. Thereby, those $G_i$'s with high weight $p_i$ will be probed more often on the average.
3.7. Random Search with Fuzzy Property Sets

It is recalled that a fuzzy set in an abstract space $\Xi$ is a class of objects with a continuum of grades of membership. Loosely speaking, it is a set with vague boundaries [114]. A fuzzy set $S$ in $\Xi$ is characterized by its membership function, i.e. a function $\mu$ which maps each point $x$ of $\Xi$ into a real number $\mu(x)$ in $[0,1]$ representing its grade of membership in $S$. In particular, the membership function of an ordinary (nonfuzzy) set $S$ is simply its indicator function $1_S$. For a full account of fuzzy set theory, see [39] and [114].

Fuzzy sets constitute a natural way to represent imprecise or partially defined pieces of a priori knowledge. Thus, they were used in [31] to treat the set theoretic digital signal restoration problem in the presence of inaccurate a priori information. The approach taken there is naturally generalized to arbitrary set theoretic estimation problems in $\mathbb{R}^k$ as follows. Let $\Gamma=\{S_1,\ldots,S_m\}$ be a collection of fuzzy subsets of $\mathbb{R}^k$ representing the a priori knowledge for an estimation problem and let $\{\mu_1,\ldots,\mu_m\}$ be the corresponding collection of membership functions. There are several ways to define the fuzzy solution set $S$ for $\Gamma$ as several intersection operators exist for fuzzy sets [39]. For our purposes, we choose to define $S$ through the membership function

$$\mu(x) = \prod_{i=1}^{m} \mu_i(x) \quad (\forall x \in \mathbb{R}^k)$$

Under the assumption that all the fuzzy sets have been modeled with continuous membership functions, this choice guarantees that $\mu$ is also continuous. A set theoretic estimate for the problem is a point which possesses the largest grade of membership in $S$, i.e. a point which maximizes $\mu$ (since $\mu$ is continuous, such a point exists if $S$ is compact). Thus, the search for a set theoretic estimate is a global optimization problem. In [31], several deterministic methods were mentioned to solve this problem. However, as seen in
Section 3.3, very efficient random search methods have been developed for the global optimization of continuous functionals on $\mathbb{R}^k$, which guarantee convergence to a global solution. These random search methods should prove worthwhile in the fuzzy set theoretic estimation problem.

3.8. The Applications of the MORS in Digital Signal Processing

The method of random search developed in this chapter is a simple and easily implementable alternative to the method of successive projections (MOSP) for generating set theoretic estimates in the Cartesian space. The MORS possesses features which, in many respects, make it more attractive than the MOSP. Nonetheless, its performance is low if the search region poorly approximates the solution set, a problem which was seen to become more acute as the dimension of the problem (i.e. the number of real parameters to estimate) increases. Under these considerations, the MORS is unconditionally recommended for low dimensional problems. Although we do not preclude its use for higher dimensions, the user should be warned of its potential inefficiency in such contexts. Before we comment on specific digital signal processing applications, let us make an important remark on the rôle of modelization in connection with the use of the MORS.

In general, the estimation of an abstract object involves the determination of a (sometimes infinite) number of real variables. On the other hand, the MORS is best fitted for problems in which the number of parameters is low. Hence, a necessary step towards the use of the MORS is modeling, i.e. the reduction of a high dimensional problem into one involving a small number of independent real parameters. Clever use of the a priori knowledge is central to efficient modeling. Admittedly, reducing the number of parameters needed to describe the object also requires approximations, which gives rise to some
degree of error. Nonetheless, such models are very valuable so long as the assumptions are coherent with the physical behavior of the system at hand. For sake of illustration, let us give a few examples in digital signal processing.

First of all, let us consider spectral estimation. Theoretically, estimating the spectral distribution of a stochastic process amounts to the determination of a function at an infinite number of points. However, if the original process is known to be autoregressive, then only the regression coefficients need be estimated. Generally speaking, for practical digital signal processing problems in which the data process can be described by an autoregressive model (e.g. processing of speech, seismic, radar, or sonar signals), excellent results are obtained with models whose order is limited to single digits and higher order models are known to bring little improvement. Another example is found in the problem of blur identification in digital image processing. A typical discrete blurring kernel is represented by a matrix with a large number of entries. Most elementary blurs can however be parametrized by very few coefficients [3]. For instance, a linear motion blur is totally determined by its length and its direction with respect to the image plane. The blur induced by an out-of-focus lens can be described by the value of the first zero crossing in the Fourier domain.

Needless to say, there are many problems which cannot be reduced to such low dimensionality and for which the use of the MORS should be avoided unless a low rejection rate can be guaranteed by geometric considerations. Such is the case of digital image recovery, where, typically, the number of pixels in the digitized image is \( k = 512 \times 512 \), and where further reduction of the number of independent parameters needed to faithfully describe this object is not justifiable on physical grounds.
3.9. Summary

A simple and easily implementable method of random search (MORS) to synthesize set theoretic estimates in the Cartesian space has been presented. It alleviates the theoretical and computational shortcomings of the MOSP and provides total flexibility with regard to the incorporation of the \textit{a priori} knowledge. Because of its potential low efficiency in high dimensional problems, the MORS is best fitted for applications in which the number of unknown parameters is low.

The MORS will be applied to the problem of harmonic retrieval in Chapter VI.
4.1. Introduction

The data provided by the observation of a physical system are always corrupted by additive noise. In most practical instances, some of the probabilistic attributes of the noise process are either known \textit{a priori} or measurable \textit{a posteriori} from the data. Such pieces of knowledge may include moments, absolute moments, second and higher order properties.

This chapter describes how probabilistic information pertaining to the noise process can be used in a general set theoretic estimation framework. In this approach, the sample statistics of the estimation residual are constrained to be consistent with those probabilistic properties of the noise which are available and sets are constructed accordingly in the solution space.

The idea of imposing noise-based constraints on the estimation residual was first formulated in a version of the constrained least-squares problem in [79] and was then applied to least-squares image restoration in [51]. There, the sample variance of the residual was forced to match that of the noise. This particular constraint has also been employed in other restoration techniques [3], [103]. In [104], this concept was reformulated in a set theoretic format and new residual-based constraints were introduced by considering other pieces of information (mean, spectral density) under the assumption that the noise was white and Gaussian.
We shall demonstrate that a larger amount of probabilistic knowledge relative to the noise process can be used to create sets in the solution space $\Xi$. The topological and geometric properties of the sets based on these probabilistic constraints are investigated and the benefit of their use toward the synthesis of better set theoretic estimates is discussed. The general framework in which the analysis takes place allows for a wider class of problems to be treated, well beyond the field of signal restoration. Moreover, the developments are not restricted to Gaussian white noise.

4.2. Methodology

Throughout this chapter, $\Xi$ has the structure of a topological vector space. All the random variables (r.v.'s) are defined on the same probability space $(\Omega, \Sigma, P)$. Lower case letters will be used to denote the value of a r.v. at a given elementary event $\omega$ in $\Omega$, representing a particular realization of the process. We recall (see Section B.3) that, for all $p$ in $\mathbb{R}_+^*$, $L^p(P)$ is the vector space of (classes of equivalence of) r.v.'s with finite $p$-th absolute moment. Following Loève [69], we shall define the probabilistic properties of a family of r.v.'s as those which can be expressed in terms of the joint distribution functions of its finite subfamilies. The Lebesgue measure in $\mathbb{R}$ will be denoted by $\lambda$. As to the measure and probability theoretic notions employed thereafter, the reader is referred to Appendix B.

As was seen in Chapter I, in most digital signal processing problems, the ultimate goal is to estimate an object $h$ from the data provided by observing a discrete stochastic process $\{X_n\}_{n \in \mathbb{Z}}$. A general mathematical model for the generation of these data is the discrete stochastic equation
In that model, $T_n$ is the signal formation operator and $\{U_n\}_{n \in \mathbb{Z}}$ is a discrete stochastic process which will be called noise and that represents any irrelevant information which happens to perturb the observations. Given an estimate $a$ for $h$, the estimation residual is defined as

$$
(\forall n \in \mathbb{Z}) \quad Y_n = X_n - T_n(a) \tag{4.2}
$$

In an ideal situation where the true object would be estimated with no error, i.e. $a = h$, one would get $T_n(a) = T_n(h)$, for every integer $n$. It follows easily from (4.1) and (4.2) that then

$$
(\forall n \in \mathbb{Z}) \quad Y_n = U_n \quad \text{a.s.} \tag{4.3}
$$

Thus, the residual and noise processes are equivalent and, as a result, they share the same probabilistic properties. Therefore, if $\Psi$ is a known probability theoretic property of the noise process, the estimate $a$ should lie in the subset $S$ of the solution space $\Xi$ defined by

$$
S_\Psi = \left\{ a \in \Xi \mid \{Y_n\}_{n \in \mathbb{Z}} \text{ satisfies } \Psi \right\} \tag{4.4}
$$

In practice, however, only $n$ samples $\{X_i \mid 1 \leq i \leq n\}$ of the data process are observed, yielding a finite sample path $\{x_i = X_i(\omega) \mid 1 \leq i \leq n\}$. Consequently, the residual process is traceable only through the finite sample path $\{y_1,...,y_n\}$ which can be computed from the available sample path $\{x_1,...,x_n\}$ in terms of $a$ by (4.2). Hence, it follows that the set which will actually be employed is

$$
S_\Psi = \left\{ a \in \Xi \mid \{y_1,...,y_n\} \text{ is consistent with } \Psi \right\} \tag{4.5}
$$
It is the set of all \( a \)'s which will produce a finite residual sample path consistent, within some confidence coefficient, with the property \( \Psi \).

Several relevant properties of these sets will be established, in particular closedness and convexity. These two properties are of great interest with regard to the synthesis of a set theoretic solution by projection methods according to the techniques discussed in Chapter II. We recall that if, for instance, \( \Xi \) is a finite dimensional normed vector space, then every nonempty closed set is proximinal by (vi) in Theorem 2.8. If \( \Xi \) is a uniformly convex Banach space, every nonempty closed and convex subset is a Chebyshev set on account of Theorem 2.9.

The operator \( T \) is defined as follows

\[
T: \Xi \rightarrow \mathbb{R}^n \\
a \mapsto \left( T_1(a), \ldots, T_n(a) \right)
\]  

(4.6)

4.3. Examples in Digital Signal Processing

Before we proceed with the analysis, we shall provide four examples of digital signal processing problems whose data generation model is that displayed in (4.1).

4.3.1. Digital Signal Restoration

In digital signal restoration, a popular model assumes that the degraded signal is an observation of a data process \( \{X_n\}_n \) obtained by convolving the original signal \( h=(h_1,\ldots,h_q) \) with some causal blurring kernel \( (t_1,\ldots,t_l) \) and by addition of noise. This noise is usually induced by the channel over which the signal is transmitted and the recording process. The \( n \)-th sample of the degraded signal is given by
Here, the goal is to restore the original signal, i.e. to estimate $h$.

### 4.3.2. Autoregressive Estimation

In a wide range of problems (e.g. system identification, speech prediction, spectral estimation, seismic, radar, or sonar data processing) the information signal $\{X_n\}_n$ is assumed to have been generated according to an autoregressive model of order $q$. The $n$-th data sample is given by

$$X_n = \sum_{k=1}^{q} h_k X_{n-k} + U_n$$

where $\{U_n\}_n$ is a random excitation sequence. The problem is then to identify the regression parameters, i.e. to estimate the $q$-tuple $h=(h_1,\ldots,h_q)$.

### 4.3.3. Processing of Radar Signals

A typical sample of a returned radar signal can be written as [106]

$$X_n = A \cos \left( \frac{2\pi(v + h_1)}{\tau} (n - h_2) + \phi \right) + U_n$$

where $A$ is the amplitude of the received signal, $\tau$ the sampling period of the receiver, $v$ the frequency of the transmitted signal, $h_1$ the Doppler shift induced by the motion of the target, $h_2$ some delay which is proportional to the distance to the target, and $\phi$ some phase reference. The noise process $\{U_n\}_n$ represents the interference. In such a context, one seeks to estimate the velocity and the range of the target, i.e. $h=(h_1,h_2)$. 
4.3.4. Harmonic Retrieval

The problem of estimating the frequency of sinusoidal signals in additive noise arises in many signal processing situations. A model for the $n$-th sample of the data process is

$$X_n = \sum_{k=1}^{q} b_k \sin(2\pi h_k \tau n + \theta_k) + U_n$$

where $q$ is the number of sinusoids, $b_k$ their amplitude, $\theta_k$ their phase, $\tau$ the sampling period, and $\{U_n\}$ the noise process. The object is then to estimate the unknown frequencies, i.e. the $q$-tuple $h = \{h_1, ..., h_q\}$.

4.4. Sets Based on Range Information

The purpose of this section is to construct sets by imposing bounds on the range of the values of the residual samples $\{y_i = z_i - T_i(a) \mid 1 \leq i \leq n\}$. It is assumed that all the r.v.'s in the noise process $\{U_n\}_{n \in \mathbb{Z}}$ are distributed as a nondegenerate r.v. $U$ with distribution function (d.f.) $F$.

Let us fix a confidence coefficient $1 - \epsilon$ in $[0,1]$. Then it is always possible to find two real numbers $\kappa$ and $\lambda$ such that

$$1 - \epsilon = P\left(\{\omega \in \Omega \mid \kappa \leq U(\omega) \leq \lambda\}\right)$$

(4.11)

Since $\{Y_n\}_{n \in \mathbb{Z}}$ and $\{U_n\}_{n \in \mathbb{Z}}$ are equivalent processes, each point in the finite sample path of the residual should also lie in the confidence interval $[\kappa,\lambda]$, which leads to the set

$$S_r = \bigcap_{i=1}^{n} C_i \quad \text{where} \quad C_i = \left\{ a \in \Xi \mid \kappa \leq z_i - T_i(a) \leq \lambda \right\}$$

(4.12)

The set $S_r$ can be regarded as the set of all $a$'s which produce a sample residual whose range is consistent, up to a $1 - \epsilon$ confidence coefficient, with that of the noise.
It is worth noting that if the range of $U$ is almost surely bounded, say $\kappa \leq U \leq \lambda$ a.s., $\epsilon$ can be taken to be zero and, hence, a 100% confidence coefficient is achievable. For example, this situation occurs when $U$ has a uniform or a binomial distribution. Another instance is when the range of $U$ takes a.s. the form $]-\infty, \lambda]$ or $[\kappa, +\infty[$. For example, in the latter case, $S_r$ becomes

$$S_r = \bigcap_{i=1}^{n} \left\{ a \in \mathbb{R} \mid \kappa \leq x_i - T_i(a) \right\}$$  \hspace{1cm} (4.13)$$

In particular, if $U$ has a Poisson or an exponential distribution, $\kappa = 0$ in the above.

Let us define

$$f_i : \Xi \to \mathbb{R}$$

$$a \mapsto x_i - T_i(a)$$ \hspace{1cm} (4.14)$$

Hence, (4.12) can be written as

$$S_r = \bigcap_{i=1}^{n} C_i \quad \text{where} \quad C_i = f_i^{-1}([\kappa, \lambda])$$ \hspace{1cm} (4.15)$$

**Proposition 4.1.**

$S_r$ is closed in $\Xi$ if $T$ is continuous.

**Proof.** All the $T_i$'s are continuous since $T$ is continuous. Hence, by (ii) in Proposition 2.1, each $C_i = f_i^{-1}([\kappa, \lambda])$ is closed in $\Xi$ as the inverse image of the closed subset $[\kappa, \lambda]$ of $\mathbb{R}$, and so is their intersection in (4.15). \quad \Box

**Proposition 4.2.**

$S_r$ is convex in $\Xi$ if $T$ is linear.
Proof. Let \( a \) and \( b \) be two arbitrary vectors in \( S_r \), let \( \alpha \) be an arbitrary real number in \( ]0,1[ \), and let \( i \) be an arbitrary integer in \( \{1,...,n\} \). Clearly, from (4.12), \( a \) and \( b \) both belong to \( C_i \) and thus

\[
\begin{cases}
\alpha \kappa \leq \alpha (x_i - T_i(a)) \leq \alpha \lambda \\
(1-\alpha)\kappa \leq (1-\alpha)(x_i - T_i(b)) \leq (1-\alpha)\lambda
\end{cases}
\]

(4.16)

But \( T_i \) is linear since \( T \) is linear. Thus, summing up the two expressions in (4.16) yields

\[
\kappa \leq x_i - T_i(\alpha a + (1-\alpha)b) \leq \lambda
\]

(4.17)

Hence, each \( C_i \) is convex and so is their intersection in (4.12) by (i) in Proposition 2.3, which concludes the proof. \( \square \)

It is noted that in cases where \( F \) is not available but where, for some \( p \) in \( \mathbb{R}^+ \), the \( p \)-th absolute moment of \( U \) is known, a probabilistic bound can be placed on \( |U| \) by invoking Markov's inequality \[69\]

\[
(\forall \kappa \in \mathbb{R}^+) \quad P \{ \omega \in \Omega \mid |U(\omega)| \geq \kappa \} \leq \kappa^{-p} E|U|^p
\]

(4.18)

4.5. Sets Based on Absolute Moment Information

4.5.1. Introduction

The use of absolute moment information relative to the noise process has been limited in the literature to the moment of order two \[3\], \[51\], \[103\], \[104\], leaving a broad class of potential \textit{a priori} knowledge unexploited. One reason for this may be that the second moment can be associated with the concrete notion of energy, which is a physically measurable quantity. There is, however, no reason to limit the analysis to order two moment information and it is of both theoretical and practical interest to investigate the use of absolute moments of arbitrary order in a general set theoretic framework.
4.5.2. Preliminaries

It is assumed that the noise sequence \( \{U_n\}_{n \in \mathbb{Z}} \) consists of independent and identically distributed (i.i.d.) r.v.'s, all distributed as a nondegenerate r.v. \( U \) with d.f. \( F \). Moreover, it is supposed that, for a fixed positive real number \( p \), the \( p \)-th absolute moment of \( U \), i.e.

\[
E|U|^p = \int_{\Omega} |U(\omega)|^p \, P(d\omega) = \int_{\mathbb{R}} |u|^p \, dF(u) \tag{4.19}
\]

is known along with \( E|U|^{2p} \) and that \( U \) belongs to \( L^{2p}(P) \). The latter restriction should not cause concern because the stochastic processes which arise from physical systems are a.s. uniformly bounded, which guarantees the finiteness of all the absolute moments. Consequently, the distributions which model these systems, even if they have unbounded support, are expected to admit finite absolute moments of any order (e.g. Gaussian, Poisson, etc.).

The \( n \)-tuples with \( i \)-th component \( x_i \) and \( y_i \) are denoted by \( x \) and \( y \), respectively. We define the functional \( N_p \) as follows

\[
(N_p(x) = \left( \sum_{i=1}^{n} |x_i|^p \right)^{1/p} \tag{4.20}
\]

For \( p \geq 1 \), \( N_p \) is a norm on \( \mathbb{R}^n \), as was seen in Section 2.4.1.

We shall also need the following theorem.

**Theorem 4.1.** [69]

Let \( U \) be a r.v. on \( (\Omega, \Sigma, P) \). Then, the function \( p \mapsto E^{1/p} |U|^p \) from \( \mathbb{R}_+^* \) into \([0, +\infty]\) is non-decreasing.

A consequence of this theorem is that \( L^q(P) \subset L^p(P) \), for all \( p \leq q \) in \( \mathbb{R}_+^* \).
4.5.3. Distribution of the Sample Absolute Moment

By hypothesis, the residual and the noise processes are equivalent and the $U_n$'s are i.i.d.r.v.'s with $p$-th absolute moment $E|U|^p$. Hence, the $Y_n$'s are also i.i.d.r.v.'s with $p$-th absolute moment $E|U|^p$. We recall that $\{y_1, \ldots, y_n\}$ is the segment of the sample path of the residual process which is available and from which the $p$-th absolute moment of the residual must be determined, within some confidence interval. The $p$-th sample absolute moment of the residual is defined as

$$M_p = \frac{1}{n} \sum_{i=1}^{n} |Y_i|^p$$  \hspace{1cm} (4.21)

Its expected value is given by

$$EM_p = \frac{1}{n} \sum_{i=1}^{n} E|Y_i|^p = \frac{1}{n} \sum_{i=1}^{n} E|U|^p = E|U|^p$$ \hspace{1cm} (4.22)

and its variance by

$$\sigma_p^2 = \frac{1}{n^2} \sum_{i=1}^{n} \text{Var}(Y_i)^p = \frac{1}{n^2} \sum_{i=1}^{n} \text{Var}|U|^p = \frac{\text{Var}|U|^p}{n} = \frac{E|U|^{2p} - E^2|U|^p}{n}$$ \hspace{1cm} (4.23)

It can be shown that, under the above hypotheses, as the sample size $n$ tends to infinity, the r.v. $M_p$ is asymptotically normal with mean $E|U|^p$ and variance $\sigma_p^2$ [44]. It will be assumed that $n$ is large enough so that the normal approximation is legitimate. Therefore, given a confidence coefficient $1-\epsilon$ in $]0,1[$, we can compute the constant $\alpha$ from the tables of the normal distribution so that

$$1-\epsilon = P\left(\{\omega \in \Omega \mid |M_p(\omega) - E|U|^p| \leq \alpha \sigma_p\}\right)$$ \hspace{1cm} (4.24)
4.5.4. Construction of the Absolute Moment Set

From (4.21), the value of the sample absolute moment at the elementary event \( \omega \) can be written as

\[
M_p(\omega) = \frac{1}{n} \sum_{i=1}^{n} |y_i|^p = \frac{1}{n} N_p(y) = \frac{1}{n} N_p(x - T(a))
\] (4.25)

Thus, by (4.24), the set of all \( a \)'s in \( \Xi \) which yield a sample absolute moment within the desired confidence interval is

\[
S_p = \left\{ a \in \Xi \mid \frac{1}{n} N_p(x - T(a)) - E|U|^p \leq \alpha \sigma_p \right\}
\] (4.26)

After some algebra, (4.26) yields

\[
S_p = \left\{ a \in \Xi \mid \eta_p \leq N_p(x - T(a)) \leq \xi_p \right\}
\] (4.27)

where

\[
\eta_p = \begin{cases} 
 n^{1/p} (E|U|^p - \alpha \sigma_p)^{1/p} & \text{if } E|U|^p \geq \alpha \sigma_p \\
 0 & \text{otherwise}
\end{cases}
\] (4.28)

and

\[
\xi_p = n^{1/p} (E|U|^p + \alpha \sigma_p)^{1/p}
\] (4.29)

The lower bound \( \eta_p \) is so defined because the functional \( N_p \) does not take on negative values. Let us define

\[
S_p^+ = \left\{ a \in \Xi \mid N_p(x - T(a)) \leq \xi_p \right\}
\]

\[
S_p^- = \left\{ a \in \Xi \mid N_p(x - T(a)) < \eta_p \right\}
\]

(4.30)

It is convenient to express \( S_p \) as the following difference
Thus, the construction of an absolute moment set requires that $U$ be in $L^{2p}(P)$ and that the $p$-th and $2p$-th absolute moments of $U$ be known.

4.5.5. Properties of the Absolute Moment Set

In this section and the following, whenever an absolute moment set $S_p$ is dealt with, it is understood that $U$ lies in $L^{2p}(P)$. We define the function $f_p$ by

$$f_p : \Xi \rightarrow \mathbb{R}$$

$$a \mapsto N_p(x - T(a))$$

Consequently we can write

$$S_p = f_p^{-1}([\eta_p, \xi_p]), \quad S_p^+ = f_p^{-1}([\infty, \xi_p]), \quad S_p^- = f_p^{-1}([-\infty, \eta_p])$$

Proposition 4.3.

$S_p$ and $S_p^+$ are closed in $\Xi$ and $S_p^-$ is open in $\Xi$ if $T$ is continuous.

Proof. Suppose $T$ is continuous. Then, since $N_p$ is continuous, so is $f_p$, as an elementary composition of continuous operators. It is noted that the sets $[\eta_p, \xi_p]$ and $[\infty, \xi_p]$ are closed in $\mathbb{R}$ and that the set $]-\infty, \eta_p[$ is open in $\mathbb{R}$. By (ii) in Proposition 2.1 and (4.33), the proof is complete. \(\square\)

Proposition 4.4.

$S_p^+$ and $S_p^-$ are convex in $\Xi$ if $T$ is linear and $p \geq 1$.

Proof. By (4.33) and (iii) in Proposition 2.3, it is enough to prove that $f_p$ is a convex function. Let $\alpha$ be an arbitrary real number in $]0,1[$ and let $a$ and $b$ be two arbitrary
vectors in $\Xi$. By linearity of $T$

$$f_p \left( \alpha a + (1-\alpha) b \right) = N_p \left( \alpha (x - T(a)) + (1-\alpha)(x - T(b)) \right)$$

(4.34)

But, for $p \geq 1$, $N_p$ is a norm on $\mathbb{R}^n$ and it is therefore convex. Whence

$$N_p \left( \alpha (x - T(a)) + (1-\alpha)(x - T(b)) \right) \leq \alpha N_p \left( x - T(a) \right) + (1-\alpha) N_p \left( x - T(b) \right)$$

(4.35)

Consequently

$$f_p \left( \alpha a + (1-\alpha) b \right) \leq \alpha f_p \left( a \right) + (1-\alpha) f_p \left( b \right)$$

(4.36)

which is the desired result. $\square$

**Proposition 4.5.**

Suppose that for some positive real number $p$

$$E |U|^2^p \geq \left( 1 + \frac{n}{\alpha^2} \right) E^2 |U|^p$$

(4.37)

Then, for all real numbers $q$ greater than or equal to $p$, $S_q = S_q^+$. 

**Proof.** Let $q$ be any real number such that $q \geq p$. Then it can be shown that [109]

$$\frac{E |U|^{2q}}{E^2 |U|^q} \geq \frac{E |U|^{2p}}{E^2 |U|^p}$$

(4.38)

Hence, if (4.37) is satisfied for $p$, we obtain that

$$\frac{E |U|^{2q}}{E^2 |U|^q} \geq 1 + \frac{n}{\alpha^2}$$

(4.39)

Which in turn yields

$$E |U|^q \leq \alpha \sqrt{E |U|^{2q} - E^2 |U|^q} = \alpha \sigma_q$$

(4.40)

It follows from (4.28) that $\eta_q = 0$ and, whence, $S_q = S_q^+$. $\square$
4.5.6. Combining Absolute Moment Information

In many practical instances enough a priori information is available to construct more than one absolute moment set. For example, the r.v. $U$ could be normally distributed with mean zero and standard deviation $\sigma$. Then

$$(\forall p \in \mathbb{R}_+^*) \quad E|U|^p = \frac{2^{p/2}}{\pi^{1/2}} \Gamma\left(\frac{p+1}{2}\right) \sigma^p$$

(4.41)

Hence, from the knowledge of any absolute moment, one can compute the parameter $\sigma$ and infer the value of any other absolute moment. This property would also hold if $U$ had a Rayleigh, Maxwell, or exponential distribution or any distribution whose absolute moments can be expressed unambiguously in terms of a single parameter. The motivation for using several absolute moment sets is that as the number of property sets increases their intersection gets smaller, which yields a better defined solution. The two basic issues that arise in connection with the use of two or more absolute moment sets are consistency and redundancy. To illustrate this point, consider two absolute moment sets $S_p$ and $S_q$. In order for them to carry consistent information, they should possess a nonempty intersection. On the other hand, suppose that $S_p$ is a subset of $S_q$. Then $S_q$ would be superfluous as a property set since it would not contribute to a smaller solution set.

Let $p$ and $q$ be two positive real numbers. In our analysis, we shall proceed as follows. First, we shall establish those conditions on the coefficients $\eta_p$, $\xi_p$, $\eta_q$, and $\xi_q$ that would ensure that the sets $S_p$ and $S_q$ would possess a nonempty intersection (Proposition 4.6) and, further, those conditions under which $S_p$ would either be a subset or a superset of $S_q$ (Proposition 4.7). We shall then discuss whether the coefficients defined in (4.28) and (4.29) satisfy these conditions.
Lemma 4.1.

Let \( p \) and \( q \) be two positive real numbers such that \( p \leq q \). Then

\[
(\forall x \in \mathbb{R}^n) \quad N_q(x) \leq N_p(x) \leq n^{1/p - 1/q} N_q(x) \tag{4.42}
\]

**Proof.** Let \( x = (x_1, \ldots, x_n) \) be an arbitrary point in \( \mathbb{R}^n \) and let \( 0 < p \leq q < +\infty \) be two real numbers. We first show that

\[
(\forall y \in \mathbb{R}^n) \quad N_p(y) = 1 \implies N_q(y) \leq 1 \tag{4.43}
\]

Let \( y \) be an arbitrary point in \( \mathbb{R}^n \) such that \( N_p(y) = 1 \). Then

\[
(\forall i \in \{1, \ldots, n\}) \quad |y_i| \leq 1 \quad \text{and} \quad |y_i|^q \leq |y_i|^p \tag{4.44}
\]

Consequently

\[
\sum_{i=1}^{n} |y_i|^q \leq \sum_{i=1}^{n} |y_i|^p = 1 \tag{4.45}
\]

which establishes (4.43). If \( N_p(x) = 0 \), the left-hand-side inequality is trivial. If not, let \( y = x/N_p(x) \). Then \( N_p(y) = 1 \) and, by (4.43), we get

\[
\frac{1}{N_p(x)} N_q(x) = N_q \left( \frac{1}{N_p(x)} x \right) = N_q(y) \leq 1 \tag{4.46}
\]

It follows that

\[
N_q(x) \leq N_p(x) \tag{4.47}
\]

To prove the right-hand side inequality, we write Hölder's inequality in \( \mathbb{R}^n \) (by applying (B.12) with the integral of (B.13))

\[
(\forall t \in [1, +\infty])(\forall (y, z) \in \mathbb{R}^n \times \mathbb{R}^n) \quad \sum_{i=1}^{n} |y_i z_i| \leq \left( \sum_{i=1}^{n} |y_i|^t \right)^{t-1} \left( \sum_{i=1}^{n} |z_i|^t \right)^{1/t} \tag{4.48}
\]
In particular, for \( t = q/p \), \( y = (1, \ldots, 1) \), and \( z = (x_1^p, \ldots, x_n^p) \), we get

\[
\sum_{i=1}^{n} |x_i|^p \leq n^{\frac{q-1}{q}} \left( \sum_{i=1}^{n} |x_i|^q \right)^{p/q}
\]

(4.49)

Whence

\[
N_p(z) \leq n^{1/p - 1/q} N_q(z)
\]

(4.50)

which is the desired result. \( \Box \)

For sake of convenience, we define

\[
(\forall t \in \mathbb{R}^*_+) \quad A_t = \left\{ x \in \mathbb{R}^n \mid \eta_t \leq N_t(x) \leq \xi_t \right\}
\]

(4.51)

Proposition 4.6.

Let \( p \) and \( q \) be two positive real numbers such that \( p \leq q \) and suppose that the range of \( a \mapsto x - T(a) \) contains \( A_p \cap A_q \). Then \( S_p \cap S_q \neq \emptyset \) if and only if the two following conditions hold

(i) \( \eta_q \leq \xi_p \).

(ii) \( \eta_p \leq n^{1/p - 1/q} \xi_q \).

Proof. Let \( p \leq q \) be any two numbers in \( \mathbb{R}^*_+ \). The range of \( a \mapsto x - T(a) \) contains \( A_p \cap A_q \). Hence, if \( A_p \cap A_q \) is not empty neither is \( S_p \cap S_q \). Therefore, it is enough to show that \( A_p \cap A_q \neq \emptyset \) if and only if (i) and (ii) hold. We first prove that (i) and (ii) are necessary. The sets \( A_p \) and \( A_q \) can be regarded as concentric full toroids in \( \mathbb{R}^n \). Let \( y \) be an arbitrary point in \( \mathbb{R}^n \). Then, by Lemma 4.1

\[
N_q(y) \leq N_p(y) \leq n^{1/p - 1/q} N_q(y)
\]

(4.52)
Hence, if we further assume that \( y \) is in \( A_q \), (4.52) leads to

\[
\eta_q \leq N_p(y) \leq n^{1/p-1/q} \xi_q \tag{4.53}
\]

Thus, if \( \eta_q > \xi_p \), (4.53) implies that \( N_p(y) > \xi_p \) and hence \( y \notin A_p \). Thus, the intersection of the toroids will be empty (geometrically, \( A_q \) will be exterior to \( A_p \)). Likewise, if \( \eta_p > n^{1/p-1/q} \xi_q \), then (4.53) implies that \( N_p(y) < \eta_p \). Hence \( y \notin A_p \). Thus, the intersection of the toroids will be empty (geometrically, \( A_q \) will be interior to \( A_p \)). This proves necessity. To prove sufficiency, consider the limit cases. First, let \( \eta_q = \xi_p \). Then (ii) clearly holds since \( \eta_p \leq \xi_p = \eta_q \leq n^{1/p-1/q} \xi_q \). Moreover, the toroids are then tangent in the \( 2n \) points \( y \) whose components are all zero except for one which is \( \pm \eta_q \). Indeed, for such a \( y \), it is easily verified that \( N_p(y) = N_q(y) = \eta_q = \xi_p \). Second, let \( \eta_p = n^{1/p-1/q} \xi_q \). Then (i) clearly holds since \( \eta_q \leq \xi_q = n^{1/p-1/q} \xi_q = \eta_p \leq \xi_p \). Moreover, the two toroids are then tangent in the \( 2^n \) points of the form \( y = (\pm \eta_p/n^{1/p}, \ldots, \pm \eta_p/n^{1/p}) \). Indeed, for such a \( y \), it is easily verified that \( N_q(y) = n^{1/q-1/p} \eta_p = \xi_q \) and \( N_p(y) = \eta_p \). In between these two extreme cases, i.e. when both (i) and (ii) hold, the toroids \( A_p \) and \( A_q \) overlap in an infinite number of points. \( \square \)

**Proposition 4.7.**

Let \( R \) denote the range of \( a \mapsto x - T(a) \) and let \( Q \) be any of the \( 2^n \) quadrants of \( \mathbb{R}^n \).

Then, for every positive real numbers \( p \) and \( q \) such that \( p \leq q \), we have

(i) If \( A_p \cap Q \subset R \), \( S_p \subset S_q \) \( \iff \) \( n^{1/p-1/q} \eta_q \leq \eta_p \) and \( \xi_p \leq \xi_q \).

(ii) If \( A_q \cap Q \subset R \), \( S_q \subset S_p \) \( \iff \) \( \eta_p \leq \eta_q \) and \( \xi_q \leq n^{1/q-1/p} \xi_p \).

**Proof.** (i) Suppose that \( n^{1/p-1/q} \eta_q > \eta_p \) and let \( y \) be the point in \( Q \) of the form \( y = (\pm \eta_p/n^{1/p}, \ldots, \pm \eta_p/n^{1/p}) \). Then \( y \in A_p \) as \( N_p(y) = \eta_p \). But, since \( A_p \cap Q \subset R \), there exists an \( a \) in \( S_p \) such that \( y = x - T(a) \). However, this \( a \) does not lie in \( S_q \) for
\[ N_q(y) = n^{1/q - 1/p} \eta_p < \eta_q. \] Hence, for \( S_p \) to be a subset of \( S_q \), it is necessary that \( n^{1/p - 1/q} \eta_q \leq \eta_p \). To show the necessity of the second condition, suppose that \( \zeta_p > \zeta_q \) and let \( y \) be any of the points in \( Q \) whose components are all zero except for one which is \( \pm \zeta_p \). Then \( y \in A_p \) since \( N_p(y) = \zeta_p \). But, since \( A_p \cap Q \subset R \), there exists an \( a \) in \( S_p \) such that \( y = z - T(a) \). However, this \( a \) is not in \( S_q \) for \( N_q(y) = \zeta_p > \zeta_q \). To prove sufficiency, let \( a \) be an arbitrary point in \( S_p \) and let \( y = z - T(a) \). Then \( \eta_p \leq N_p(y) \leq \zeta_p \). But, from Lemma 4.1

\[ n^{1/q - 1/p} N_p(y) \leq N_q(y) \leq N_p(y) \]  

(4.54)

Hence, it follows that

\[ n^{1/q - 1/p} \eta_p \leq N_q(y) \leq \zeta_p \]  

(4.55)

But, by hypothesis, \( \eta_q \leq n^{1/q - 1/p} \eta_p \) and \( \zeta_p \leq \zeta_q \). Hence, (4.55) yields

\[ \eta_q \leq N_q(y) \leq \zeta_q \]  

(4.56)

Thus, \( a \in S_q \), which establishes (i). (ii) Suppose that \( \eta_p > \eta_q \) and let \( y \) be any of the points in \( Q \) whose components are all zero except for one which is \( \pm \eta_q \). Then \( y \in A_q \) since \( N_q(y) = \eta_q \). But, since \( A_q \cap Q \subset R \), there exists an \( a \) in \( S_q \) such that \( y = z - T(a) \). However, this \( a \) is not in \( S_p \) for \( N_p(y) = \eta_q < \eta_p \). Thus, for \( S_q \) to be a subset of \( S_p \), it is necessary that \( \eta_p \leq \eta_q \). To show the necessity of the second condition, suppose that \( \zeta_q > n^{1/q - 1/p} \zeta_p \) and let \( y \) be the point in \( Q \) of the form \( y = (\pm \zeta_q/n^{1/q}, \ldots, \pm \zeta_q/n^{1/q}) \). Then \( y \in A_q \) as \( N_q(y) = \zeta_q \). But, since \( A_q \cap Q \subset R \), there exists an \( a \) in \( S_q \) such that \( y = z - T(a) \). However, this \( a \) is not in \( S_p \) for \( N_p(y) = n^{1/p - 1/q} \zeta_q > \zeta_p \). To prove sufficiency, let \( a \) be an arbitrary point in \( S_q \) and let \( y = z - T(a) \). Then \( \eta_q \leq N_q(y) \leq \zeta_q \). Hence, by Lemma 4.1

\[ \eta_q \leq N_p(y) \leq n^{1/p - 1/q} \zeta_q \]  

(4.57)
But, by hypothesis, $\eta_p \leq \eta_q$ and $n^{1/p - 1/q} \xi_q \leq \xi_p$. Hence, (4.57) entails

$$\eta_p \leq N_p(y) \leq \xi_p \quad (4.58)$$

Therefore $a \in S_p$. □

4.5.7. Discussion

The conditions on the range $R$ of $a \mapsto x - T(a)$ are, for example, trivially satisfied if the operator $T$ is surjective (onto). Let $\eta_p$ and $\xi_q$ be defined as in (4.28) and (4.29) respectively. We now prove that (ii) in Proposition 4.6 is always satisfied.

**Proposition 4.8.**

Let $p$ and $q$ be two positive real numbers such that $p \leq q$. Then

$$\eta_p \leq n^{1/p - 1/q} \xi_q.$$  

**Proof.** The case $\eta_p = 0$ is trivial. If $\eta_p > 0$, we have

$$n^{-1/p} \eta_p = \left( E|U|^p - \alpha \sqrt{\frac{E|U|^{2p} - E^2|U|^p}{n}} \right)^{1/p} \leq E^{1/p} |U|^p \quad (4.59)$$

On the other hand

$$E^{1/q} |U|^q \leq \left( E|U|^q + \alpha \sqrt{\frac{E|U|^{2q} - E^2|U|^q}{n}} \right)^{1/q} = n^{-1/q} \xi_q \quad (4.60)$$

But, by Theorem 4.1, $E^{1/p} |U|^p \leq E^{1/q} |U|^q$ since $p \leq q$. Hence

$$\eta_p \leq n^{1/p - 1/q} \xi_q \quad (4.61)$$

which is the desired result. □

We now claim that $\eta_p$ and $\xi_p$ are decreasing functions of $p$ for reasonable values of the sample size $n$ (say $n \geq 4$). Although a formal proof for this claim is not available, it can be
supported by the following analysis. First of all, let us rewrite (4.28) and (4.29) as

\[ \eta_p = n^{1/p} \times \left( 1 - \frac{\alpha}{\sqrt{n}} \sqrt{\frac{E|U|^{2p}}{E^{2}|U|^p} - 1} \right)^{1/p} \times E^{1/p} |U|^p \]  \tag{4.62}

and

\[ \zeta_p = n^{1/p} \times \left( 1 + \frac{\alpha}{\sqrt{n}} \sqrt{\frac{E|U|^{2p}}{E^{2}|U|^p} - 1} \right)^{1/p} \times E^{1/p} |U|^p \]  \tag{4.63}

In these products, as \( p \) increases, the first factor clearly decreases whereas the last factor increases by Theorem 4.1. As to the variations of the second factor, no conclusion seems possible. Numerical computations for various distributions have indicated that, for \( n \geq 4 \), \( n^{1/p} \) decreases faster than the other factors in (4.62) and (4.63) increase. Hence, \( \eta_p \) and \( \zeta_p \) are decreasing functions of \( p \). Thus, (i) in Proposition 4.6 is always satisfied since \( \eta_q < \eta_p \leq \zeta_p \) and the inequalities \( \zeta_p \leq \zeta_q \) and \( \eta_p \leq \eta_q \) never hold in Proposition 4.7.

Consequently, under proper assumptions on the range of \( a \mapsto x - T(a) \), two different absolute moment sets intersect without one being a subset or a superset of the other. The practical significance of this result is that different absolute moment sets carry consistent but not redundant information. Hence, by adding as many moment sets as the \textit{a priori} knowledge permits to the existing collection of property sets used in the description of the solution, the feasible set gets smaller by (1.3), which will yield better estimates.

4.6. Sets Based on Moment Information

4.6.1. Introduction

It is assumed that the noise sequence \( \{U_n\}_{n \in \mathbb{Z}} \) consists of i.i.d.r.v.'s, all distributed as a nondegenerate r.v. \( U \) with d.f. \( F \). It is also supposed that, for a fixed positive integer \( k \), \( U \)
belongs to $L^{2k}(P)$, and that the $k$-th moment of $U$, i.e.

$$EU^k = \int \Omega U^k(\omega)P(d\omega) = \int R u^k dF(u)$$

(4.64)

is known along with $EU^{2k}$. The former condition guarantees that $EU^k$ exists and is finite and that the variance of $U^k$ is finite since

$$|EU^k| \leq EU|U|^k \leq E^{1/2}|U|^{2k} < +\infty$$

(4.65)

For our purposes, moments are not as valuable as absolute moments. First of all, moments are defined only at positive integral order, which limits the range of their properties. Second, in many instances, moments and absolute moments coincide and, therefore, the results of Section 4.5 apply. Such is the case when $U$ is a.s. nonnegative (e.g. $U$ has a Poisson distribution or is absolutely continuous with an exponential, Maxwell or Rayleigh density), or when the moment under consideration is of even order. Hence, only odd moments of noise processes for which the r.v. $U$ assumes negative values on a set of positive $P$-measure need be considered thereafter.

4.6.2. Construction of the Moment Set

The procedure followed here is basically the same as in Section 4.5. For convenience, the same notations will be used but it is understood that the positive integral index $k$ refers to a moment (as the positive real index $p$ referred to an absolute moment).

The $k$-th sample moment based on the portion $\{Y_1,\ldots,Y_n\}$ of the residual process is defined as

$$M_k = \frac{1}{n} \sum_{i=1}^{n} Y_i^k$$

(4.66)

By repeating the analysis of Section 4.5.3, it follows that $M_k$ is asymptotically normal
with mean and standard deviation respectively given by

\[ EM_k = EU^k \quad \text{and} \quad \sigma_k = \sqrt{\frac{EU^{2k} - E^2U^k}{n}} \]  

(4.67)

Hence, assuming that the normal approximation is legitimate and given a confidence coefficient $1 - \epsilon$ in $]0,1[$, one can compute the constant $\alpha$ such that

\[ 1 - \epsilon = P\left(\{\omega \in \Omega \mid |M_k(\omega) - EU^k| \leq \alpha \sigma_k\}\right) \]  

(4.68)

Consequently, the set of all $a$'s in $\Xi$ which yield a $k$-th sample moment within the desired confidence interval is

\[ S_k = \left\{ a \in \Xi \mid \gamma_k \leq \sum_{i=1}^{n} \left( x_i - T_i(a) \right)^k \leq \delta_k \right\} \]  

(4.69)

where

\[
\begin{align*}
\gamma_k &= n(\mathbb{E}U^k - \alpha \sigma_k) \\
\delta_k &= n(\mathbb{E}U^k + \alpha \sigma_k)
\end{align*}
\]  

(4.70)

4.6.3. Properties of the Moment Set

Whenever a moment set $S_k$ is dealt with, it is understood that $U$ is in $L^{2k}(P)$. If $k$ is even or if $U$ is a.s. nonnegative, the results of Sections 4.5.5 and 4.5.6 apply. We define

\[ f_k: \Xi \rightarrow \mathbb{R} \]

\[ a \mapsto \sum_{i=1}^{n} \left( x_i - T_i(a) \right)^k \]  

(4.71)

Thus, from (4.69), $S_k$ can be rewritten as

\[ S_k = f_k^{-1}(\gamma_k, \delta_k) \]  

(4.72)
Proposition 4.9.

$S_k$ is closed in $\Xi$ if $T$ is continuous.

**Proof.** If $T$ is continuous, so is $f_k$. By (ii) in Proposition 2.1, it follows from (4.72) that $S_k$ is closed in $\Xi$ as the inverse image of the closed subset $[\gamma_k, \delta_k]$ of $\mathbb{R}$. $\Box$

Proposition 4.10.

$S_1$ is convex in $\Xi$ if $T$ is linear.

**Proof.** Let $\alpha$ be an arbitrary real number in $]0,1[$ and let $a$ and $b$ be two arbitrary vectors in $S_1$. Then

\[
\begin{align*}
\alpha \gamma_1 &\leq \alpha \sum_{i=1}^{n} \left( z_i - T_i(a) \right) \leq \alpha \delta_1 \\
(1-\alpha) \gamma_1 &\leq (1-\alpha) \sum_{i=1}^{n} \left( z_i - T_i(b) \right) \leq (1-\alpha) \delta_1
\end{align*}
\]

By summing up these two expressions and using the linearity of $T$, it follows at once that

\[
\gamma_1 \leq \sum_{i=1}^{n} \left( z_i - T_i(\alpha a + (1-\alpha)b) \right) \leq \delta_1
\]

Whence, $\alpha a + (1-\alpha)b$ also belongs to $S_1$ which is therefore convex. $\Box$

Proposition 4.11.

Suppose that $U$ is such that

\[
(\forall c \in \mathbb{R}) \quad F(c) + F(-c) = 1 - P\left( \{\omega \in \Omega \mid U(\omega) = c \} \right)
\]

Then, if $k$ is odd, $S_k$ can be written as

\[
S_k = \left\{ a \in \Xi \mid \left| \sum_{i=1}^{n} \left( z_i - T_i(a) \right)^k \right| \leq \alpha \sqrt{nE_{2k}} \right\}
\]
Proof. The odd moments of a r.v. $U$ satisfying (4.75) are zero [44]. Therefore (4.69) reduces to (4.76). □

The conditions of Proposition 4.11 are satisfied, in particular, when the r.v. $U$ is absolutely continuous with a $\lambda$-a.e. even density. Such is the case when the noise is zero mean and uniform, Laplacian, or Gaussian, which are common assumptions in signal processing applications.

4.7. Sets Based on Second Order Information

4.7.1. Introduction

In the previous sections, we have focused on the first order properties of the noise process $\{U_n\}_{n \in \mathbb{Z}}$. We shall now turn our attention to second order properties, i.e. properties which can be defined or determined by means of the mixed second moments $E U_n U_m$. From Section B.2 (see also [12] and [70] for an extensive discussion), it is known that $L^2(P)$ is a real Hilbert space, with scalar product

$$<U, V> = \sum_{\Omega} UVdP = EUV$$

It follows from the Cauchy-Schwarz inequality that if the $U_n$'s are in $L^2(P)$, their mixed second moments exist and are finite. Two r.v.'s $U$ and $V$ in $L^2(P)$ are uncorrelated if $EUV = EUVE$; they are orthogonal if $EUV = 0$. A central concept in the forthcoming developments is that of the spectral distribution of a stationary process and a discussion of this topic is in order.

4.7.2. Spectral Analysis of Stationary Discrete Stochastic Processes

For a full account of the spectral properties of complex stochastic processes the reader is
referred to Blanc-Lapierre and Picinbono [12], Doob [38], Loève [70], and Rosenblatt [85].

Since we are interested only in real processes, we shall follow mainly Doob's treatment, which provides great insight into such processes.

4.7.2.1. The Correlation Function

Let \( \{U_n\}_{n \in \mathbb{Z}} \) be a (real) wide sense stationary discrete stochastic process (w.s.s.d.s.p.), i.e., a process for which all the \( U_n \)'s are in \( L^2(P) \), the \( EU_n \)'s do not depend on \( n \), and such that the correlation function \( r \) does not depend on \( n \), where

\[
  r: \mathbb{Z} \to \mathbb{R} \\
  m \mapsto EU_n U_n + m 
\]  

(4.78)

Thus, by centering each r.v. of a w.s.s.d.s.p. at its expectation, a zero mean w.s.s.d.s.p. is obtained. Therefore, only zero mean processes need be considered thereafter.

**Theorem 4.2.** [38]

Let \( r \) be the correlation function of a w.s.s.d.s.p. Then

(i) \( (\forall m \in \mathbb{Z}) \ r(m) = r(-m). \)

(ii) \( (\forall m \in \mathbb{Z}) \ |r(m)| \leq r(0). \)

(iii) For every finite collection of real numbers \( \{b_1, \ldots, b_n\} \), \( \sum_{i=1}^{n} \sum_{j=1}^{n} b_i r(i-j) b_j \geq 0. \)

4.7.2.2. The Spectral Distribution and the Spectral Density

The notions of spectral distribution and spectral density will be introduced through the following theorems. We recall (see Section B.1) that a d.f. on \( \mathbb{R} \) is a real-valued, nondecreasing, left continuous function.
Theorem 4.3. [38]

A function $r: \mathbb{Z} \rightarrow \mathbb{R}$ is the correlation function of a w.s.s.d.s.p. if and only if there exists a nondecreasing real-valued function $G$ defined on $[0,1/2]$ such that

$$(\forall m \in \mathbb{Z}) \quad r(m) = \int_0^{1/2} \cos(2\pi \nu m) dG(\nu) \quad (4.79)$$

The function $G$ in the above Fourier-Stieltjes integral is called the spectral distribution of the process in question. It actually is a d.f. on $[0,1/2]$ since it can be adjusted so that, for every $\nu$ in $]0,1/2]$, $G(\nu) = G(\nu^-)$, hence making it left continuous. The total power of the process is given by

$$r(0) = \int_0^{1/2} dG(\nu) \quad (4.80)$$

As a d.f., $G$ is defined up to an additive constant. It is convenient to normalize it so that $G(0) = 0$ and $G(1/2) = r(0)$. The spectrum of a w.s.s.d.s.p. consists of every point $\nu$ in $[0,1/2]$ in whose neighborhood the spectral distribution increases. Hence, if $\bar{G}$ denotes the function which coincides with $G$ on $[0,1/2]$ and is zero elsewhere, the spectrum is given by

$$\sigma = \left\{ \nu \in [0,1/2] \mid (\forall \epsilon \in \mathbb{R}^+) \quad \bar{G}(\nu + \epsilon) > \bar{G}(\nu - \epsilon) \right\} \quad (4.81)$$

The following theorem gives the general decomposition of a d.f. on $[0,1/2]$. It applies in particular to the spectral distribution of a w.s.s.d.s.p.

Theorem 4.4. [69]

A d.f. $G$ on $[0,1/2]$ can be decomposed as

$$G = G_1 + G_2 + G_3 \quad (4.82)$$

where
(i) $G_1$ is absolutely continuous: there exists a nonnegative measurable function $g$ on $[0,1/2]$ such that

$$(\forall \nu \in [0,1/2]) \quad G_1(\nu) = \int_{[0,\nu]} gd\lambda \quad \text{and} \quad dG_1 = gd\nu$ \quad \lambda\text{-a.e.} \quad (4.83)$$

(ii) $G_2$ is a step function with at most a countable number of jumps. Namely, if $v_i$ is the location of the $i$-th jump and $\kappa_i$ its height, then

$$(\forall \nu \in [0,1/2]) \quad G_2(\nu) = \sum_{v_i < \nu} \kappa_i \quad (4.84)$$

(iii) $G_3$ is continuous with zero derivative $\lambda$-a.e.

The function $g$ in (i) is called the spectral density of the process. The following theorem gives a sufficient condition under which a w.s.s.d.s.p. is absolutely continuous.

**Theorem 4.5.** [38]

Let $r$ be the correlation function of a w.s.s.d.s.p. $\{U_n\}_{n \in \mathbb{Z}}$. Then the spectral distribution $G$ of $\{U_n\}_{n \in \mathbb{Z}}$ is absolutely continuous if

$$\sum_{m = -\infty}^{+\infty} |r(m)| < +\infty \quad (4.85)$$

The spectral density of $\{U_n\}_{n \in \mathbb{Z}}$ is then given by

$$(\forall \nu \in [0,1/2]) \quad g(\nu) = 2r(0) + 4 \sum_{m = 1} r(m) \cos(2\pi \nu m) \quad (4.86)$$

and its correlation function by

$$(\forall m \in \mathbb{Z}) \quad r(m) = \int_0^{1/2} \cos(2\pi \nu m) g(\nu) d\nu \quad (4.87)$$

Loosely speaking, Theorem 4.5 states that if the correlation function damps out rapidly as $|m|$ increases, the spectral distribution reduces to its absolutely continuous component.
(i.e. \( G = G_1 \) in Theorem 4.4), meaning that the process has a spectral density.

### 4.7.2.3. Particular Cases

#### 4.7.2.3.1. White Noise Processes

A discrete white noise (or purely random) process with total power \( \sigma^2 \) is a collection \( \{U_n\}_{n \in \mathbb{Z}} \) of pairwise uncorrelated zero mean (and therefore pairwise orthogonal) r.v.'s with, for all integer \( n \), \( E|U_n|^2 = \sigma^2 \). This constitutes a w.s.s.d.s.p. with an absolutely continuous spectral distribution whose correlation function and spectral density are respectively given by

\[
(\forall m \in \mathbb{Z}) \quad r(m) = \begin{cases} 
\sigma^2 & \text{if } m = 0 \\
0 & \text{else}
\end{cases} \quad \text{and} \quad (\forall \nu \in [0,1/2]) \quad g(\nu) = 2\sigma^2 \quad (4.88)
\]

Hence, the spectral density of such a process is constant.

#### 4.7.2.3.2. Processes with a Spectral Density

The class of w.s.s.d.s.p.'s which possess a spectral density (i.e. \( G = G_1 \) in Theorem 4.4) includes stationary autoregressive processes, stationary moving average processes, and stationary autoregressive moving average processes [81]. A completely general model for a w.s.s.d.s.p. \( \{U_n\}_{n \in \mathbb{Z}} \) whose spectral distribution is absolutely continuous is

\[
(\forall n \in \mathbb{Z}) \quad U_n = \sum_{i= -\infty}^{+\infty} b_i V_{n-i} \quad \text{with} \quad \sum_{i= -\infty}^{+\infty} b_i^2 < +\infty \quad (4.89)
\]

where \( \{V_n\}_{n \in \mathbb{Z}} \) is a discrete white noise process. The spectral density of \( \{U_n\}_{n \in \mathbb{Z}} \) is then given by [81]

\[
(\forall \nu \in [0,1/2]) \quad g(\nu) = 2\sigma^2 \left| \sum_{i= -\infty}^{+\infty} b_i \exp(-j2\pi \nu i) \right|^2 \quad (4.90)
\]
4.7.2.3.3. Harmonic Processes

Harmonic processes constitute an important example of processes whose spectral distribution consists only of jumps (i.e. $G = G_2$ in Theorem 4.4). An expression for such a process is [81]

$$(\forall n \in \mathbb{Z}) \quad U_n = \sum_{k=1}^{q} b_k \sin(2\pi \nu_k n + \theta_k) \quad \text{with} \quad 0 \leq \nu_1, \ldots, \nu_q \leq 1/2$$ (4.91)

where the $\theta_k$'s are uniformly distributed r.v.'s over $]-\pi, \pi[$. Its correlation function is

$$(\forall m \in \mathbb{Z}) \quad r(m) = \frac{1}{2} \sum_{k=1}^{q} b_k^2 \cos(2\pi \nu_k m)$$ (4.92)

Moreover, its spectral distribution $G$ consists of a step function with $q$ jumps, the $k$-th jump being located at $\nu_k$ and having height $b_k^2/2$. Therefore, the spectrum of this process is $\sigma = \{\nu_1, \ldots, \nu_q\}$. A common abuse of language is to say that such a process has a line spectrum because, if $G$ was to be regarded as a generalized function (i.e. a distribution in the sense of Schwartz [88]), its derivative would be a sum of $q$ Dirac distributions with respective masses $\{b_1^2/2, \ldots, b_q^2/2\}$ concentrated at frequencies $\{\nu_1, \ldots, \nu_q\}$.

4.7.3. The Case of Gaussian White Noise

We recall that a stochastic process $\{U_n\}_{n \in \mathbb{Z}}$ is said to be Gaussian if the d.f. of every finite subfamily $\{U_{n_1}, \ldots, U_{n_k}\}$ is multivariate Gaussian. In this section, it is assumed that $\{U_n\}_{n \in \mathbb{Z}}$ is a Gaussian white noise process of total power $\sigma^2$, as defined in Section 4.7.2.3.1. It is well known that two uncorrelated jointly Gaussian r.v.'s are independent. Hence, the $U_n$'s are Gaussian i.i.d.r.v.'s with mean zero and variance $\sigma^2$.

\footnote{The distributions of Laurent Schwartz have nothing to do with d.f.'s.}
4.7.3.1. Construction of the Spectral Set

The residual process \( \{Y_n\}_{n \in \mathbb{Z}} \) is equivalent to the noise process \( \{U_n\}_{n \in \mathbb{Z}} \) and, therefore, it should also be white, with constant spectral density \( 2\sigma^2 \). In the following, the chi-square distribution with \( p \) degrees of freedom is denoted by \( \chi_p^2 \).

**Theorem 4.6.** [81]

Let \( \{Y_n\}_{n \in \mathbb{Z}} \) be a Gaussian discrete white noise process with total power \( \sigma^2 \) and let \( n \) be an even number. Define

\[
(\forall k \in \{0, \ldots, n/2\}) \quad I_k = \frac{2}{n} \left| \sum_{i=1}^{n} Y_i \exp(-j\frac{2\pi}{n} ki) \right|^2
\]

(4.93)

Then

(i) \( \{I_0, \ldots, I_{n/2}\} \) are independent r.v.'s.

(ii) \( I_0/2\sigma^2 \) and \( I_{n/2}/2\sigma^2 \) have a \( \chi_1^2 \) distribution.

(iii) \( \{I_1/\sigma^2, \ldots, I_{n/2-1}/\sigma^2\} \) all have a \( \chi_2^2 \) distribution.

The \( I_k \)'s constitute the discrete periodogram of the residual process, based on the \( n \) r.v.'s \( \{Y_1, \ldots, Y_n\} \). It follows from the above theorem that

\[
(\forall k \in \{0, \ldots, n/2\}) \quad EI_k = 2\sigma^2 \quad \text{and} \quad \text{Var} I_k = \begin{cases} 
8\sigma^4 & \text{if } k = 0 \text{ or } n/2 \\
4\sigma^4 & \text{if } 0 < k < n/2
\end{cases}
\]

(4.94)

As in the previous sections, \( \{Y_1, \ldots, Y_n\} \) represents the portion of the residual process which has been observed. For convenience, we assume that \( n \) is even (if it is odd, \( n/2 \) should be replaced by \( (n-1)/2 \) in Theorem 4.6 and in the following equations). Although the periodogram is known to be a poor estimate of the spectral density [81], Theorem 4.6 provides a simple means to constrain the finite sample path \( \{y_1, \ldots, y_n\} \) of the residual to be
consistent with the property that \( \{Y_n\}_{n \in \mathbb{Z}} \) is a Gaussian white noise process. Indeed, we can compute from the tables of the \( \chi_1^2 \) distribution and from the expression of the \( \chi_2^2 \) distribution (which is simply an exponential distribution with parameter 1/2) the constants \( \beta_1 \) and \( \beta_2 \) such that, for some \( \epsilon \) in \( ]0,1[ \)

\[
1 - \epsilon = \begin{cases} 
P(\{\omega \in \Omega \mid \frac{I_k(\omega)}{2\sigma^2} \leq \beta_1\}) & \text{if } k = 0 \text{ or } n/2 \\
P(\{\omega \in \Omega \mid \frac{I_k(\omega)}{\sigma^2} \leq \beta_2\}) & \text{if } 0 < k < n/2
\end{cases}
\]  

(4.95)

Let \( \omega \) be the elementary event corresponding to the observed realization of the process. Since the finite residual sample path is given by \( \{y_i = x_i - T_i(a) \mid 1 \leq i \leq n\} \), the observed values of the discrete periodogram are given by

\[
(\forall k \in \{0, \ldots, n/2\}) \quad I_k(\omega) = \frac{2}{n} \left| \sum_{i=1}^{n} (x_i - T_i(a)) \exp(-j\frac{2\pi}{n} ki) \right|^2
\]  

(4.96)

It follows that the set of all \( a \)'s which produce a finite residual path consistent, within a \( 1-\epsilon \) confidence coefficient, with the whiteness and normality of the noise process is

\[
S_d = \bigcap_{k=0}^{n/2} D_k
\]  

(4.97)

where

\[
(\forall k \in \{0, \ldots, n/2\}) \quad D_k = \left\{ a \in \Xi \mid \left| \sum_{i=1}^{n} (x_i - T_i(a)) \exp(-j\frac{2\pi}{n} ki) \right|^2 \leq \xi_k \right\}
\]  

(4.98)

with

\[
(\forall k \in \{0, \ldots, n/2\}) \quad \xi_k = \begin{cases} 
\frac{n\sigma^2}{2} \beta_1 & \text{if } k = 0 \text{ or } n/2 \\
\frac{n\sigma^2}{2} \beta_2 & \text{if } 0 < k < n/2
\end{cases}
\]  

(4.99)
As was mentioned earlier, a $\chi^2_2$ distribution is simply an exponential distribution with parameter $1/2$. Hence, given $\epsilon$, we get easily that $\beta_2 = -2\ln(\epsilon)$. Consequently

\[(\forall k \in \{1, \ldots, n/2-1\}) \quad \xi_k = -n \sigma^2 \ln(\epsilon) \] 

(4.100)

4.7.3.2. Properties of the Spectral Set

For every $k$ in $\{0, \ldots, n/2\}$, let us define

\[f_k: \Xi \rightarrow \mathbb{R} \]

\[a \mapsto \left| \sum_{i=1}^{n} (\tau_i - T_i(a)) \exp(-j \frac{2\pi}{n} k i) \right|^2 \]

(4.101)

Then

\[(\forall k \in \{0, \ldots, n/2\}) \quad D_k = f_k^{-1}([-\infty, \xi_k]) \] 

(4.102)

**Proposition 4.12.**

$S_d$ is closed in $\Xi$ if $T$ is continuous.

**Proof.** Let $k$ be an arbitrary integer in $\{0, \ldots, n/2\}$. If $T$ is continuous so are the $T_i$'s. It is easily seen that $f_k$ will thereby be continuous, as the composition and the sum of continuous functions. Therefore, by (4.102), since $[-\infty, \xi_k]$ is closed in $\mathbb{R}$, each $D_k$ is closed in $\Xi$ by (ii) in Proposition 2.1, and so is their intersection $S_d$ by (i) in Proposition 2.1. □

**Proposition 4.13.**

$S_d$ is convex in $\Xi$ if $T$ is linear.

**Proof.** Let $k$ be an arbitrary integer in $\{0, \ldots, n/2\}$. It is enough to show that $f_k$ is convex. Indeed, if it is, then $D_k$ will be convex by (4.102) and (iii) in Proposition 2.3 and so will the intersection $S_d$ by (i) in Proposition 2.3. Let $\alpha$ be an arbitrary real number in $]0,1[$
and let $a$ and $b$ be two arbitrary elements in $\Xi$. Then, by hypothesis, for each $i$ in $\{1, \ldots, n\}$, $T_i$ is linear and thus

$$x_i - T_i \left( a + (1-\alpha)b \right) = \alpha \left( x_i - T_i(a) \right) + (1-\alpha) \left( x_i - T_i(b) \right)$$

(4.103)

Consequently, by multiplying through by $w_{ik} = \exp(-j\frac{2\pi n}{n} ki)$, summing over all $i$'s, and taking the square of the magnitude, we get

$$f_k \left( a + (1-\alpha)b \right) = \left| \alpha \sum_{i=1}^{n} (x_i - T_i(a))w_{ik} + (1-\alpha) \sum_{i=1}^{n} (x_i - T_i(b))w_{ik} \right|^2$$

(4.104)

But it is well known that $z \mapsto |z|^2$ is a convex function on $\mathbb{C}$. Upon applying the corresponding convexity inequality to the right hand side, we get

$$f_k \left( a + (1-\alpha)b \right) \leq \alpha f_k \left( a \right) + (1-\alpha)f_k \left( b \right)$$

(4.105)

Thus, $f_k$ is convex. $\square$

4.7.4. The Case of Non-Gaussian White Noise

If $\{U_n\}_{n \in \mathbb{Z}}$ is a non-Gaussian discrete white noise process, it is no longer guaranteed that the $U_n$'s are independent, which tremendously limits the analysis. In this section, we shall place ourselves in a more restricted situation and assume that $\{U_n\}_{n \in \mathbb{Z}}$ is a discrete process with i.i.d.r.v.'s all distributed as a zero mean r.v. $U$ in $L^4(P)$, with variance $\sigma^2$. Such a process is white with total power $\sigma^2$ and constant spectral density $2\sigma^2$.

Let $\{I_0, \ldots, I_{n/2}\}$ be the discrete periodogram of the residual process based on the r.v.'s $\{Y_1, \ldots, Y_n\}$, as defined in the previous section. It can be shown that the $I_k$'s are asymptotically (i.e. as $n$ goes to infinity) pairwise uncorrelated with $[81]$
\[
\text{Var} I_k = \begin{cases} 
8\sigma^4 + \frac{4(E|U|^4 - 3\sigma^4)}{n} & \text{if } k = 0 \text{ or } n/2 \\
4\sigma^4 + \frac{4(E|U|^4 - 3\sigma^4)}{n} + O\left(\frac{1}{n^2}\right) & \text{if } 0 < k < n/2
\end{cases}
(4.106)
\]

Moreover, by invoking the Central Limit Theorem, it can be shown that the r.v.'s in (ii) and (iii) of Theorem 4.6 are asymptotically distributed as a \(\chi_1^2\) and a \(\chi_2^2\) respectively [52]. These results indicate that, under relatively mild conditions, the conclusions of Theorem 4.6 hold in an asymptotic sense. Consequently, provided that \(n\) is large enough, the distributions of the r.v.'s in the discrete periodogram can be approximated by their asymptotic limits and the results of Section 4.7.3 apply.

4.7.5. The General Case of Correlated Noise

In this section, we shall drop the whiteness assumption. If \(\{U_n\}_{n \in \mathbb{Z}}\) is a zero mean strictly stationary discrete process whose span of dependence is small enough (formally speaking, it is required that the process be strongly mixing with absolutely summable second and fourth order cumulants), then Theorem 4.6 holds asymptotically [85]. Second order properties can also be enforced in instances when, for some lags \(m\), the values of the correlation function \(r(m)\) of the noise process are known. Since the residual process is equivalent to the noise process, these values should also be that of the correlation function of the residual process at these lags and the finite sample path \(\{y_1, ..., y_n\}\) should be consistent with them, up to some confidence coefficient. Unfortunately, the distribution theory for the sample correlations is extremely complicated. Under relatively involved assumptions, it can be shown that the r.v.'s defined by

\[
(\forall m \in \{0, ..., n - 1\}) \quad R_m = \frac{1}{n} \sum_{i=1}^{n-m} Y_i Y_{i+m}
(4.107)
\]
are asymptotically jointly Gaussian [85]. Exact results have also been established for low order correlation lags of special processes, which are referenced in [81].

In practice, all the assumptions mentioned so far may be very difficult to justify. For an approximate result based on more workable assumptions, we now follow [81]. Let \( \{U_n\}_{n \in \mathbb{Z}} \) be a zero mean w.s.s.d.s.p. Let us define the normalized correlation function as
\[
\tilde{r}(\cdot) = \frac{r(\cdot)}{r(0)}
\]
and the normalized sample correlation function as
\[
\tilde{R}_m = \frac{\sum_{i=1}^{n-|m|} Y_i Y_{i+|m|}}{\sum_{i=1}^{n} |Y_i|^2}
\]
(4.108)

Now suppose that \( |\tilde{r}(m)| \) goes to zero as \( |m| \) goes to infinity and that all the normalized correlation coefficients are known. Then a crude estimate for the asymptotic distribution of \( \tilde{R}_m \) is a normal distribution with mean \( \tilde{r}(m) \) and variance
\[
\sigma_r^2 = \frac{1}{n} \sum_{m=-\infty}^{+\infty} |\tilde{r}(m)|^2
\]
(4.109)

A more complex expression for \( \sigma_r^2 \) can be found in [81] for cases where \( |\tilde{r}(m)| \) does not go to zero as \( |m| \) goes to infinity. If only one of the true normalized correlation coefficients is known, say for a given lag \( m \) in \( \{-n+1, \ldots, n-1\} \), then the observed realization of the sample correlation r.v. defined in (4.108) can be used in lieu of \( \tilde{r}(m) \) in the above equation, leading to
\[
\sigma_r^2 = \frac{1}{n} \sum_{m=1-n}^{n-1} \left| \sum_{i=1}^{n-|m|} y_i y_{i+|m|} \right|^2 \frac{1}{\sum_{i=1}^{n} |y_i|^2}
\]
(4.110)

From the tables of the normal distribution, we can compute the constant \( \alpha \) so that a
given confidence coefficient

\[ 1 - \epsilon = P \left( \{ \omega \in \Omega \mid | \bar{R}_m(\omega) - \bar{r}(m) | \leq \alpha \sigma_r \} \right) \]

(4.111)

is achieved. It follows that the set of all \( a \)'s which give a residual sample path within the corresponding confidence interval is

\[
S_c = \left\{ a \in \Xi \mid \left\{ \frac{1}{n} \sum_{i=1}^{n} \left( x_i - T_i(a) \right) \left( x_{i+1} - T_{i+1}(a) \right) - \bar{r}(m) \right\} \leq \alpha \sigma_r \right\}
\]

(4.112)

As in the proof of Proposition 4.12, it can be shown that \( S_c \) is closed if \( T \) is continuous.

4.8. Sets Based on Higher Order Information

4.8.1. Introduction

If a process \( \{ U_n \}_{n \in \mathbb{Z}} \) is Gaussian, all the information about \( \{ U_n \}_{n \in \mathbb{Z}} \) is contained in the means \( EU_n \) and the second mixed moments \( EU_n U_{n+r} \). First and second order probabilistic attributes, however, do not provide a complete description of non-Gaussian processes. It is therefore of interest to investigate the use of the information available from higher order properties, i.e. properties defined or determined via mixed moments of order greater than two.

4.8.2. Higher Order Properties of Discrete Stochastic Processes

4.8.2.1. Preliminaries

Historically, the idea of higher order spectral (polyspectral) densities was suggested by Kolmogorov and was first introduced in [93]. A general discussion of their properties can
be found in [20]. The asymptotic theory of polyspectral estimates is discussed in [21] and their computational aspects in [22]. In the forthcoming brief exposition of higher order properties, we shall mainly follow Rosenblatt's treatment [85]. The reader is referred to [20] and [85] for a more complete account. It will also be useful to regard some of the results of Section 4.7.2 as particular cases of those below.

Let $k$ be a positive integer. We shall call a $k$-th order stationary discrete stochastic process (k.s.d.s.p.) a process $\{U_n\}_{n \in \mathbb{Z}}$ which is stationary up to order $k$ and for which all the $U_n$'s are in $L^k(P)$. It follows that for every $q$ in $\{1, \ldots, k\}$ and for every subfamily $\{U_{n_1}, \ldots, U_{n_q}\}$ of such a process, the $q$-th order mixed moment

$$m(n_1, \ldots, n_q) = E \prod_{i=1}^{q} U_{n_i},$$

exists, is finite, and depends only on the lags $\{n_2-n_1, \ldots, n_q-n_1\}$.

4.8.2.2. Cumulants

Let $\{U_n\}_{n \in \mathbb{Z}}$ be a k.s.d.s.p. and let $(n_1, \ldots, n_k)$ be a $k$-tuple in $\mathbb{Z}^k$. The $k$-th order cumulant is defined as

$$c(n_1, \ldots, n_k) = \sum_{\kappa} (-1)^{p-1}(p-1)! \prod_{j=1}^{p} m(\kappa_j) \quad \text{with} \quad m(\kappa_j) = E \prod_{i \in \kappa_j} U_{n_i},$$

where the summation extends over all partitions $\kappa = \{\kappa_1, \ldots, \kappa_p\}$ of $\{n_1, \ldots, n_k\}$. Given a collection of integers $\{n_1, \ldots, n_k\}$, the mixed moment can be expressed in terms of the cumulants as follows

$$m(n_1, \ldots, n_k) = \sum_{\kappa} \prod_{j=1}^{p} c(\kappa_j)$$

where $\kappa$ is as in (4.114). Because we deal with processes which are stationary up to order
$k$, we shall from now onwards simply express the cumulant $c(n_1,\ldots,n_k)$ as a function $c(\tau_1,\ldots,\tau_{k-1})$ of the lags $\{\tau_i = n_{i+1} - n_1 \mid 1 \leq i \leq k-1\}$. If all the $U_n$'s have mean zero, the cumulant function of order two is simply the correlation function

$$(\forall r \in \mathbb{Z}) \quad c(r) = EU_n U_{n+r} \quad (4.116)$$

the cumulant of order three is the third order mixed moment

$$(\forall (r,s) \in \mathbb{Z}^2) \quad c(r,s) = EU_n U_{n+r} U_{n+s} \quad (4.117)$$

and the cumulant of order four is given by

$$(\forall (r,s,t) \in \mathbb{Z}^3) \quad c(r,s,t) = EU_n U_{n+r} U_{n+s} U_{n+t} - EU_n U_{n+r} EU_{n+s} U_{n+t}$$
$$- EU_n U_{n+s} EU_{n+r} U_{n+t} - EU_n U_{n+t} EU_{n+r} U_{n+s} \quad (4.118)$$

The cumulants are symmetric functions of their arguments. For instance, the third order cumulant satisfies

$$(\forall (r,s) \in \mathbb{Z}^2) \quad c(r,s) = c(s,r) = c(-r,s-r) = c(r-s,-s) \quad (4.119)$$

Similar relations apply to higher order cumulants.

4.8.2.3. Polyspectral Densities

Let $\{U_n\}_{n \in \mathbb{Z}}$ be a k.s.d.s.p. such that

$$\sum_{\tau_1 = -\infty}^{+\infty} \cdots \sum_{\tau_{k-1} = -\infty}^{+\infty} |c(\tau_1,\ldots,\tau_{k-1})| < +\infty \quad (4.120)$$

The $k$-th order spectral density of $\{U_n\}_{n \in \mathbb{Z}}$ is defined for every $(k-1)$-tuple $(\nu_1,\ldots,\nu_{k-1})$ in $[-1/2,1/2]^{k-1}$ as

$$g(\nu_1,\ldots,\nu_{k-1}) = \sum_{\tau_1 = -\infty}^{+\infty} \cdots \sum_{\tau_{k-1} = -\infty}^{+\infty} c(\tau_1,\ldots,\tau_{k-1}) \exp \left( -j2\pi \sum_{i=1}^{k-1} \nu_i \tau_i \right) \quad (4.121)$$

In particular, for $k=3$, we obtain the bispectral density
\[(\forall (\nu_1, \nu_2) \in [-1/2, 1/2]^2) \quad g(\nu_1, \nu_2) = \sum_{r = -\infty}^{+\infty} \sum_{s = -\infty}^{+\infty} c(r, s) \exp \left( -j2\pi(\nu_1 r + \nu_2 s) \right) \quad (4.122)\]

From the properties displayed in (4.119), the following symmetry properties of the bispectral density are obtained

\[(\forall (\nu_1, \nu_2) \in [-1/2, 1/2]^2) \quad g(\nu_1, \nu_2) = g(\nu_2, \nu_1) = g(-\nu_1, -\nu_2) = g(-\nu_1 - \nu_2, \nu_2) = g(\nu_1, -\nu_1 - \nu_2) \quad (4.123)\]

These relations imply that the values of the bispectral density are completely specified by the values defined over triangles such as

\[\Delta = \left\{ (\nu_1, \nu_2) \in [0, 1/2]^2 \mid \nu_1 \geq \nu_2 \quad \text{and} \quad 2\nu_1 + \nu_2 \leq 1 \right\} \quad (4.124)\]

Similar relations hold for polyspectral densities of order greater than three, hence limiting the domain over which the polyspectral density needs be specified.

### 4.8.2.4. Polyspectral Density Estimates

Let \(\{Y_n\}_{n \in \mathbb{Z}}\) be a strictly stationary process such that for every integer \(k\) greater than two

\[(\forall i \in \{1, \ldots, k - 1\}) \quad \sum_{\tau_i = -\infty}^{+\infty} \cdots \sum_{\tau_{k-1} = -\infty}^{+\infty} (1 + |\tau_i|) \left| c(\tau_1, \ldots, \tau_{k-1}) \right| < +\infty \quad (4.125)\]

The \(k\)-periodogram based on the segment \(\{Y_1, \ldots, Y_n\}\) of the process is defined as

\[I_{\nu_1, \ldots, \nu_k} = \frac{1}{n} \prod_{i=1}^{k} \sum_{j=1}^{n} Y_i \exp(-j2\pi \nu_i i) \quad \text{for} \quad \sum_{i=1}^{k} \nu_i = 0 \quad \text{(modulo 1)} \quad (4.126)\]

Under the provision that \((\nu_1, \ldots, \nu_k)\) does not lie in a proper subspace, the \(k\)-periodogram is an asymptotically unbiased estimate of the \(k\)-th order spectral density as \(n\) goes to infinity. It is however inconsistent and must be smoothed by averaging weighted versions. Under proper assumptions on the weighting function and the process, these estimates can
be shown to be asymptotically normal as $n$ goes to infinity.

4.8.3. Sets Constructed from Higher Order Statistics

In this section, we briefly show how to construct sets based on properties accessible through higher order statistics. It is assumed that the noise process $\{U_n\}_{n \in \mathbb{Z}}$ satisfies the assumptions of Section 4.8.2.4.

4.8.3.1. The Normality Set

Suppose that $\{U_n\}_{n \in \mathbb{Z}}$ is a Gaussian process. Then, all its cumulants of order greater than two are zero and it follows that all its spectral densities of higher order than the second vanish. Thus, for every integer $k$ greater than two, the observed value of the $k$-periodogram based on the sample path $\{y_1, \ldots, y_n\}$ of the residual process should be within some interval around zero. Given a confidence coefficient, the bounds of this interval can be computed by invoking the asymptotic normal distribution of the $k$-periodogram. The corresponding set is that of all $a$'s which yield a residual whose observed $k$-periodogram falls in the confidence interval.

4.8.3.2. The Linearity Set

In this section, we follow [100]. Suppose that $\{U_n\}_{n \in \mathbb{Z}}$ is a linear process, i.e.

$$(\forall n \in \mathbb{Z}) \quad U_n = \sum_{i=-\infty}^{+\infty} b_i V_{n-i} \quad \text{with} \quad \sum_{i=-\infty}^{+\infty} b_i^2 < +\infty$$

(4.127)

where the $V_n$'s are zero mean i.i.d.r.v.'s in $L^3(P)$ with $EV_n^2 = \mu_2$ and $EV_n^3 = \mu_3$. Then, if $g(.)$ and $g(.,.)$ denote respectively the spectral and the bispectral density of $\{U_n\}_{n \in \mathbb{Z}}$, it can be shown that for every frequency $\nu_1$ and $\nu_2$
\[ z(v_1, v_2) = \frac{|g(v_1, v_2)|^2}{g(v_1)g(v_2)g(v_1 + v_2)} = \frac{\mu_3^2}{\mu_2} \quad (4.128) \]

Let \( \hat{z}(v_1, v_2) \) be the estimate of \( z(v_1, v_2) \) computed by replacing \( g(v) \) and \( g(v_1, v_2) \) by the observed values of the respective periodograms based on the residual sample path \( \{y_1, \ldots, y_n\} \). Then, the constancy of \( \hat{z}(v_1, v_2) \) over some grid \( G \) in the frequency plane can be used to test the linearity of the residual process. This test was developed in [100] and can be used to construct the set of all \( a \)'s which produce a residual such that \( \hat{z}(v_1, v_2) \) is within some confidence interval around the constant value \( \mu_3^2/\mu_2 \) for all \( (v_1, v_2) \) in \( G \).

4.8.3.3. The Independence Set

Suppose that \( \{U_n\}_{n \in \mathbb{Z}} \) consists of i.i.d.r.v.'s distributed as a r.v. \( U \) whose cumulants at all order exist and are finite, the \( k \)-th being given by (4.114) for \( n_1 = \ldots = n_k \) and denoted by \( c_k \). The cumulant function of this process is given by

\[
(\forall (n_1, \ldots, n_k) \in \mathbb{Z}^k) \quad c(n_1, \ldots, n_k) = \begin{cases} 
  c_k & \text{if } n_1 = \ldots = n_k = 0 \\
  0 & \text{otherwise} 
\end{cases} \quad (4.129)
\]

Hence, the real part of the \( k \)-th order spectral density is the constant \( c_k \) and the imaginary part is zero [22]. It is noted that the flatness of the spectral density merely translates the uncorrelatedness of the \( U_n \)'s (see Section 4.7.2.3.1). On the other hand, the independence of the \( U_n \)'s shows up in the flatness of the higher order spectral densities. For a given \( k \), the corresponding set is that of all \( a \)'s which produce a residual whose observed \( k \)-periodogram falls within the confidence interval around the expected constant value. Again, given a confidence coefficient, the bounds of this interval can be computed by invoking the asymptotic normal distribution of the \( k \)-periodogram.
4.8.3.4. The Reversibility Set

Suppose that \( \{U_n\}_{n \in \mathbb{Z}} \) is time reversible, meaning that the finite dimensional d.f.'s of \( \{U_{-n}\}_{n \in \mathbb{Z}} \) are the same as that of \( \{U_n\}_{n \in \mathbb{Z}} \). As noted in [22], we then have

\[
(\forall (n_1, \ldots, n_k) \in \mathbb{Z}^k) \quad c(n_1, \ldots, n_k) = c(-n_1, \ldots, -n_k) \tag{4.130}
\]

and the imaginary part of the \( k \)-th order spectral density is therefore identically zero. For a given \( k \), the corresponding set is that of all \( a \)'s which produce a residual whose observed \( k \)-periodogram has an imaginary part in some confidence interval (determined as above) around zero.

4.9. Sets Based on Other Probabilistic Information

Suppose that all the r.v.'s in the noise process \( \{U_n\}_{n \in \mathbb{Z}} \) are distributed as a r.v. \( U \) with d.f. \( F \). In statistics, procedures exist to test various hypotheses concerning \( F \). Testing an hypothesis \( \Psi \) typically consists in choosing a nonnegative statistic \( Z \) whose d.f. (under the assumption that \( \Psi \) is true) is available and setting a significance level \( \epsilon \). The observed value of the statistic is then computed from a finite sample path of the process and, if it is less than some constant \( \alpha \) determined by \( \epsilon \) and the d.f. of \( Z \), the hypothesis is accepted.

These tests can be used to construct sets in the solution space. In this context, an hypothesis is a given piece of a priori knowledge \( \Psi \) pertaining to \( F \). As before, \( \{Y_i = X_i - T_i(a) \mid 1 \leq i \leq n\} \) is the finite segment of the residual process which is observed. Let \( Z(Y_1, \ldots, Y_n) \) be the statistic of the test associated with \( \Psi \) and let \( \alpha \) be the constant determined by some significance level \( \epsilon \). Then, since the residual process should possess the property \( \Psi \), a set can be constructed by constraining its available sample path \( \{y_i = x_i - T_i(a) \mid 1 \leq i \leq n\} \) to yield an observed statistic \( Z(x - T(a)) \) less than \( \alpha \), namely
\[ S_\Psi = \left\{ a \in \Xi \mid Z(x - T(a)) < \alpha \right\} \] (4.131)

Hence, \( S_\Psi \) is the set of all \( a \)'s which give rise to a residual which is consistent, within a \( 1 - \epsilon \) confidence coefficient, with the known property \( \Psi \).

The piece of information \( \Psi \) may be relative to the functional form of \( F \). In particular, techniques are available to test departure from a uniform [88], normal [77], and exponential [92] distribution. In other instances, it may only be known that \( F \) is symmetric (e.g. about zero as in (4.75)). The symmetry test of [33] can then be used.

4.10. Summary

Property sets based on the knowledge of various pieces of probabilistic information relative to the noise process have been constructed and analyzed. Among the pieces of information considered were the range of the r.v.'s, the moments and absolute moments of arbitrary order, the correlation function at some lags, and the spectral density. It was also shown how sets based on information available through higher order cumulants and polyspectral densities could be constructed. The possibility of constructing sets via tests based on various other types of probabilistic knowledge such as the functional form of the d.f. of the noise or its symmetry properties was also indicated. The use of these property sets in the framework of set theoretic digital signal restoration and set theoretic harmonic retrieval will be demonstrated in the next two chapters.
CHAPTER V

APPLICATION TO DIGITAL SIGNAL RESTORATION

5.1. Introduction

The digital restoration problem is to estimate the original form of a blurred and noise corrupted discrete signal. This chapter is concerned with the application of set theoretic estimation to this problem. In this context, the original signal is the object which must be recovered from a degraded version and some a priori knowledge. Simulation results are provided to illustrate an application of the method of successive projections (MOSP) developed in Chapter II. Some of the sets derived in Chapter IV are also used. The novelty of these simulations resides in the incorporation of a nonconvex property set in the description of the solution.

5.2. Assumptions and Models

We are concerned with problems in which the original signal has been blurred by a linear shift-invariant operator and by addition of wide-sense stationary signal-independent noise. Most instances of signal degradation can be well approximated by this model. Under the above assumptions, the blurring operation can be described by convolution. To conform to previous notations, the data process for this problem is denoted by \( \{X_n\}_{n \in \mathbb{Z}} \), the kernel of the blur operator by \( \{t_n\}_{n \in \mathbb{Z}} \), and the noise process by \( \{U_n\}_{n \in \mathbb{Z}} \). Moreover, the object to be estimated is the original signal \( h = \{h_n\}_{n \in \mathbb{Z}} \). Hence, the model for the formation of the
data process is

$$(\forall n \in \mathbb{Z}) \quad X_n = \sum_{i=-\infty}^{+\infty} t_i h_{n-i} + U_n$$

(5.1)

In practice, the original signal and the blurring kernel have finite extent, say $n - l + 1$ and $l$ respectively. Thus, $n$ will represent the length of the degraded signal and, for convenience, $l-1$ zeroes will be added at the end of the original signal so that its length is also $n$. Therefore, the degraded signal (i.e. the observed realization of the segment \{X_1, \ldots, X_n\} of the data process) can be written as

$$x = Th + u$$

(5.2)

where $x$ is the $n \times 1$ degraded signal vector, $T$ the $n \times n$ blur matrix, $h$ the $n \times 1$ original signal vector, and $u$ the $n \times 1$ noise vector.

5.3. Conventional Restoration Methods

Conventional restoration methods are based on classical estimation theory, in particular on Bayesian techniques. They seek to produce an estimate of the original signal which is optimum in terms of a predefined criterion. Least-squares (Wiener filtering), constrained least-squares, maximum a posteriori, and maximum entropy are among the most common restoration criteria [3]. Other criteria have also been considered in the literature, such as maximum power [102]. In Section 1.2, we expressed some reservations towards the use of classical estimation methods because they rely on an arbitrary notion of optimality and they often require biased statistical assumptions about the object to be estimated. Moreover, they usually make poor use of the a priori knowledge.
5.4. Set Theoretic Restoration

In set theoretic digital signal restoration, the \textit{a priori} knowledge is represented by a collection \(\{S_1, \ldots, S_m\}\) of property sets and the solution space \(\Xi\) is simply the Cartesian space \(\mathbb{R}^n\). A restored signal is any vector in the intersection \(S\) of these sets.

The basic digital signal restoration problem of (5.2) was first treated in a set theoretic framework in [90] and [113]. This concept was further formalized in [104], where it was shown that noise properties constituted very valuable pieces of \textit{a priori} knowledge. In [31], the set theoretic restoration problem in the presence of partially defined or inaccurate \textit{a priori} knowledge was approached via fuzzy set theory. Finally, stochastic blurring operators were considered in [32].

It is important to note that the problems presented in [32], [90], [104], and [113] were solved by using the algorithm of Theorem 2.14 in the Euclidean space (in the signal recovery literature, this algorithm is often referred to as POCS, for projection onto convex sets). Therefore, only convex property sets could be considered in these studies, which precluded the incorporation of some important pieces of \textit{a priori} information. There are however many useful properties which yield nonconvex sets in \(\mathbb{R}^n\). In this regard, we can mention the set of signals whose Euclidean norm is bounded from below (minimum energy constraint), the set of signals whose entropy is bounded from above (maximum entropy constraint), the set of signals which do not have more than a fixed number of nonzero values, the set of signals with a fixed number of zero crossings, and the set of signals which have a prescribed Fourier transform magnitude.
5.5. Simulation Results

5.5.1. Signal Degradation

In the results shown here, the original signal is a simulated X-ray fluorescence spectrum. Such signals feature high resolution patterns together with large zero regions and have been used in previous work to test restoration methods (e.g. [31], [103], [104]). The original signal \( h \) has \( n = 64 \) points and is displayed in Figure 5.1. It can be regarded as a plot of photon count versus frequency. It was blurred by convolution with a Gaussian shaped impulse response with a standard deviation of two points. This type of impulse response constitutes a good model for the finite resolution of the measurement instruments. Gaussian white noise (as defined in Section 4.7.2.3.1) with variance \( \sigma^2 = 0.002 \) was then added to obtain the degraded signal \( z \) seen in Figure 5.2. In the following, it is assumed that the blur operator and the aforementioned characteristics of the noise are known. For purpose of reference, the restoration produced by the standard Wiener filter [3] is shown in Figure 5.3. Only a small amount of the original information has been recovered.

5.5.2. Property Sets for Set Theoretic Restoration

In this section, we shall construct the property sets that will be used subsequently in the set theoretic restoration of the degraded signal. Sets based on the properties of the noise and that of the original signal will be considered.

As was seen in Chapter IV, property sets can be constructed from the information that the noise is white and Gaussian with variance \( \sigma^2 \). From (4.2) and (5.2), if \( a \) denotes the restored signal, the finite sample path of the residual process can be represented by the vector
In the following, $T_i(a)$ represents the $i$-th component of the product $Ta$ and $x_i$ the $i$-th component of $x$. We shall denote by $S_1$ the set based on range information of Section 4.4, by $S_2$ the second absolute moment set of Section 4.5.4, and by $S_3$ the spectral set of Section 4.7.3.1. It is noted that the properties of the noise process are consistent with all the assumptions made to derive the expression of these sets. The confidence coefficient is fixed to $1-\varepsilon=95\%$. From (4.12), by centering the confidence interval at zero, $S_1$ can be written as

$$S_1 = \bigcap_{i=1}^{n} C_i \quad \text{where} \quad C_i = \left\{ a \in \mathbb{R}^n \mid |x_i - T_i(a)| \leq \lambda \right\}$$

(5.4)

From the tables of the normal distribution, we get $\lambda=1.96\sigma=0.088$. The expression of $S_2$ is given in (4.27) as

$$S_2 = \left\{ a \in \mathbb{R}^n \mid \eta_2 \leq ||x-Ta|| \leq \xi_2 \right\}$$

(5.5)

where $||.||$ denotes the Euclidean norm in $\mathbb{R}^n$, i.e. the function $N_2$ defined in (4.20). From the tables of the normal distribution we get that $\alpha=1.96$ in (4.28) and (4.29). Consequently, by (4.41), $\eta_2=0.289$ and $\xi_2=0.415$. Finally, from (4.97), $S_3$ can be written as

$$S_3 = \bigcap_{k=0}^{n/2} \left\{ a \in \mathbb{R}^n \mid \left| \sum_{i=1}^{n} (x_i - T_i(a))\exp(-j\frac{2\pi}{n}kt) \right|^2 \leq \xi_k \right\}$$

(5.6)

From the tables of the $\chi_1^2$ distribution, we get that $\beta_1=3.84$ and therefore, from (4.99), $\xi_k=0.492$ for $k=0$ and $k=n/2$. On the other hand, from (4.100), we get that $\xi_k=0.383$ for $0<k<n/2$. 

$$y = x - Ta$$

(5.3)
As an X-ray fluorescence spectrum, the original signal is nonnegative and does not possess more than a few nonzero values, say \( z \) (given that \( n = 64 \), a standard value is \( z = 9 \)). The set of all vectors with nonnegative components whose number of nonzero values does not exceed \( z \) will be denoted by \( S_4 \). In order to investigate the properties of \( S_4 \), let us denote by \( \{e_i \mid 1 \leq i \leq n\} \) the standard orthonormal basis of \( \mathbb{R}^n \) and by \( S_5 \) the first quadrant of \( \mathbb{R}^n \), i.e. the set of vectors with nonnegative components. The set \( S_4 \) can be described as the intersection of \( S_5 \) with the union of the \( n!/z!/(n-z)! \) vector subspaces of \( \mathbb{R}^n \) generated by \( z \) distinct \( e_i \)'s. These finite dimensional vector subspaces and \( S_5 \) are trivially closed and, therefore, so is \( S_4 \). Thus, by Corollary 2.1, \( S_4 \) is proximinal. A projection of a vector \( a \) in \( \mathbb{R}^n \) onto \( S_4 \) is simply obtained by retaining the \( z \) largest nonnegative components of \( a \) and by setting the remaining points to zero. Because the \( z \) largest nonnegative components of a vector are not necessarily uniquely determined, the set \( M_4 \) of multifurcation points of \( S_4 \) is nonempty and it follows from (iv) in Proposition 2.4 that \( S_4 \) is not convex.

5.5.3. Set Theoretic Restoration by the POCS

The solution space \( \mathbb{R}^n \) is equipped with the Euclidean norm \( \| \cdot \| \) and hence becomes a Hilbert space.

It was proved in Section 4.4 that, under our hypotheses, each \( C_i \) in (5.4) is closed and convex. Likewise, \( S_3 \) will also be closed and convex by the results of Section 4.7.3.2. According to the results of Section 4.5.5, \( S_2 \) is closed but not convex and it must be replaced by its convex hull \( S_2^+ \) defined in (4.30). This is not a problem because experience has shown that, regardless of the starting point, the tail of a sequence of successive projections never visits the convex deficiency \( S_2^- \) of \( S_2 \) defined in (4.30). Finally, the set \( S_4 \) cannot be used in the POCS because it is not convex. We can however use the closed and convex superset...
$S_5$ of signals with nonnegative components. The convex sets $C_i$ of (5.4), $S_2^+$, $S_3$, and $S_5$ are similar to that used in [104] and their respective projection operators can be found there.

The POCS was implemented by using the degraded signal as a starting point and by projecting sequentially onto the $C_i$'s, $S_2^+$, $S_3$, and $S_5$. The coefficient of the stopping rule of Section 2.6.4.4 was set to $\epsilon = 10^{-2}$. Convergence to the estimate displayed in Figure 5.4 was obtained in 45 iterations.

5.5.4. Set Theoretic Restoration by the MOSP

The version of the MOSP given by Corollary 2.2 was then used. The MOSP allows us to incorporate the nonconvex set $S_4$. Corollary 2.2 states that, in order for the MOSP to converge, the iterations should be started at a point of attraction of the system of property sets. In Section 2.6.4.3, it was noted that points of attraction were more likely to be found in the vicinity of the solution set $S$. Because the degraded signal $z$ still retains some of the features of the original signal, it constitutes a sensible choice for a starting point. The MOSP was implemented by projecting sequentially onto the $C_i$'s of (5.4), $S_2^+$, $S_3$, and $S_4$. As above, $\epsilon$ was set to $10^{-2}$ in (2.36). Convergence to the feasible signal displayed in Figure 5.5 was achieved in 60 iterations. Since more a priori knowledge has been used, it is not surprising that the MOSP gave a better estimate than the POCS. Indeed, the three peaks on the left are more sharply recovered and the separation between the two main peaks has been improved. Moreover, all the artifacts which appeared in the flat regions of the signal have been removed.
Figure 5.1. The Original Signal.

Figure 5.2. The Degraded Signal.
Figure 5.3. Restoration by Wiener Filtering.

Figure 5.4. Restoration by the POCS.
In this chapter, set theoretic estimation has been applied to digital signal restoration. In using the standard POCS algorithm to generate a set theoretic restoration one is limited to convex property sets. The MOSP developed in Chapter II makes it possible to incorporate nonconvex property sets under the provision that the iterations be started at a point of attraction of the system of property sets. It was argued that the degraded signal was a good candidate for a point of attraction. In the simulations, a nonconvex property set was used to illustrate the advantage of the MOSP over the POCS.
CHAPTER VI

APPLICATION TO HARMONIC RETRIEVAL

6.1. Introduction

The problem of estimating the frequency of sinusoidal signals in additive noise is of special interest in a broad range of signal processing applications. In this chapter, this problem is approached via set theoretic estimation. Aside from being a new area of application of set theoretic estimation, the frequency estimation problem allows us to illustrate two main theoretical developments of previous chapters. First of all, Monte Carlo experiments are performed to give a pictorial description of various noise property sets derived in Chapter IV and to assess their individual contribution to the solution set. Second, a practical implementation of the method of random search (MORS) of Chapter III is demonstrated.

6.2. Background

In digital signal processing, a wide class of spectral analysis problems involve the determination of the spectrum of a harmonic process i.e., as seen in Section 4.7.2.3.3, the location of the jump points of its spectral distribution. For instance, in applications such as sonar or radar signal processing, the points in the spectrum represent physical quantities such as speed or bearing. In geophysics, they play a central rôle in the study of free oscillations of the earth and microseisms.
In this chapter, we shall consider a data process \( \{X_n\}_{n \in \mathbb{Z}} \) consisting of \( q \) real sinusoids corrupted by an additive white noise process \( \{U_n\}_{n \in \mathbb{Z}} \), i.e.

\[
(\forall n \in \mathbb{Z}) \quad X_n = \sum_{k=1}^{q} b_k \sin(2\pi v_k n) + U_n
\]

where \( b_k \) and \( v_k \) represent respectively the amplitude and the digital frequency of the \( k \)-th sinusoid. Physically, the above process is obtained by sampling a continuous process at a rate not less than the Nyquist frequency. Therefore, the digital frequencies \( \{v_1, \ldots, v_q\} \) lie in \([0,1/2]\). The problem is then estimate the \( v_k \)'s from a finite sample path \( \{x_1, \ldots, x_n\} \) of the data process. This task becomes especially difficult when the frequencies are closely spaced, the length \( n \) of the data record short, or the signal-to-noise ratio (SNR) low. We recall that if \( \sigma^2 \) denotes the total power of the noise process, the SNR for the \( k \)-th sinusoid is given by

\[
\text{SNR}_k = 10 \log_{10} \left( \frac{b_k^2}{2\sigma^2} \right)
\]

The harmonic retrieval problem has a long history and no general method clearly stands out in terms of performance and computational complexity. Conventional frequency domain methods operate via the Discrete Fourier Transform (DFT) and are limited to a frequency resolution of \( 1/n \). On the other hand, the performance of modern high resolution frequency estimation techniques (among the most established \([25], [80], \) and \([105])\) is known to degrade severely at low SNR and/or when the length of the data record is short. Finally, under the assumption that the noise is Gaussian, the maximum likelihood method has proven satisfactory but it is computationally involved.
6.3. The Set Theoretic Approach

In the set theoretic approach, the vector whose components are the parameters of the sinusoids (amplitudes, frequencies) constitutes the object to be estimated. Property sets can be constructed by considering two main sources of a priori knowledge. First of all, one can use constraints on the parameters. For instance, it may be known that some of the amplitudes are equal, or in a given ratio, or are within some bounds. Bounds may also be available for the frequencies, or the spacing between them; in some problems, the sinusoids may be harmonically related. As discussed in Chapter IV, noise properties constitute the second source of a priori information. Any common point of all the property sets available for the problem is a solution.

A characteristic of the harmonic retrieval problem is that it is nonlinear with respect to the frequencies and, therefore, deriving the projection operators onto various property sets may be cumbersome. Moreover, since the number of sinusoids is typically small, it is a problem of low dimension. These two factors strongly favor employing the MORS rather than the MOSP for the synthesis of a set theoretic estimate.

6.4. Representation of Noise Property Sets by Monte Carlo Simulation

6.4.1. Introduction

In Chapter IV, we have used various probabilistic properties of the noise process to create sets in the solution space. In the context of harmonic retrieval, we shall now give a pictorial representation of these sets. This experiment will provide great insight into the actual contribution of each noise property in this particular set theoretic estimation problem while illustrating some general statements made in Chapter IV. Because the signal
formation model is nonlinear in terms of the unknown frequencies, it would be very
difficult to get a precise representation of the noise property sets via analytical pro-
cedures. On the other hand, Monte Carlo simulations will allow us to get an almost exact
representation of these sets in the solution space.

6.4.2. Data Formation Model

In order to represent explicitly the property sets, it is best to restrict ourselves to a prob-
lem of dimension two, i.e. $\Xi = \mathbb{R}^2$. For this reason, we shall consider a problem with two
sinusoids in noise, where the amplitudes are known and the two frequencies are the unk-
owns. A more general model with unknown amplitudes will be considered in the next sec-
tion, where the harmonic retrieval problem is actually addressed. Data are obtained by
observing the process

$$
X_n = \sin(2\pi h_1 n) + \sin(2\pi h_2 n) + U_n
$$

In this model the true values of the frequencies are $h_1 = 0.100$ and $h_2 = 0.140$. The noise
process $\{U_n\}_{n \in \mathbb{Z}}$ is white and Gaussian with variance $\sigma^2 = 0.05$, which yields a SNR of 10
dB on each sinusoid. As usual, the observed data finite sample path is denoted by
$\{x_1, \ldots, x_n\}$. The number of available data points is $n = 16$.

6.4.3. Methodology

To represent a given noise property set $S_i$ in the $(h_1, h_2)$ frequency plane, we perform the
following Monte Carlo experiment. Let $a = (a_1, a_2)$ denote the estimate of $h$. A large
number of points $a$ are drawn at random from a uniform distribution over the square
$[0, 1/2] \times [0, 1/2]$. Next, the residual points are computed according to
A simple acceptance/rejection procedure then takes place. Only those \( a \)'s which produce a residual path \( \{y_1, \ldots, y_n\} \) consistent with the noise property in question are retained. The smallest frequency is assigned to \( a_1 \) and the largest to \( a_2 \). The confidence coefficient on the sets is fixed to \( 1 - \epsilon = 95\% \). The scatter plot of these points in the frequency plane represents approximately the set \( S_i \). The extent of a given scatter plot can be used as an empirical measure of how discriminating the corresponding piece of noise information is when the only unknowns are the frequencies.

In constructing the following sets, we use the knowledge that the \( U_n \)'s are Gaussian i.i.d.r.v.'s with mean zero and variance \( \sigma^2 = 0.05 \).

### 6.4.4. The Range Set

From (4.12), by centering the confidence interval at zero, \( S_r \) can be written as

\[
S_r = \bigcap_{i=1}^{n} C_i \quad \text{with} \quad C_i = \left\{ a \in \mathbb{R}^2 \mid |z_i - T_i(a)| \leq \lambda \right\}
\]

From the tables of the normal distribution, we get \( \lambda = 1.96\sigma = 0.438 \). The range set is displayed in Figure 6.1.

### 6.4.5. The Absolute Moment Sets

The expression for the absolute moment set of order \( p \) is given in (4.27) as

\[
S_p = \left\{ a \in \mathbb{R}^2 \mid \eta_p \leq N_p(\mathbf{z} - T(a)) \leq \xi_p \right\}
\]

We shall represent \( S_p \) for \( p = \frac{1}{2}, 1, 2, 4, \) and 8. From (4.41), we compute the corresponding values of the \( E|U|^{p} \)'s and we get that \( \alpha = 1.96 \) in (4.28) and (4.29) from the tables of the
normal distribution. Consequently, we obtain

\begin{align*}
\eta_{1/2} &= 2.42 \times 10^1 \\
\eta_1 &= 1.80 \\
\eta_2 &= 4.96 \times 10^{-1} \\
\eta_4 &= 0 \\
\eta_8 &= 0 \\
\zeta_{1/2} &= 5.65 \times 10^1 \\
\zeta_1 &= 3.91 \\
\zeta_2 &= 1.16 \\
\zeta_4 &= 7.47 \times 10^{-1} \\
\zeta_8 &= 7.29 \times 10^{-1}
\end{align*}

Figures 6.2 through 6.6 represent the set $S_p$ for these values of $p$.

### 6.4.6. The Moment Sets

By Proposition 4.11, the expression for the odd moment set of order $k$ is given in (4.76) as

\begin{equation}
S_k = \left\{ a \in \mathbb{R}^2 \mid \left| \sum_{i=1}^n (x_i - T_i(a))^k \right| \leq \delta_k \right\} \quad \text{where} \quad \delta_k = \alpha \sqrt{nE U^{2k}}
\end{equation}

We shall represent $S_k$ for $k=1$ and $k=3$. We get that $\alpha = 1.96$ from the tables of the normal distribution. Thus $\delta_1 = 1.75$ and $\delta_3 = 3.39 \times 10^{-1}$. The moment sets for $k=1$ and $k=3$ are shown in Figures 6.7 and 6.8 respectively.

### 6.4.7. The Spectral Set

From (4.97), the spectral set is given by

\begin{equation}
S_d = \bigcap_{k=0}^{n/2} D_k \quad \text{with} \quad D_k = \left\{ a \in \mathbb{R}^2 \mid \left| \sum_{i=1}^n (x_i - T_i(a)) \exp\left( -j \frac{2\pi}{n} ki \right) \right|^2 \leq \xi_k \right\}
\end{equation}

From the tables of the $\chi^2_1$ distribution, we get that $\beta_1 = 3.84$ and therefore, from (4.99), $\xi_k = 3.07$ for $k=0$ and $k=n/2$. For $0 < k < n/2$, (4.100) yields $\xi_k = 2.40$. The spectral set is displayed in Figure 6.9.
Figure 6.1. The Range Set.

Figure 6.2. The Absolute Moment Set of Order \(1/2\).
Figure 6.3. The Absolute Moment Set of Order 1.

Figure 6.4. The Absolute Moment Set of Order 2.
Figure 6.5. The Absolute Moment Set of Order 4.

Figure 6.6. The Absolute Moment Set of Order 8.
Figure 6.7. The Moment Set of Order 1.

Figure 6.8. The Moment Set of Order 3.
6.4.8. Discussion

Quantitatively, these simulations show that various noise properties lead to sets of very different shapes and areas. We shall however take care not to draw any general conclusion as to which set is the most useful in determining the solution. For instance, in this particular example, the spectral set is contained in all the absolute moment sets, which become therefore obsolete in its presence. This is however not always true. In general, even in linear problems, deciding which pieces of noise information are most effective in defining a set theoretic estimate is a delicate task for it amounts to establishing that a given property set is a subset of another. Such a decision should therefore be weighed by the experience of the user with similar problems.

It is interesting to note that the spectral set does not contain the true solution for that
particular data frame. This behavior should be expected in 5% of the cases since the confidence coefficient was set to 95%. This fact was confirmed by testing a large number of different data frames. Finally, let us remark that the shape of the third moment set in Figure 6.8 proves that the diameter of a nonconvex set is a poor measure of its effectiveness in a set theoretic framework. Indeed, a set can possess distant extreme points while being "narrow" in the neighborhood of the solutions.

6.5. Harmonic Retrieval by the MORS

In this section, we shall demonstrate an application of the MORS to the harmonic retrieval problem. We shall compare the results of the MORS against that of the well established Pisarenko method [80]. Let us briefly recall that in Pisarenko's method one forms the covariance matrix $C$ of the process based on the exact values of the correlation function at $2q+1$ lags, including the zero lag. The zeroes of the polynomial whose coefficients are the components of the eigenvector associated with the smallest eigenvalue of $C$ occur in conjugate pairs of the form $Z_k = \exp(\pm j2\pi\nu_k)$. Thus, the arguments of the poles give the angular frequencies of the sinusoids. In practice, the covariance matrix is estimated from a finite segment of the data process $\{x_1, ..., x_n\}$ and, therefore, this result is only approximate.

6.5.1. Data Formation Model

The data are obtained by observing $n = 16$ points of the process

$$X_n = h_3 \sin(2\pi h_1 n) + h_4 \sin(2\pi h_2 n) + U_n$$  \hspace{1cm} (6.10)

In this model, the true values of the frequencies are $h_1 = 0.100$ and $h_2 = 0.140$, and the true values of the amplitudes are $h_3 = h_4 = 1.00$. It is noted that the frequency spacing is well
beyond the resolution \(1/n\) of DFT methods. As before, the noise process \(\{U_n\}_{n \in \mathbb{Z}}\) is white and Gaussian and the SNR on each sinusoid is 10 dB. Since the amplitudes and the frequencies are the unknowns, the true object \(h\) lies in \(\mathbb{R}^4\), which is thus the solution space for this set theoretic estimation problem.

### 6.5.2. Property Sets for the Problem

To construct property sets for the problem, three standard pieces of \textit{a priori} knowledge will be assumed.

(i) The data were obtained by sampling at least at twice the Nyquist rate.

(ii) The power of one sinusoid is not less than ten times that of the other.

(iii) The noise process is white and Gaussian with known power \(\sigma^2\).

We obtain from (i) that the true frequencies \(h_1\) and \(h_2\) lie in \([0,0.25]\). Now, let \(p_z\) be the total power of the data process. We have

\[
p_z = \frac{h_3^2 + h_4^2}{2} + \sigma^2 \tag{6.11}
\]

It follows from (ii) and (6.11) that

\[
B \leq h_3, h_4 \leq \sqrt{10B} \quad \text{where} \quad B = \sqrt{2(p_z - \sigma^2)/11} \tag{6.12}
\]

An estimate for the total power of the data process is

\[
\hat{p}_z = \frac{1}{n} \sum_{i=1}^{n} x_i^2 \tag{6.13}
\]

Extensive simulation using 16-point data frames formed in accordance with (6.10) provides numerical evidence that the probability that \(|p_z - \hat{p}_z|/\hat{p}_z\) be greater than 0.30 is negligible. Hence, since our data record gives \(\hat{p}_z = 0.803\), \(p_z\) ranges from 0.56 to 1.05. By substituting these extreme values in (6.12), we get that \(0.30 \leq h_3, h_4 \leq 1.35\). The amplitudes
should therefore be in the interval \([0.30,1.35]\). Hence, the property set based on the information relative to the amplitudes and the frequencies is

\[
S_1 = [0,0.25] \times [0,0.25] \times [0.30,1.35] \times [0.30,1.35]
\]  

(6.14)

The probabilistic information pertaining to the noise process given in (iii) can also be used to construct property sets for the problem. Given an estimate \(a\) of \(h\) in \(\mathbb{R}^4\), the residual points are computed according to

\[
(\forall i \in \{1,\ldots,n\}) \quad y_i = x_i - T_i(a) = x_i - a_3 \sin(2\pi a_1 i) - a_4 \sin(2\pi a_2 i)
\]  

(6.15)

The noise property sets used here will be that described in Section 6.4. Their expression (with the understanding that \(T_i\) is as in (6.15) and \(a\) lies in \(\mathbb{R}^4\)) and the value of their parameters will remain the same since the confidence coefficient is kept at 95% and the noise is as in Section 6.4.2 in the present experiment.

6.5.3. Determination of the Search Region

We need to select a subset \(G\) of \(\mathbb{R}^4\) according to the criteria established in Section 3.4.4. Although \(S_1\) is not the set of smallest Lebesgue measure, it will be chosen as a search region because, as a hyperrectangle, it allows uniform generation of points at minimum computational cost. Thus, \(G = S_1\) and the search region for the problem is given by (6.14). In the presence of more \textit{a priori} knowledge, the search region could be significantly improved. Such would be the case if it was known that the sinusoids had equal power, or were harmonically related, or if a more accurate range for the frequencies was available.

6.5.4. Results

In order to give a precise description of the performance of the MORS, we need to generate hundreds of solutions. It is understood that in practice one should stop at the first
feasible point produced by the MORS. To this end, $10^8$ points were drawn from a uniform distribution over $G$. Among these, 3093 were found to be feasible, which indicates that the expected number of trials required to obtain a solution for this problem is $\bar{\eta} \approx 3.2 \times 10^4$. The MORS produces quadruples $a = (a_1, a_2, a_3, a_4)$ in $\mathbb{R}^4$. Since we are interested in the frequencies here, i.e. the two first components of $a$, let $S$ be the intersection of the solution set given by the MORS with the $(a_1, a_2)$ plane. This set is displayed in Figure 6.10, the smallest frequency being assigned to $a_1$ and the largest to $a_2$. The statistics of the points in $S$ are as follows. The mean and the standard deviation are $m_1 = 9.90 \times 10^{-2}$ and $\sigma_1 = 2.30 \times 10^{-3}$ for $a_1$, $m_2 = 13.92 \times 10^{-2}$ and $\sigma_2 = 1.61 \times 10^{-3}$ for $a_2$. Since the MORS is equally likely to produce any of these solutions, it is important to consider the worst case: in terms of the Euclidean distance from the true object $(h_1, h_2)$, the worst feasible frequencies produced by the MORS are $(a_1, a_2) = (0.094, 0.136)$. This compares favorably with the standard Pisarenko method which gave $(a_1, a_2) = (0.095, 0.189)$.

To study the behavior of the MORS in a very noisy environment, we reran the same experiment with a SNR of 5 dB on each sinusoid, i.e. $\sigma^2 = 0.158$. Figure 6.11 shows the set $S$. Its statistics are $m_1 = 9.78 \times 10^{-2}$ and $\sigma_1 = 5.70 \times 10^{-3}$ for $a_1$, $m_2 = 13.84 \times 10^{-2}$ and $\sigma_2 = 3.14 \times 10^{-3}$ for $a_2$. Since the power of the noise has increased, the parameters of the noise property sets vary accordingly and the solution set becomes larger. This simply confirms that under more uncertainty, the set of feasible points, i.e. the class of objects which may have given rise to the data, increases. The worst frequencies found by the MORS are $(a_1, a_2) = (0.086, 0.131)$. Pisarenko's method gave $(a_1, a_2) = (0.098, 0.200)$.

Finally, let us note that, in this particular problem, the same results were obtained by using only the range and spectral sets and leaving the moment and absolute moment sets out.
Figure 6.10. The Set Theoretic Solutions - SNR = 10 dB.

Figure 6.11. The Set Theoretic Solutions - SNR = 5 dB.
6.6. Summary

In this chapter, we have used the problem of harmonic retrieval to illustrate some of the concepts and methods developed earlier. In the simulations, the sum of two closely spaced sinusoids in white noise was used to represent various noise property sets derived in Chapter IV and to demonstrate an application of the MORS developed in Chapter III. In severe conditions (16 sample points and SNR's of 10 and 5 dB's) the MORS gave very promising results and resolved the sinusoids far better than the standard Pisarenko method. These results could be further improved by incorporating any additional piece of a priori information which may be available in specific applications. Another option to improve the results in applications where computation time is not a crucial factor is to let the MORS generate more than one solution and to average them, under the provision that the solution set is convex and well centered about the true solution, as was the case in the experiment presented above. Finally, these simulations also showed that the MORS is an easily implementable tool for the synthesis of set theoretic estimates, regardless of the complexity of the analytical expressions which describe the a priori knowledge.
CHAPTER VII

CONCLUSIONS

Although set theoretic methods had been used in scattered applications in the literature, they lacked abstracting concepts and mathematical methods. The main objective of this dissertation was to alleviate these shortcomings in order to consolidate the foundations of set theoretic estimation and extend the range of its applications.

The first task was to formally define set theoretic estimation as a technique in which consistency with the \textit{a priori} knowledge serves as an estimation criterion. Each piece of \textit{a priori} knowledge is represented by a set in a solution space. Set theoretic estimates differ from classical estimates in that they are not characterized in terms of optimality with respect to some criterion. Rather, they are objects which do not violate those constraints about the problem which are known \textit{a priori}.

The only method available for the synthesis of set theoretic estimates was that of successive projections onto closed and convex sets in a Hilbert space. This method clearly imposes stringent conditions on the sets and the underlying solution space. We have introduced two new methods, which relax these restrictive assumptions. First, the method of successive projections (MOSP) was generalized to approximately compact sets in arbitrary metric spaces and convergence results were established. Second, a method of random search (MORS) was developed, in which a solution is sought by randomly searching a restricted region of the solution space. It is worth noting that these results have a natural extension to the synthesis of fuzzy set theoretic estimates, a problem for which two
methods were proposed in [31]. In the first approach, the problem was formulated as that of finding a common point of the $\alpha$-level cuts of the fuzzy sets and successive projections onto these sets were employed to this end. This required that all the $\alpha$-level cuts be convex, an assumption which can be released through the use of the MOSP or of the MORS. The second approach was to carry out a direct deterministic optimization of the membership function of the fuzzy intersection set. As was seen in Section 3.7, it may be more advantageous to employ stochastic optimization to perform this task.

In the past, noise properties had been employed to a very limited extent to construct sets in the solution space, as they had been restricted to signal restoration problems in the presence of additive white Gaussian noise. The use of noise properties has been generalized in three respects. First of all, a general data formation model has been considered. Second, the analysis has departed from the assumption that the noise was white and Gaussian. Finally, new pieces of information such as moments and absolute moments of arbitrary order, second and higher order properties have been considered.

The methods presented in this dissertation have been developed in their full generality and have potential applications in a wide spectrum of fields, from economics to engineering. They are well suited for digital signal processing, an area where estimation problems with specific pieces of a priori information abound. In particular, applications in digital signal restoration and harmonic retrieval were demonstrated and set theoretic estimates were seen to outperform their conventional counterparts. Among the other areas of digital signal processing in which set theoretic estimation is anticipated to meet with the same success, we can mention system identification, spectral estimation, coding, signal reconstruction, radar and sonar processing, and seismic data processing.
From a theoretical standpoint, there remain several open questions which deserve further investigation and we shall briefly outline some of them. The concept of region of attraction was seen to be central in the convergence properties of the MOSP. In the very general context of approximately compact sets in which the MOSP was developed, it is almost impossible to describe such a region. Hence, it may be worth considering less general classes of sets (e.g. sets which are nonconvex but have some regularity properties) for which the region of attraction could be determined more explicitly. This would in turn provide the user with a less heuristic means to determine where to start the iterations in order to obtain a solution. The performance of the MORS was seen to be prone to fast deterioration as the dimension of the problem increases. Therefore, the question of adapting the search could be investigated and better techniques to determine the search region could be studied. Finally, there does not seem to be a general method to identify those noise properties which will be redundant in defining a set theoretic estimate. Such an analysis could, however, probably be done in simple cases, e.g. linear signal formation model.

REFERENCES


Throughout this dissertation, the following notations will be employed. \( \mathbb{Z} \) is the set of integers, \( \mathbb{N} \) is the set of nonnegative integers, \( \mathbb{R} \) is the set of real numbers, \( \mathbb{R}^+ \) is the set of nonnegative real numbers, \( \mathbb{R}_+ \) is the set of positive real numbers, \([a,b]\) is a closed interval, \([a,b[\) is an open interval, \(]a,b]\) and \([a,b[\) are half closed intervals. \( \mathbb{C} \) denotes the set of complex numbers. To conform to engineering notations, the \( \pi/2 \) rotation operator in \( \mathbb{C} \) is denoted by \( j \) instead of \( i \). If \( z \) is a complex number, \( \bar{z} \) denotes its conjugate and \( \text{Re}(z) \) its real part. The quantifiers \( \forall, \exists, \) and \( \exists! \) mean respectively "for all", "there exists at least one", and "there exists exactly one".

A.1. Set Theory

Let \( \Xi \) be a space, i.e. a nonempty abstract set composed of elements called points. The relation \( x \in \Xi \) means that \( x \) is an element of \( \Xi \). Its negation is written \( x \notin \Xi \). If \( S \) is another set, the relation \( S \subseteq \Xi \) means that every element of \( S \) is an element of \( \Xi \). Then, one says that \( \Xi \) contains \( S \) or that \( S \) is a subset of \( \Xi \). The subset of \( \Xi \) which contains no element is denoted by \( \emptyset \) and is called the empty set. The class of all subsets of \( \Xi \) is denoted by \( \mathbb{P} \Xi \). Let \( A \) be a nonempty index set and let \( \{S_a\}_{a \in A} \) be an arbitrary class of sets in \( \Xi \). Set union, intersection, and complementation are respectively defined as
De Morgan's laws read

\[
\left( \bigcup_{\alpha \in A} S_\alpha \right)^c = \bigcap_{\alpha \in A} S_\alpha^c \quad \text{and} \quad \left( \bigcap_{\alpha \in A} S_\alpha \right)^c = \bigcup_{\alpha \in A} S_\alpha^c
\]

(A.2)

\(\{S_\alpha\}_{\alpha \in A}\) is said to cover the set \(S\) if \(S \subseteq \bigcup_{\alpha \in A} S_\alpha\). A sequence \(\{S_n\}_{n \geq 0}\) of sets in \(\Xi\) is said to be increasing [respectively decreasing] if

\[
(\forall n \in \mathbb{N}) \quad S_n \subseteq S_{n+1} \quad [\text{respectively } S_{n+1} \subseteq S_n] \quad \text{(A.3)}
\]

If \(S_2 \subseteq S_1\), then \(S_1 \cap S_2^c\) is denoted by \(S_1 - S_2\). The Cartesian product of two spaces \(\Xi_1\) and \(\Xi_2\) is defined as

\[
\Xi_1 \times \Xi_2 = \left\{ (x_1, x_2) \mid x_1 \in \Xi_1, x_2 \in \Xi_2 \right\} \quad \text{(A.4)}
\]

The inverse image under a function \(X: \Xi_1 \rightarrow \Xi_2\) of a subset \(S\) of \(\Xi_2\) is defined as

\[
X^{-1}(S) = \left\{ x \in \Xi_1 \mid X(x) \in S \right\} \quad \text{(A.5)}
\]

For every class \(\{S_\alpha\}_{\alpha \in A}\) of subsets of \(\Xi_2\), we have

(i) \(X^{-1}\left( \bigcap_{\alpha \in A} S_\alpha \right) = \bigcap_{\alpha \in A} X^{-1}(S_\alpha)\)

(ii) \(X^{-1}\left( \bigcup_{\alpha \in A} S_\alpha \right) = \bigcup_{\alpha \in A} X^{-1}(S_\alpha)\)

(iii) \((\forall \alpha \in A) \quad X^{-1}(S_\alpha^c) = (X^{-1}(S_\alpha))^c\)

References: N. Bourbaki [14], F. Hausdorff [49].
A.2. Topology

Let \( \Xi \) be a set. A class \( \tau \) of subsets of \( \Xi \) is said to be a topology on \( \Xi \) if

(i) \( \emptyset \in \tau \) and \( \Xi \in \tau \).

(ii) \( \tau \) is closed under finite intersection.

(iii) \( \tau \) is closed under arbitrary union.

The class \( \tau \) is also called the class of open sets of the topological space \((\Xi, \tau)\). A subset \( V \) of \((\Xi, \tau)\) is said to be a neighborhood of a point \( z \) in [respectively of a nonempty subset \( S \) of] \((\Xi, \tau)\) if there exists an open set \( T \) such that \( z \in T \subseteq V \) [respectively \( S \subseteq T \subseteq V \)]. A subset \( S \) of \((\Xi, \tau)\) is open if and only if it is a neighborhood of each of its points. The interior of \( S \) is the largest open set \( S^o \) contained in \( S \). A subclass \( \beta \) of \( \tau \) is said to be a base for \( \tau \) if for every point \( z \) in \((\Xi, \tau)\) and each neighborhood \( V \) of \( z \) there exists a set \( B \) in \( \beta \) such that \( z \in B \subseteq V \). A subset \( S \) of \((\Xi, \tau)\) is also a topological space with the relative (or induced) topology \( \tau_S = \{ S \cap T \mid T \in \tau \} \). Let \( \{ z_n \}_{n \geq 0} \) be a sequence of points in \((\Xi, \tau)\) and let \( x \) be a point in \((\Xi, \tau)\). Then \( x \) is said to be a cluster point of \( \{ z_n \}_{n \geq 0} \) if for every neighborhood \( V \) of \( x \) and every \( m \) in \( \mathbb{N} \) there exists an integer \( n \) greater than or equal to \( m \) such that \( z_n \in V \). The sequence \( \{ z_n \}_{n \geq 0} \) is said to converge to \( x \) if for every neighborhood \( V \) of \( x \) there exists an \( m \) in \( \mathbb{N} \) such that for every integer \( n \) greater than or equal to \( m \) we have \( z_n \in V \). \((\Xi, \tau)\) is said to be a Hausdorff space if two distinct points have disjoint neighborhoods. All of the topological spaces to be considered in this dissertation are Hausdorff. The dual class of complements of open sets is called the class of closed sets and is closed under arbitrary intersections and finite unions. Let \( S \) be a subset of \((\Xi, \tau)\). The closure of \( S \) is the smallest closed set \( \overline{S} \) containing \( S \). \( S \) is said to be dense in \( \Xi \) if \( \overline{S} = \Xi \). The class of all nonempty closed subsets of \( S \) is denoted by \( 2^S \). Finally, \( S \) is called compact if every class of open sets covering \( S \) contains a finite subclass also covering \( S \).
Let \((\Xi_1, \tau_1)\) and \((\Xi_2, \tau_2)\) be two topological spaces. Then the product topology on \(\Xi_1 \times \Xi_2\) has \(\{T_1 \times T_2 \mid T_1 \in \tau_1, T_2 \in \tau_2\}\) as a base. Moreover, a function \(X\) from \(\Xi_1\) into \(\Xi_2\) is said to be continuous if the inverse image of every open set is an open set, that is

\[
(\forall T_2 \in \tau_2) \quad X^{-1}(T_2) \in \tau_1
\]  

(A.6)

A nonnegative real-valued function \(d(.,.)\) with domain \(\Xi \times \Xi\) is called a distance if

\begin{enumerate}
  \item \((\forall (x,y) \in \Xi^2)\) \quad d(x,y) = 0 \iff x = y
  \item \((\forall (x,y) \in \Xi^2)\) \quad d(x,y) = d(y,x)
  \item \((\forall (x,y,z) \in \Xi^3)\) \quad d(x,z) \leq d(x,y) + d(y,z)
\end{enumerate}

\((\Xi, d)\) is called a metric space. The diameter of a nonempty set \(S\) in \((\Xi, d)\) is the number \(\delta\) in \([0, +\infty]\) defined by \(\delta = \text{Sup} \{d(x, y) \mid x \in S, y \in S\}\). \(S\) is said to be bounded if its diameter is finite. Let \(a\) be a point in \((\Xi, d)\) and \(r\) a positive real number. The open and closed balls of center \(a\) and radius \(r\) in \((\Xi, d)\) are respectively defined as

\[
B[a, r] = \left\{ x \in \Xi \mid d(a, x) < r \right\} \quad \text{and} \quad B[a, r] = \left\{ x \in \Xi \mid d(a, x) \leq r \right\}
\]  

(A.7)

The topology of a metric space is determined by the requirement that the class of all open balls be a base. Hence, in \((\Xi, d)\), a set \(S\) is open if

\[
(\forall a \in S)(\exists r \in \mathbb{R}_+) \quad B[a, r] \subset S
\]  

(A.8)

Moreover, a sequence \(\{x_n\}_{n \geq 0}\) of points in \((\Xi, d)\) converges to a point \(x\) in \((\Xi, d)\) if \(\{d(x_n, x)\}_{n \geq 0}\) converges to zero in (the usual topology of) \(\mathbb{R}\), i.e.

\[
(\forall r \in \mathbb{R}_+)(\exists p \in \mathbb{N})(\forall n \in \mathbb{N} \mid n \geq p) \quad d(x_n, x) < r
\]  

(A.9)

It is called a Cauchy sequence if \(d(x_m, x_n)\) goes to zero as \(m\) and \(n\) go to \(+\infty\), i.e.

\[
(\forall r \in \mathbb{R}_+)(\exists p \in \mathbb{N})(\forall m, n \in \mathbb{N} \mid m \geq p, n \geq p) \quad d(x_m, x_n) < r
\]  

(A.10)

A metric space \(\Xi\) is called complete if every Cauchy sequence in \(\Xi\) converges to a point in
A set $S$ in a metric space is closed if every convergent sequence with elements in $S$ has its limit in $S$. Let $\{x_n\}_{n \geq 0}$ be a sequence of points in a metric space $\Xi$ and let $x$ be a point in $\Xi$. Then $x$ is a cluster point of $\{x_n\}_{n \geq 0}$ if there exists a subsequence $\{x_{n_k}\}_{k \geq 0}$ of $\{x_n\}_{n \geq 0}$ converging to $x$. A set $S$ in a metric space is compact if every sequence with elements in $S$ admits at least one cluster point in $S$.

References: N. Bourbaki [15], [16], J. Kelley [56], K. Kuratowski [60], [61], L. Schwartz [89].

A.3. Topological Vector Spaces

A vector space $\Xi$ over a field $K$ is an abstract space of object called vectors which is endowed with an operation '$+$' from $\Xi \times \Xi$ into $K$ called addition such that

(i) $(\forall (x,y) \in \Xi^2) \quad x + y = y + x$

(ii) $(\forall (x,y,z) \in \Xi^3) \quad x + (y + z) = (x + y) + z$

(iii) $(\exists 0 \in \Xi)(\forall x \in \Xi) \quad x + 0 = x$

(iv) $(\forall x \in \Xi)(\exists (-x) \in \Xi) \quad x + (-x) = 0$

and an operation '$\cdot$' from $K \times \Xi$ into $\Xi$ called scalar multiplication such that

(v) $(\forall \alpha \in K)(\forall (x,y) \in \Xi^2) \quad \alpha.(x+y) = \alpha.x + \alpha.y$

(vi) $(\forall (\alpha,\beta) \in K^2)(\forall x \in \Xi) \quad (\alpha + \beta).x = \alpha.x + \beta.x$

(vii) $(\forall (\alpha,\beta) \in K^2)(\forall x \in \Xi) \quad (\alpha \beta).x = \alpha.(\beta.x)$

(viii) $(\forall x \in V) \quad 1.x = x$

where $1$ is the unit element of $K$. A nonempty subset $S$ of $\Xi$ is a vector subspace if

$$(\forall \alpha \in K)(\forall (x,y) \in \Xi^2) \quad \alpha.x + y \in S \quad (A.11)$$

A real-valued function $X$ defined on $\Xi$ is convex if
and it is concave if $-X$ is convex. A subset $S$ of $\mathbb{R}$ is convex if

$$(\forall \alpha \in [0,1])(\forall (x,y) \in S^2) \quad \alpha x + (1 - \alpha) y \in S$$

(A.13)

An arbitrary intersection of convex sets is a convex set. The convex hull of a set $S$ is the smallest convex set containing $S$. From now on, the dot will be omitted in scalar multiplications and $K$ will be either $\mathbb{R}$ or $\mathbb{C}$. A topological vector space is a vector space $\mathbb{E}$ endowed with a topology such that the two mappings

$$\mathbb{E} \times \mathbb{E} \rightarrow \mathbb{E} \quad \text{and} \quad K \times \mathbb{E} \rightarrow \mathbb{E}$$

(A.14)

$$(x, y) \mapsto x + y \quad (\alpha, x) \mapsto \alpha x$$

are continuous ($K$ being equipped with the usual topology). A norm on $\mathbb{E}$ is a nonnegative real-valued function $\| \cdot \|$ with domain $\mathbb{E}$ such that

(i) $(\forall x \in \mathbb{E}) \quad \| x \| = 0 \quad \iff \quad x = 0$

(ii) $(\forall \alpha \in K)(\forall x \in \mathbb{E}) \quad \| \alpha x \| = |\alpha| \| x \|$

(iii) $(\forall (x,y) \in \mathbb{E}^2) \quad \| x + y \| \leq \| x \| + \| y \|$

$(\mathbb{E}, \| \cdot \|)$ is called a normed vector space (n.v.s.). For every norm $\| \cdot \|$, the relation

$$(\forall (x,y) \in \mathbb{E}^2) \quad d(x,y) = \| x - y \|$$

(A.15)

defines a distance. Thus, every n.v.s. is a metric vector space. Conversely, only those metric vector spaces $(\mathbb{E}, d)$ where

$$\begin{cases} (\forall (x,y,z) \in \mathbb{E}^3) \quad d(x + z, y + z) = d(x, y) \\ (\forall \alpha \in K)(\forall (x,y) \in \mathbb{E}^2) \quad d(\alpha x, \alpha y) = |\alpha| d(x, y) \end{cases}$$

(A.16)

are n.v.s.'s. The norm must then be defined as

$$(\forall x \in \mathbb{E}) \quad \| x \| = d(x,0)$$

(A.17)
Every n.v.s. is a topological vector space with the topology defined by the distance induced by its norm. In a n.v.s., the convergence of (A.9) is called strong convergence. A Banach space is a complete n.v.s. Every finite dimensional n.v.s. is a Banach space. A scalar product on \( \Xi \) is a \( K \)-valued function \( \langle \cdot, \cdot \rangle \) with domain \( \Xi \times \Xi \) which satisfies

(i) \( (\forall x \in \Xi) \quad x \neq 0 \implies \langle x, x \rangle > 0 \)

(ii) \( (\forall \alpha \in K) (\forall (z, y) \in \Xi^2) \quad \langle \alpha z, y \rangle = \alpha \langle z, y \rangle \)

(iii) \( (\forall (z, y, z) \in \Xi^3) \quad \langle z + y, z \rangle = \langle z, z \rangle + \langle y, z \rangle \)

(iv) \( (\forall (x, y) \in \Xi^2) \quad \langle y, x \rangle = \overline{\langle x, y \rangle} \)

A pre-Hilbert space is a vector space \( \Xi \) endowed with a scalar product. In a pre-Hilbert space \( (\Xi, \langle \cdot, \cdot \rangle) \), the scalar product induces a norm as follows

\[
(\forall z \in \Xi) \quad \|z\| = \sqrt{\langle z, z \rangle} \tag{A.18}
\]

Thus, a pre-Hilbert space is a n.v.s.; conversely, only those n.v.s.'s \( (\Xi, \|\cdot\|) \) for which the parallelogram property

\[
(\forall (z, y) \in \Xi^2) \quad \|z + y\|^2 + \|z - y\|^2 = 2 \left( \|z\|^2 + \|y\|^2 \right) \tag{A.19}
\]

holds are pre-Hilbert spaces. The scalar product is then given by the polarization identity

\[
(\forall (x, y) \in \Xi^2) \quad \langle x, y \rangle = \begin{cases} 
\frac{1}{4} \left( \|z + y\|^2 - \|z - y\|^2 \right) & \text{if } K = \mathbb{R} \\
\frac{1}{4} \left( \sum_{n=1}^{4} \langle j^n x + j^n y, z \rangle \right) & \text{if } K = \mathbb{C} 
\end{cases} \tag{A.20}
\]

Two vectors \( x \) and \( y \) are said to be orthogonal if \( \langle x, y \rangle = 0 \), and one writes \( x \perp y \). A Hilbert space is a complete pre-Hilbert space. Alternatively, a Hilbert space is a Banach space whose norm has the parallelogram property.

References: N. Bourbaki [17], L. Schwartz [89], K. Yosida [112].
A.4. Functional Analysis

In this section, $\Xi$ is a vector space over $K$ ($\mathbb{R}$ or $\mathbb{C}$). A functional is an operator $T$ from $\Xi$ into $K$. $T$ is a linear functional if

\begin{equation}
(\forall (\alpha, \beta) \in K^2)(\forall (x, y) \in \Xi^2)
T(\alpha x + \beta y) = \alpha T(x) + \beta T(y)
\end{equation}

(A.21)

The dual space of a n.v.s. $(\Xi, \|\cdot\|)$ is denoted by $\Xi'$ and is the set of all continuous linear functionals defined on $\Xi$. It is a Banach space with the norm

\begin{equation}
(\forall T \in \Xi')
\|T\| = \text{Sup} \left\{|T(x)| \mid x \in \Xi, \|x\| \leq 1\right\}
\end{equation}

(A.22)

A sequence $\{x_n\}_{n=0}$ in a n.v.s. $(\Xi, \|\cdot\|)$ is said to converge weakly to a point $x$ in $(\Xi, \|\cdot\|)$ if, for every $T$ in $\Xi'$, the sequence $\{T(x_n)\}_{n=0}$ converges to $T(x)$ in $K$. Then $x$ is called the weak limit. If a sequence converges strongly to a point, it converges weakly to that point. In finite dimensional spaces, the converse is also true. A set $S$ in a n.v.s. is said to be sequentially weakly closed if the weak limit of every sequence with elements in $S$ belongs to $S$. If $S$ is sequentially weakly closed it is closed. A set $S$ in a n.v.s. is said to be weakly compact if every sequence $\{x_n\}_{n=0}$ with elements in $S$ has a subsequence $\{x_{n_k}\}_{k=0}$ converging weakly to an element in $S$. If $S$ is compact it is weakly compact. If $(\Xi, \langle \cdot, \cdot \rangle)$ is a Hilbert space, $\{x_n\}_{n=0}$ converges weakly to $x$ if and only if $\{\langle x_n, y \rangle\}_{n=0}$ converges to $\langle x, y \rangle$ in $K$, for every $y$ in $\Xi$.

References: N. Bourbaki [17], M. Day [34], K. Yosida [112].
APPENDIX B

ELEMENTS OF MEASURE THEORY, INTEGRATION, PROBABILITY THEORY, AND STOCHASTIC PROCESSES

The set of extended real numbers is defined as \( \mathbb{R} = \mathbb{R} \cup \{+\infty\} \cup \{-\infty\} \). The algebraic conventions are \( \infty + \infty = \infty, \ 0 \times \infty = 0 \). The expression \( \infty - \infty \) is not defined.

B.1. Measure Theory

A class \( \Sigma \) of subsets of a space \( \Omega \) is said to be a \( \sigma \)-algebra if

(i) \( \emptyset \in \Sigma \) and \( \Omega \in \Sigma \).

(ii) \( \Sigma \) is closed under complementation.

(iii) \( \Sigma \) is closed under countable unions.

Let \( \Sigma \) be a \( \sigma \)-algebra of subsets of \( \Omega \). Then \( (\Omega, \Sigma) \) is called a measurable space and the sets in \( \Sigma \) are said to be measurable. The smallest \( \sigma \)-algebra containing a class \( \Gamma \) of subsets of \( \Omega \) is denoted by \( \sigma(\Gamma) \). If \( \tau \) is a topology on \( \Omega \), \( \sigma(\tau) \) is called the Borel \( \sigma \)-algebra. Henceforth, the Borel sets in \( \mathbb{R}^k \) will be that generated by the Euclidean topology. A measure on a \( \sigma \)-algebra \( \Sigma \) is an extended real-valued nonnegative set function \( \mu \) such that

(i) \( \mu(\emptyset) = 0 \).

(ii) For every disjoint sequence \( \{S_n\}_{n \geq 0} \) of sets in \( \Sigma \), \( \mu(\bigcup_{n \geq 0} S_n) = \sum_{n \geq 0} \mu(S_n) \).

\( (\Omega, \Sigma, \mu) \) is called a measure space. A property is said to hold \( \mu \)-almost everywhere (\( \mu \)-a.e.) if it holds everywhere except on a subset of a set \( S \) in \( \Sigma \) such that \( \mu(S) = 0 \).
Let \( \nu \) be another measure on \((\Omega, \Sigma)\). Then \( \nu \) is said to be absolutely continuous with respect to \( \mu \) if
\[
(\forall S \in \Sigma) \quad \mu(S) = 0 \implies \nu(S) = 0 \tag{B.1}
\]
Let \( \Delta \) be the Borel \( \sigma \)-algebra of \( \mathbb{R} \) and let \( \mu \) be a measure on \( \Delta \) which assigns finite values to bounded intervals. Consider the \( \sigma \)-algebra
\[
\Delta_\mu = \left\{ B \cup S \mid B \in \Delta, S \subset N \in \Delta \text{ with } \mu(N) = 0 \right\} \tag{B.2}
\]
Then \( \mu \) can be extended from a measure on \( \Delta \) to a measure on \( \Delta_\mu \), called Lebesgue-Stieltjes measure and that will also be denoted by \( \mu \), by assigning to every set in \( \Delta_\mu \) the \( \mu \)-measure of the Borel set from which it differs by a subset of a \( \mu \)-null Borel set. A function \( F \) from \( \mathbb{R} \) into \( \mathbb{R} \) which is nondecreasing and left continuous is called a distribution function (d.f.) on \( \mathbb{R} \). For all real numbers \( a \leq b \) the relation
\[
F(b) - F(a) = \mu([a, b]) \tag{B.3}
\]
establishes a one-to-one correspondence between the Lebesgue-Stieltjes measures and the d.f.'s defined up to an additive constant. In particular, if \( (\forall c \in \mathbb{R}) \ F(c) = c \), then \( \mu \) is called the Lebesgue measure on \( \mathbb{R} \) and assigns to an elementary intervals \([a, b]\) its length \( b - a \). The Lebesgue measure in \( \mathbb{R}^k \) is defined in a similar manner, as the extension of the measure which assigns to the product of elementary \( k \)-dimensional intervals its \( k \)-dimensional volume.

B.2. Integration

A real-valued function \( X \) on a measurable space \((\Omega, \Sigma)\) is said to be (Borel) measurable if the set \( X^{-1}(B) \) belongs to \( \Sigma \) for every Borel subset \( B \) of \( \mathbb{R} \) or, equivalently, if
A continuous function of measurable functions is measurable. Henceforth, \( X \) is a measurable function on the measure space \((\Omega, \Sigma, \mu)\). The integral of \( X \) over \( \Omega \) with respect to \( \mu \) will be denoted by \( \int_{\Omega} X(\omega) \mu(d\omega) \), or simply \( \int_{\Omega} X d\mu \).

The indicator of a subset \( S \) of \( \Omega \) is the function denoted by \( 1_S \) which takes value 1 on \( S \) and 0 on \( S^c \). \( X \) is said to be a nonnegative simple function if it can be written as a finite sum \( X = \sum_{i=1}^{k} x_i 1_{S_i} \), where the \( x_i \)'s are in \( \mathbb{R}_+ \) and the \( S_i \)'s are disjoint sets in \( \Sigma \). Then

\[
\int_{\Omega} X d\mu = \sum_{i=1}^{k} x_i \mu(S_i) \tag{B.5}
\]

If \( X \) is nonnegative, there exists an increasing sequence \( \{X_n\}_{n \geq 0} \) of nonnegative simple functions converging to \( X \) and

\[
\int_{\Omega} X d\mu = \lim_{n \to +\infty} \int_{\Omega} X_n d\mu \tag{B.6}
\]

Now, consider the two nonnegative measurable functions defined by \( X^+ = \max\{X, 0\} \) and \( X^- = \max\{-X, 0\} \). Then \( |X| = X^+ + X^- \), \( X = X^+ - X^- \), and

\[
\int_{\Omega} X d\mu = \int_{\Omega} X^+ d\mu - \int_{\Omega} X^- d\mu \tag{B.7}
\]

This integral exists provided at least one of the terms in the difference is finite; if both are, \( X \) is said to be integrable. For every \( p \) in \( \mathbb{R}^+ \), \( L^p(\mu) \) is defined as

\[
L^p(\mu) = \left\{ X \text{ measurable} \mid \int_{\Omega} |X|^p d\mu < +\infty \right\} \tag{B.8}
\]

Clearly, the elements in \( L^p(\mu) \) are defined up to a \( \mu \)-null set. In addition, \( L^\infty(\mu) \) is defined as the space of all measurable functions which are \( \mu \)-a.e. bounded. \( L^p(\mu) \) is a
complete metric vector space for the following distances

\[
\begin{align*}
\text{if } 0 < p < 1, & \quad d_p(X, Y) = \int_{\Omega} |X - Y|^p \, d\mu \\
\text{if } 1 \leq p < +\infty, & \quad d_p(X, Y) = \left( \int_{\Omega} |X - Y|^p \, d\mu \right)^{1/p} \\
\text{if } p = +\infty, & \quad d_p(X, Y) = \inf \left\{ c \in \mathbb{R} \mid \mu \left( \{ \omega \in \Omega \mid |X(\omega) - Y(\omega)| > c \} \right) = 0 \right\}
\end{align*}
\] (B.9)

For \( p \geq 1 \), \( d_p \) comes from a norm and thus \((L^p(\mu), d_p)\) is a real Banach space. For \( p = 2 \), \( d_2 \) comes from the scalar product

\[
<X, Y> = \int_{\Omega} X Y \, d\mu
\] (B.10)

Thus \((L^2(\mu), d_2)\) is a real Hilbert space. For every \( p \) in \([1, +\infty[\), every \( X \) in \( L^p(\mu) \), and every \( Y \) in \( L^{p-1}(\mu) \), Hölder’s (or, if \( p = 2 \), Cauchy-Schwarz’) inequality reads

\[
\int_{\Omega} |XY| \, d\mu \leq \left( \int_{\Omega} |X|^p \, d\mu \right)^{1/p} \left( \int_{\Omega} |Y|^{p-1} \, d\mu \right)^{p-1/p}
\] (B.11)

A useful integral is when \( \Omega = \{1, \ldots, k\} \), \( \Sigma = \{1\} \) and, for all \( S \) in \( \Sigma \), \( \mu(S) \) is the number of elements in \( S \). Let \( X \) be the function defined by \( X(i) = z_i \), for all \( i \) in \( \Omega \). Then

\[
\int_{\Omega} X \, d\mu = \sum_{i=1}^{k} z_i
\] (B.12)

Let \( \mu \) be a Lebesgue-Stieltjes measure on \( \mathbb{R} \) and let \( F \) be the corresponding d.f. given by (B.3). Then, if \( g \) is an integrable function on \((\mathbb{R}, \Delta_\mu, \mu)\), the integral \( \int_{\mathbb{R}} g(x) \mu(dx) \), also written \( \int_{\mathbb{R}} g(x) dF(x) \), is called a Lebesgue-Stieltjes integral.

References: P. Halmos [47], M. Rao [82].
B.3. Probability Theory

In probability theory, we follow the natural measure theoretic framework set forth by Andrei N. Kolmogorov in 1933 in the fundamental monograph referenced below. A probability space is a measure space \((\Omega, \Sigma, P)\) such that \(P(\Omega) = 1\). \(P\) is called a probability measure, the points in \(\Omega\) are called elementary events, and the sets in \(\Sigma\) are called events. If a property holds \(P\text{-a.e.}\), it is said to hold almost surely \((P\text{-a.s.})\) or with probability one. A random variable \((r.v.)\) \(X\) is a measurable real-valued function on \((\Omega, \Sigma, P)\). If \(\Delta\) denotes the Borel \(\sigma\)-algebra of \(\mathbb{R}\), the probability measure induced by the r.v. \(X\) on \(\Delta\) is defined by

\[
\forall B \in \Delta \quad P_X(B) = P\left(X^{-1}(B)\right) = P\left\{\omega \in \Omega \mid X(\omega) \in B\right\}
\]

The d.f. of \(X\) is defined by

\[
\forall c \in \mathbb{R} \quad F(c) = P\left(X^{-1}(-\infty, c]\right) = P\left\{\omega \in \Omega \mid X(\omega) < c\right\}
\]

For any two real numbers \(a \leq b\), \(F(b) - F(a) = P_X([a, b[\). Hence, \(F\) is the d.f. corresponding to the Lebesgue-Stieltjes measure \(P_X\). The expected value of \(X\), if it exists, is

\[
EX = \int_{\Omega} X(\omega)P(d\omega) = \int_{\mathbb{R}} zP_X(dz) = \int_{\mathbb{R}} zdF(z)
\]

For every \(p\) in \(\mathbb{R}^*_+\), \(E|X|^p\) is the \(p\)-th absolute moment of \(X\). Hence, \(L^p(P)\) is the space of all r.v.'s with finite \(p\)-th absolute moment. For every positive integer \(k\), the \(k\)-th moment of \(X\), if it exists, is defined as \(EX^k\). Let \(\lambda\) be the Lebesgue measure on \(\mathbb{R}\). A r.v. \(X\) is said to be absolutely continuous if there exists a nonnegative measurable function \(f\) on \(\mathbb{R}\), called density, such that

\[
\forall B \in \Delta \quad P\left\{\omega \in \Omega \mid X(\omega) \in B\right\} = \int_B f\,d\lambda
\]

More generally, if \(\lambda\) is the Lebesgue measure on \(\mathbb{R}^k\) and \(\Delta\) the Borel \(\sigma\)-algebra of \(\mathbb{R}^k\), the
random vector $X = (X_1, \ldots, X_k)$ has density $f$ on $\mathbb{R}^k$ if (B.16) holds.

References: A. Kolmogorov [59], M. Loève [69], [70], J. Neveu [72].

B.4. Stochastic Processes

A stochastic process is a quadruple $(\Omega, \Sigma, P, \{X_t\}_{t \in T})$, where $(\Omega, \Sigma, P)$ is a probability space and $\{X_t\}_{t \in T}$ a family of r.v.'s on this space, indexed on a nonempty set $T$. For every $\omega$ in $\Omega$, the function from $T$ into $\mathbb{R}$ which maps $t$ into $X_t(\omega)$ is called a path or a realization of the process. Let $I = \{t_1, \ldots, t_n\}$ be a finite subset of $T$. Then a finite dimensional d.f. of $\{X_t\}_{t \in T}$ is defined by

$$(\forall (c_1, \ldots, c_n) \in \mathbb{R}^n) \quad F_I(c_1, \ldots, c_n) = P\left(\bigcap_{i=1}^{n}\{\omega \in \Omega \mid X_{t_i}(\omega) < c_i\}\right) \quad (B.17)$$

A stochastic process is completely described by its finite dimensional d.f.'s. It is said to be strictly stationary if, for every finite subset $I = \{t_1, \ldots, t_n\}$ of $T$ and every $h$ in $\mathbb{R}$ such that $I + h = \{t_1 + h, \ldots, t_n + h\}$ remains a subset of $T$, we have

$$(\forall (c_1, \ldots, c_n) \in \mathbb{R}^n) \quad F_I(c_1, \ldots, c_n) = F_{I+h}(c_1, \ldots, c_n) \quad (B.18)$$

The r.v.'s $\{X_t\}_{t \in T}$ are said to be independent if for every finite subset $I = \{t_1, \ldots, t_n\}$ of $T$

$$(\forall (c_1, \ldots, c_n) \in \mathbb{R}^n) \quad F_I(c_1, \ldots, c_n) = \prod_{i=1}^{n} F_{t_i}(c_i) \quad (B.19)$$

where $F_{t_i}$ is the d.f. of $X_{t_i}$. If $T = \mathbb{Z}$, the stochastic process is said to be discrete. Two discrete stochastic processes $\{X_n\}_{n \in \mathbb{Z}}$ and $\{Y_n\}_{n \in \mathbb{Z}}$ defined on $(\Omega, \Sigma, P)$ are said to be equivalent if $(\forall n \in \mathbb{Z}) \quad X_n = Y_n$ a.s.

References: J. Doob [38], P. Lévy [67].