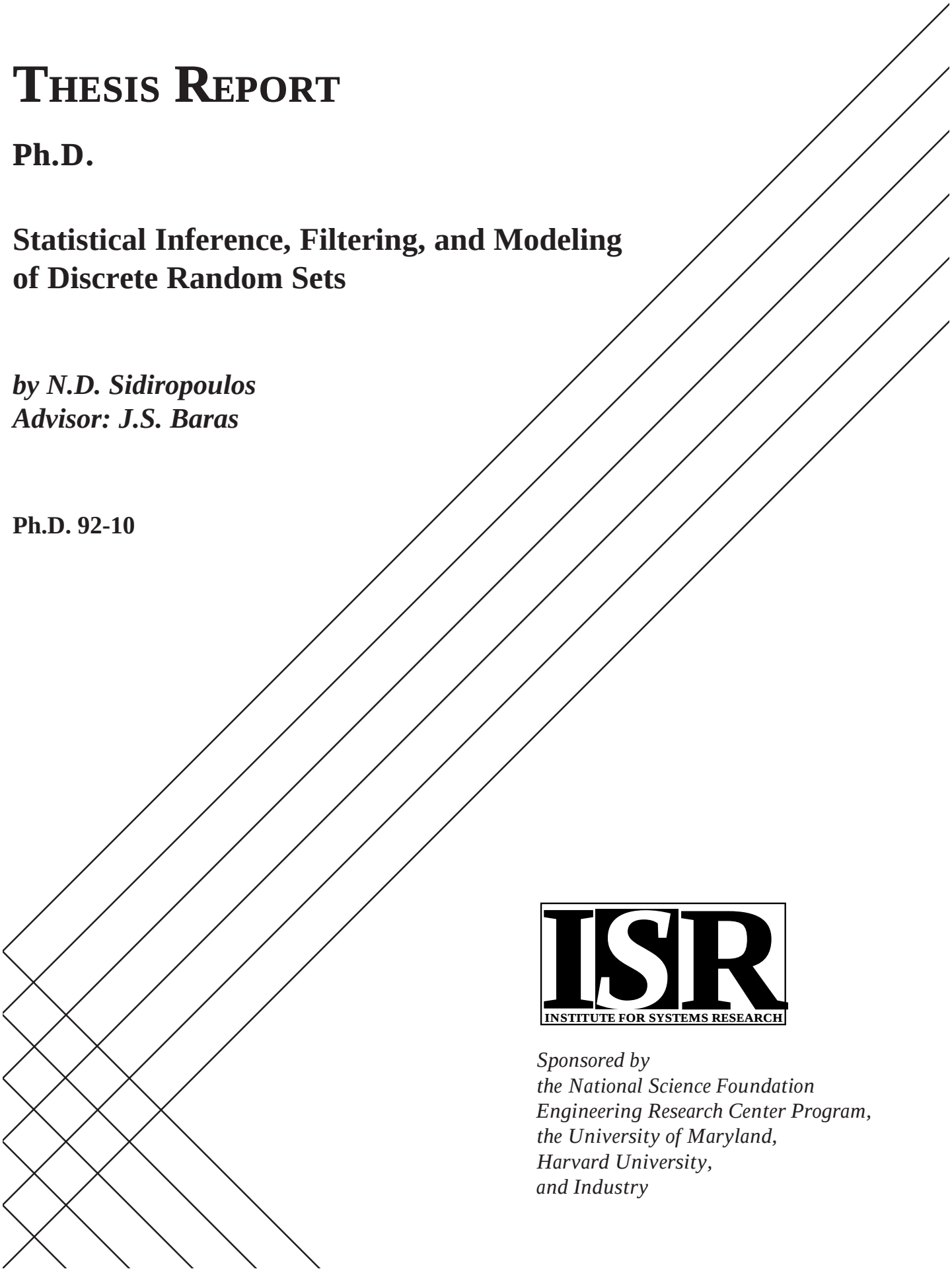
THESIS REPORT

Ph.D.

Statistical Inference, Filtering, and Modeling of Discrete Random Sets

by N.D. Sidiropoulos
Advisor: J.S. Baras

Ph.D. 92-10



Sponsored by
the National Science Foundation
Engineering Research Center Program,
the University of Maryland,
Harvard University,
and Industry

ABSTRACT

Title of Dissertation: Statistical Inference, Filtering, and Modeling of Discrete Random Sets

Nicholaos D. Sidiropoulos, Doctor of Philosophy, 1992

Dissertation directed by: Professor John Baras
Department of Electrical Engineering
Professor Carlos Berenstein
Department of Mathematics

The objective of this dissertation is the systematic study of several aspects of modeling, statistical inference, and filtering of "random" binary digital images, or, uniformly bounded discrete random sets. This study consists of two interleaved parts. In the first part, we consider some important aspects of a theory of uniformly bounded discrete random sets. The fundamental result is a strengthened version of a backbone theorem of continuous-domain random set theory, namely the uniqueness theorem of Choquet-Kendall-Matheron, for the case of uniformly bounded discrete random sets.

The vehicle through which much of the discussion is carried out is a "special" discrete random set model, the discrete radial Boolean random set, which is closely related to the theory of Morphological shape-size distributions. The continuous-domain Boolean random set has been successfully used in a variety of applications. A good portion of the first part of this dissertation is devoted to the statistical inference of the discrete radial Boolean random set. We consider three problems: parameter estimation, binary hypothesis testing, and classification of "random"

known shapes in Boolean clutter. The tools come mainly from the area of Morphological shape analysis. The focus is on inference procedures which are both computationally efficient, and statistically sound.

In the second part, we consider the problem of estimating realizations of uniformly bounded discrete random sets, distorted by a degradation process which can be described by a union/intersection noise model. Two different optimal filtering approaches are considered. The first involves a class of filters which arises quite naturally from the set-theoretic analysis of optimal filters. We call this the class of *mask filters.* The second approach deals with optimal Morphological filters. First, we provide some fresh statistical insight into certain "folk theorems" of Morphological filtering. We do so by exploiting the uniformly bounded discrete random set formulation of the filtering problem. Then we show that, by using an appropriate (under a given degradation model) expansion of the optimal filter, we can obtain "universal" characterizations of optimality, in terms of the fundamental functionals of random set theory, namely the generating functionals of the signal and the noise.

Statistical Inference, Filtering, and Modeling of Discrete Random Sets

by

Nicholaos D. Sidiropoulos

Dissertation submitted to the Faculty of The Graduate School of The University of Maryland in partial fulfillment of the requirements for the degree of Doctor of Philosophy
1992

Dissertation Committee:

Professor John Baras, Chairman/Advisor
Professor Carlos Berenstein, Advisor
Professor P. Krishnaprasad
Professor Ramalingam Chellappa
Professor Ben Kedem
Associate Professor Prakash Narayan

DEDICATION

To my beloved family, and the days to come.

ACKNOWLEDGMENTS

I am indebted to both my research advisors, Dr. John Baras, and Dr. Carlos Berenstein, for their continuous encouragement throughout the course of this work. They have given me ample freedom of choice of research topics and methods. They believed in my work, and have stood by me during difficult times. The explosive drive and motivation of Dr. Baras, and the informal, amiable style of Dr. Berenstein have blended well with me. My tenure with them has been an experience at many different levels.

I chose my research area after studying a survey paper by Dr. John Goutsias, of Johns Hopkins University. He has been an invaluable source of good advice, references, and moral support. Over the years, we have engaged in many stimulating and enthusiastic discussions, which helped me shape my ideas and focus my research. He has also been one of the very first people to read early drafts of my work. His help and friendship are greatly appreciated.

A number of other people have contributed, one way or another, to the technical aspects of this work. It is my pleasure to acknowledge the help of Dr. Andrian Papamarcou, Dr. Prakash Narayan, Dr. Armand Makowski, and my

good friend and colleague Dr. Leandros Tassioulas. I would also like to thank the members of my committee for their willingness to serve.

I gratefully acknowledge the financial support provided by the National Science Foundation, under grant NSFD CDR 8803012.

I have shared a lot of the joy, and the frustration, of life as a graduate student with my good friends Leonidas Constantinou, and Olga Poulida. They have helped me maintain my perspective, and keep my feet on the ground. Their friendship has been invaluable.

Most of all, I would like to express my deepest gratitude to my family; during hard times, they often had to overcome their sorrow to support me emotionally. My father has encouraged me rigorously to pursue Doctoral study. His conviction about the importance of education has had a fundamental impact on my life.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

NOTATION

RS	Acronym for Random Set
DRS	Acronym for (uniformly bounded) Discrete Random Set
$\mathbf{Z}^2$	The planar integer lattice
$B \subset \mathbf{Z}^2$	The "base frame", i.e. the finite lattice on which DRS's are defined
$\Sigma(B)$	The power set of B
$\Sigma(\Sigma(B))$	The power set of the power set of B
DBRS	Acronym for Discrete Boolean Random Set
DRBRS	Acronym for Discrete Radial Boolean Random Set
$Q_X(\cdot)$	Generating functional of DRS X
$T_X(\cdot)$	Capacity functional of DRS X
$X \oplus H$	Minkowski sum of the sets X, H
$X \ominus H$	Minkowski subtraction of the set H from the set X
H^s	The reflection of set H with respect to the origin
$X \oplus H^s$	Dilation of set X by structuring element H
$X \ominus H^s$	Erosion of set X by structuring element H

$X \circ H$	Opening of set X by structuring element H
$X \bullet H$	Closing of set X by structuring element H
X^c	The complement of set X with respect to the unit element of the lattice
MSF	Morphological Skeleton Function
ASF	Alternating Sequential Filter
Ker(Ψ)	The Morphological erosion (dilation) kernel of filter Ψ
Bas(Ψ)	The Morphological erosion (dilation) basis of filter Ψ
E	The expectation operator
OP	Orthogonality Principle
iid	independent and identically distributed
pmf	probability mass function
a.s.	almost surely
ML	Maximum Likelihood
MLE	Maximum Likelihood Estimator
MVUE	Minimum Variance Unbiased Estimator
MAP	Maximum A Posteriori
BSC	Binary Symmetric Channel
PPP	Poisson Point Process
SPPP	Stationary Poisson Point Process

CHAPTER
ONE

INTRODUCTION

1.1 Background and Motivation

In most modern image processing and analysis systems, low-level pictorial information is represented by a set of samples, or measurements, of a function in 2-dimensional space, over a finite lattice, B, of sampling sites. The resulting collection of measurements is a digital image. More often than not, the range ("alphabet") of sample values is finite, and it can be identified with a finite subset of the natural numbers, say $\mathcal{A} = \{0, 1, \cdots, N-1\}$. Every finite-alphabet digital image can be thought of as a subset of $B \times \mathcal{A}$, and it can be processed and analyzed using set-theoretic tools. Binary digital images are a special case, one of considerable interest in its own right. Every binary digital image can be thought of as a subset of the sampling lattice, B.

Our objective is to study "random" binary digital images. Such a study is important for two reasons. First, random binary digital images arise naturally in various important applications (e.g., modeling of spatial structures,

and pictorial pattern recognition). Second, the study of random binary digital images serves as a stepping stone towards the study of more complex random finite-alphabet digital images. The results easily generalize to the finite-alphabet case, by considering finite-alphabet digital images as subsets of $B \times \mathcal{A}$.

Random set theory is a fusion of ideas originating along the fuzzy lines between *Geometrical Probability*, *Stereology*, and *Stochastic Geometry* [65]. It has been developed independently by Kendall [34], and Matheron [45] in the early seventies, based on earlier results by Choquet [9]. It deals with the probabilistic description of "binary images". The theory of *Mathematical Morphology* has been conceived (and, to a great extent, developed), by Matheron [45], Serra [57, 21], and their collaborators. It is concerned with the quantitative analysis of shape, with an emphasis on geometric structure. The two theories have evolved interactively. As a result, random set theory is appropriate when the objective is the quantitative analysis of geometric features of random "objects" in multi-dimensional spaces, be it binary images in two-dimensional space, solid objects in three-dimensional space, or, more generally, "concepts" in more abstract spaces.

Almost a quarter century after its inception, random set theory still remains pretty much inaccessible to the vast majority of statisticians and engineers alike. This can be largely attributed to three key factors: (1) Early treatises on the subject were highly technical, sometimes even intimidating; (2) Most people feel uncomfortable when the "data" are sets; and (3) Practitioners never fully real-

ized the potential of the theory and its applications. During all these years, the theory of random sets (especially those in Euclidean spaces, henceforth referred to as continuous-domain random sets) has matured considerably, as a result of the work of a few dedicated individuals, and research groups. The basic concepts which underlie the theory are well understood, and the counterparts of many classical constructions and results of probability theory (e.g., ergodic theorems, laws of large numbers, etc.) have been established for random sets. However, there exist fundamental differences between random variables and random sets; the transition from *points* to *sets*, in n-dimensional space, forces a radical change in methodology. The sheer dimensionality of the problem makes traditional tools obsolete. In particular, the lack of a total set ordering makes life a lot more difficult. Random variables are completely characterized by their cumulative distribution function (cdf). Although a counterpart of the cdf for random sets does exist, it is notoriously difficult to work with. This is directly related to the lack of a total set ordering. Any meaningful notion of a density is lost within the technicalities. The situation calls for a change in perspective.

This dissertation is concerned with several aspects of the statistical inference, filtering, and modeling of a special class of random sets, namely *uniformly bounded discrete random sets.* Discrete random sets are defined on a (regular or irregular) lattice of points. The term *uniformly bounded* is used here to convey that this lattice consists of *finitely* many points. The class of uniformly bounded discrete random sets is a restriction of the class of continuous-domain random

sets. Although continuous-domain random set theory is fairly mature, the special class of uniformly bounded discrete random sets has not yet received due attention. This class is "special" for many reasons; by far the most important one is that, in practice, we almost always deal with finitely many data points. The vast majority of image processing, analysis, and synthesis is nowadays done with digital computers, which operate on finite, discrete image representations. Given the rate of advances in digital computing, this situation is not likely to change in the foreseeable future. Hence, from an applications point of view, a theory of uniformly bounded discrete random sets is "essentially" sufficient. This should be taken with a grain of salt, for there exist certain modeling applications which demand a continuous-domain approach. The second fundamental feature of uniformly bounded discrete random sets is that their construction involves considerably fewer technicalities than their continuous-domain counterparts, and this should make the former much more appealing to the engineering community. In particular, it means that the practitioner can focus on *his* problems (e.g., filtering, statistical analysis, etc.), rather than worry about the problems *of the model.* For example, in the case of uniformly bounded discrete random sets, we have the standard notion of probability mass, whereas no clear-cut notion of probability density exists for continuous-domain random sets. Finally, our experience has been that working with finitely many data points allows us to prove rather strong results, which typically do not work out in the case of infinitely many data points.

During the better part of this century, most breakthroughs in signal processing have been made possible by clever use of statistics. Examples are abundant: Kalman filtering, Wiener filtering, Viterbi decoding, just to name a few. It was the artful blend of signals and systems with statistical models and methods that propelled modern era signal processing into its present state. Originally, Mathematical Morphology had been primarily targeted to the statistical analysis of "images", but later on the statistical aspects were pretty much neglected. *Morphological filtering*, a successful and popular branch of Mathematical Morphology, has played a pivotal role in increasing the visibility of the theory. However, most of the research effort has been focused on the syntactical properties of Morphological filters. We maintain that, *joint optimization of both syntactical and statistical properties of Morphological filters is the most promising design approach.* The statistical analysis of these filters is complicated, because they are inherently nonlinear. However, we shall see that such an analysis is possible within the framework of the proposed theory of uniformly bounded discrete random sets.

The practice of random set analysis and synthesis has been hampered by a "handicap" of random set theory: the lack of simple, tractable, realistic models which can encode information about highly correlated spatial structures, such as man-made objects. This handicap can be alleviated, up to a certain extent, if the object in question is viewed from relatively few aspects, i.e. it can only assume a relatively small number of "realizations", in which case it can be

compactly specified by using a probability mass on the space of realizations. On the other hand, there exist good models for spatially "white" structures, namely the Boolean model and its derivatives.

Statistical inference techniques similar to *Maximum Likelihood* (ML), or *Maximum A Posteriori* (MAP), are practically non-existent in continuous-domain random set theory [23]. This is due to inherent analytical difficulties, and the fact that, despite its usefulness, continuous-domain random set theory falls short of providing the necessary tools needed to construct efficient inference procedures. The richness of the event space poses another fundamental constraint: it is very difficult (if not impossible) to derive the measure on the event space, given a constructive random set specification. In contrast, we will see that ML and MAP inference is often possible for uniformly bounded discrete random sets.

We have undertaken this study with three goals in mind. First, to demonstrate that a uniformly bounded discrete random set approach to digital image analysis and synthesis is not only feasible, but also rewarding, and meaningful. Second, to develop some aspects of *Stochastic Morphology*, and show that certain tools of Morphological shape analysis fit naturally in the stochastic setting. Third to provide sufficient motivation for people working in image processing, filtering, modeling, and estimation, pictorial pattern recognition, inspection, and classification, and other related areas, to seriously consider using a discrete random set-theoretic approach. We believe that years of exciting research, and important discoveries still lie ahead.

1.2 Organization

This dissertation addresses some issues relevant to modeling, statistical inference, and filtering of uniformly bounded discrete random sets. The outlook is as follows. The rest of this chapter is devoted to continuous-domain random sets. It provides some necessary background, as well as a brief literature survey. The purpose is to familiarize the reader with the fundamentals of random set theory, provide an account of various aspects of random set modeling and inference, and offer an up-to-date reference list. The Boolean random set model is treated in some detail, in part to motivate the analogy with the proposed discrete-domain counterpart.

In chapter 2, we develop an axiomatic formulation of uniformly bounded discrete random sets. We show how the fundamental theorem of Choquet - Kendall - Matheron can be significantly strengthened in the case of uniformly bounded discrete random sets. Next, we develop a discrete-case analog of the Boolean model. Special attention is paid to a restricted version of this model, namely the discrete radial Boolean random set, which is closely related to the Morphological analysis of images, and the associated theory of shape-size distributions. This model is central to our work. Its probabilistic specification, both in terms of its "hitting functional", and in terms of its probability mass function, is fully investigated. Some aspects of divisibility of uniformly bounded discrete random sets are also considered.

Chapter 3 deals with the statistical inference of the discrete radial Boolean random set model. Two specific problems are addressed: parameter estimation, and hypothesis testing. In principle, the problem of testing hypotheses can be solved by brute force, by means of our strengthened version of the Choquet - Kendall - Matheron theorem. In practice, this is usually impossible, because of combinatorial explosion. Parameter estimation is - naturally - more complicated. Our approach is oriented towards obtaining bounds on the Maximum Likelihood estimate. Specifically, since direct ML estimation seems impossible, our aim is to develop computationally efficient and statistically sound procedures for the determination of a "feasible region"; i.e., a "small" subset of parameter values which is guaranteed to contain the Maximum Likelihood estimate(s) of the parameter. We make extensive use of the strengthened Choquet - Kendall - Matheron theorem, as well as certain tools of Morphological shape analysis. In particular, we employ the Morphological Skeleton Transform *as an estimator* of the marked point process which underlies the discrete radial Boolean random set. This mode of use of the Morphological Skeleton Transform is entirely new. In the past, it has been extensively used for shape description, coding, and recognition. This "random skeleton" idea can be used both for hypothesis testing, and for parameter estimation purposes. Another problem which is considered is that of binary hypothesis testing between "sparse" and "dense" models, and the feasibility of making computationally cheap Maximum Likelihood decisions. This problem has applications in the automated inspection of cell populations,

particularly in diagnostic medical imaging.

In chapter 4, we consider the problem of estimating realizations of uniformly bounded discrete random sets, distorted by a degradation process which can be described by a union/intersection noise model. This kind of degradation is rather general, and it can model both "impulsive" and "smooth", uncorrelated or colored noise. We start by formulating the optimal filtering problem for the case of uniformly bounded discrete random sets. Two distinct filtering approaches are pursued. The first involves a class of filters which arises quite naturally from the set-theoretic analysis of optimal filters. We call this the class of *mask filters.* We consider both fixed and adaptive mask filters, and derive explicit formulas for the optimal mask filter under quite general assumptions on the signal and the degradation process.

The second approach is more appropriate for images which exhibit a stationary statistical behavior, and it involves a class of Morphological filters. First we provide some theoretical justification for the popularity of certain Morphological filtering schemes. In particular, we prove that if the signal is "smooth", then these schemes are optimal (in the sense of providing the MAP estimate of the signal) under a reasonable worst-case statistical scenario. These results offer some fresh statistical insight into Morphological filtering. They do so by exploiting the uniformly bounded discrete random set formulation of the filtering problem. Finally, we show that, by using an appropriate (under a given degradation model) expansion of the optimal filter, we can obtain universal char-

acterizations of optimality which do not rely on strong assumptions regarding the spatial interaction of geometrical primitives of the signal and the noise. This approach corresponds to a somewhat counter-intuitive use of fundamental Morphological operators; however, it is exactly this mode of use that enables us to arrive at characterizations of optimality in terms of the fundamental functionals of random set theory, namely the generating functionals of the signal and the noise.

Chapter 5 represents a preliminary attempt to apply the framework of Probabilistic Classification Trees to a specific known "random" shape in clutter recognition problem. This chapter is not exhaustive, and it aims at bringing together several key ideas (namely, Kendall's "trapping system", discrete random sets, Mathematical Morphology, and hit-or-miss probing of images by sets), to develop adaptive probabilistic probing, a flexible sequential procedure for the efficient classification of binary image objects in a heavily cluttered and noisy environment. The method has a distinct geometrical flavor, and it holds considerable promise for drastically reducing the runtime complexity of the decision rule. Here, we confine ourselves to developing a suitable formalism, highlighting the benefits, as well as the drawbacks, of the approach, and presenting a concrete design procedure for probing known "random" shapes in Boolean clutter.

Chapter 6 contains some concluding thoughts, along with future research directions. Finally, in chapter 7, we provide (in the form of an appendix) a self-contained summary of basic Maximum Likelihood parameter estimation, as

it pertains to our work. This (standard) material is used freely throughout the rest of this dissertation.

Some notes on style are in order. Figures and tables are grouped together at the end of each chapter. Results are labeled as follows. A *theorem* is an elegant, powerful, and, perhaps, surprising result. A *proposition* is a result which has been planned, and worked out. It is important in its own right, possibly even more so than a theorem, but it is not necessarily elegant, and its proof may involve considerable technicalities. A *lemma* is something that is used to prove something else. A *corollary* is something that comes "for free" by employing earlier results.

1.3 Continuous-domain random sets

There exists a sizable body of literature which is concerned with various aspects of continuous-domain random set theory. In this section, we formally define random sets, in the sense of Matheron [44, 45], and offer a brief survey of those parts of the literature which are directly relevant to the purposes of this dissertation. While the emphasis is on compactness and clarity of presentation, we have attempted to preserve rigor. The technical details of this section are not a prerequisite for the rest of this work. This survey is by no means comprehensive; the interested reader is referred to the original references [44, 45, 34], survey papers [11, 12, 56], the book of Stoyan, Kendall, and Mecke [65], and a recent tutorial paper [23], for a thorough exposition.

Let $\mathcal{O}, \mathcal{F}, \mathcal{K}$ denote the spaces of open, closed and compact subsets of $\mathbf{R}^n$, for some n. Consider the following collections of sets in $\mathcal{F}$

$$F_G = \{F \in \mathcal{F} \mid F \cap G \neq \emptyset\}, \quad G \in \mathcal{O}$$

$$F^K = \{F \in \mathcal{F} \mid F \cap K = \emptyset\}, \quad K \in \mathcal{K}$$

The former collection comprises of all sets in $\mathcal{F}$ which "hit" an open set G, while the latter collection comprises of all sets in $\mathcal{F}$ which "miss" a compact set K. The collection of sets $\{F_G, G \in \mathcal{O}\}$ and $\{F^K, K \in \mathcal{K}\}$ generates a Topology, $T(\mathcal{F})$, on $\mathcal{F}$. This is known as the **hit-or-miss topology**, and it allows the study of convergence and continuity in $\mathcal{F}$ [44, 45, 57]. By taking countable unions and intersections of the open sets of the topological space $(\mathcal{F}, T(\mathcal{F}))$, a σ-algebra, $\Sigma(\mathcal{F})$, is generated on $\mathcal{F}$.

Definition 1.3.1 *A random closed set, or, for brevity,* **Random Set (RS)**, X, *is a measurable mapping of a probability space* $(\Omega, \Sigma(\Omega), P)$ *into the measurable space* $(\mathcal{F}, \Sigma(\mathcal{F}))$.

By definition, a RS X induces a unique probability measure, P_X, on $\Sigma(\mathcal{F})$.

Definition 1.3.2 *The* **capacity functional**, T_X, *of a RS* X, *is defined as*

$$T_X(K) = P_X(X \in F_K) = P_X(X \cap K \neq \emptyset), \quad K \in \mathcal{K}$$

The capacity functional contains all the information about the RS X. This is the subject of the following fundamental theorem, originally due to Choquet [9],

and independently introduced in the context of random set theory by Kendall [34] and Matheron [44, 45].

Theorem 1.1 *[9, 34, 45, Choquet-Kendall-Matheron] Given $T_X(K)$, $\forall K \in \mathcal{K}$, there exists a unique probability measure, P_X, on $\Sigma(\mathcal{F})$, such that $P_X(X \in F_K) = T_X(K)$, $\forall K \in \mathcal{K}$*

The proof of this theorem is rather delicate. Choquet proved it in an abstract setting in his seminal work [9]. Matheron proved it within the context of his theory of shape-size distributions [44, 45]. Kendall was specifically interested in a statistical theory of shape, and has contributed an independent construction of random sets, along with a proof of the Choquet theorem as it pertains to random sets, based on a simple, but powerful, "trapping system" idea.

Let us now formally define an important class of random set models, the so-called germ-grain random sets, and a very special instance of this class, namely the Boolean random set model. For the sake of completeness, we reproduce some relevant definitions, taken mainly from [65].

Definition 1.3.3 *A* **Point Process**, Φ, *on* $\mathbf{R}^n$, *is a measurable mapping of a probability space $(\Omega, \Sigma(\Omega), P)$ into a measurable space $(N, \Sigma(N))$, where N is the family of all subsets ϕ, of $\mathbf{R}^n$, satisfying the following two regularity conditions:*

(i) ϕ is locally finite (each bounded subset of $\mathbf{R}^n$ must contain only a finite number of points in ϕ).

(ii) ϕ is simple (no two points in ϕ coincide).

Here $\Sigma(N)$ is the smallest σ-algebra on N which makes measurable all mappings $\phi \mapsto |\phi \cap B|$, where B runs through the bounded Borel sets of $\mathbf{R}^n$.

Informally, a Point Process on $\mathbf{R}^n$ can be thought of as a random pattern of points, scattered over $\mathbf{R}^n$. Consider the measurable space $(\mathbf{R}^n, \mathcal{B}(\mathbf{R}^n))$, where $\mathcal{B}(\mathbf{R}^n)$ denotes the Borel σ-field of $\mathbf{R}^n$, and a measure Λ on $\mathcal{B}(\mathbf{R}^n)$ such that for all bounded $B \in \mathcal{B}(\mathbf{R}^n)$ the measure of B, $\Lambda(B)$, is finite. The measure Λ is referred to as a **Radon measure**. If, in addition, Λ gives zero mass to singletons (like the Lebesgue measure), then Λ is called a **diffuse Radon measure**.

Definition 1.3.4 *A* **Poisson Point Process (PPP)**, Φ, *on* $\mathbf{R}^n$, *with diffuse Radon measure (or mean measure)* Λ *on* $\mathcal{B}(\mathbf{R}^n)$ *is a point process which is completely specified by the following two properties:*

(i) Poisson distribution of point counts; the number of points in a bounded set $B \in \mathcal{B}(\mathbf{R}^n)$ has a Poisson distribution with mean $\Lambda(B)$

$$P(\Phi(B) = m) = \frac{(\Lambda(B))^m e^{-\Lambda(B)}}{m!}, \quad m = 0, 1, 2, \ldots$$

(ii) Independent Scattering; the number of points in k disjoint Borel sets form k independent random variables.

If the diffuse Radon measure, Λ, is absolutely continuous (admits a density) with respect to the Lebesgue measure, then it can be written as

$$\Lambda(B) = \int_B \lambda(z)dz, \quad \forall B \in \mathcal{B}(\mathbf{R}^n)$$

The density $\lambda(z) \geq 0, \forall z \in \mathbf{R}^n$, is called the *intensity* of the general PPP. Henceforth, we make the assumption that Λ *is* absolutely continuous with respect to the Lebesgue measure (and, therefore, the intensity $\lambda(z), z \in \mathbf{R}^n$, exists). In case $\lambda(z) = \lambda = const$, $\forall z \in \mathbf{R}^n$, we have the special case of a *Stationary PPP (SPPP).*

Definition 1.3.5 *The RS X defined by*

$$X = \bigcup_{i=1,2,\ldots} G_i \oplus \{p_i\}$$

where $P = \{p_1, p_2, \cdots\}$ is a Point Process and $\{G_1, G_2, \cdots\}$ is a set of non-empty, bounded RS's, is a **germ-grain RS**. *The points p_i, $i = 1, 2, \ldots$, are the* **germs**, *whereas the RS's G_i, $i = 1, 2, \ldots$, are the* **primary grains** *of the germ-grain RS X (the symbol $\oplus$ stands for Minkowski set addition; here, since one of its arguments is a singleton, it amounts to a translation of the primary grain by the vector corresponding to this singleton).*

Definition 1.3.6 *Let Φ be a PPP with intensity $\lambda(z) > 0$, $\forall z \in \mathbf{R}^n$, and let $\{G_1, G_2, \cdots\}$ be a set of non-empty, bounded iid RS's, independent of Φ, and characterized by the capacity functional T_G. If*

$$E\left[\int_{\mathbf{R}^n} 1_{K \oplus G_1^s}(z)\lambda(z)dz\right] < \infty, \ \forall K \in \mathcal{K}$$

then the resulting germ-grain RS X will be called a **Boolean RS**. *Here, 1 stands for the indicator function, G_1^s stands for the reflection of G_1 with respect to the origin, and E stands for expectation.*

Informally, a Boolean random set is constructed by centering a simple random shape (set), such as a disc of random size, at each point of a Poisson field of points in the plane, and then taking the union of the resulting sets. Random shapes centered at different points of the Poisson field are assumed to be independent and statistically equivalent. In a sense, it can be argued that the Boolean model is a natural generalization of "white noise", which conveniently incorporates a random shape component.

The Boolean RS is arguably the most important random set model to date. It has received considerable attention in the literature (for example, see [56, 13, 11, 20, 4, 57, 21, 65], and references therein), both in terms of theory, and in terms of practice. Typical applications include, but are not limited to: random clumping of dust, or powder particles; modeling of geological structures, bomb fields, patterns in photographic emulsion, colloids in gel form, and structural inhomogeneities in amorphous matter [65, p.68]; tumor growth [11], and the spatial pattern of heather [13]. Other potential applications include particle counting and size analysis in images of cell cultures, and modeling of clutter in infrared imaging.

The capacity functional of a Boolean RS has a particularly nice form. It is given by

$$T_X(K) = 1 - exp\left\{-E\left[\int_{\mathbf{R}^n} 1_{K\oplus G_1^s}(z)\lambda(z)dz\right]\right\}$$

If the Poisson process is stationary, i.e., if $\lambda(z) = \lambda,\ \forall z \in \mathbf{R}^n$, then we have the special case of a stationary Boolean random set, whose capacity functional has

the following strikingly simple form

$$T_X(K) = 1 - exp\{-E[\lambda\, m_n(K \oplus G_1^s)]\}$$

where m_n stands for Lebesgue measure in $\mathbf{R}^n$.

1.4 Literature review

The existing literature on random sets[1] can be roughly separated in two overlapping parts. The first part comprises of those references which mainly address fundamental aspects of random set theory; i.e., the Choquet theorem, laws of large numbers, central limit theorem, infinite divisibility, semi-Markovianity, ergodic properties, etc.; e.g., [9, 44, 45, 34, 65, 12, 56, 57, 21, 67, 10, 23], and some others which are halfway between random set theory and Geometrical Probability; e.g., [37, 36, 38, 35, 32, 33, 29, 5]. The second part consists of those references which mainly deal with random set models and their statistical inference [4, 11, 13, 20, 56, 57, 21, 54]. Statistical inference includes model fitting (parameter estimation), hypothesis testing (classification), and signal estimation (filtering). The second part of the literature is almost entirely devoted to the stationary Boolean model. Parameter estimation is the central theme, and hypothesis testing and signal estimation receive hardly any attention.

[1] Point processes can be thought of as a special class of random sets. However, point process theory is really a distinct, mature subject, whose scope and methods diverge significantly from those of random set theory. Thus, we do not include point process theory in this survey. See [49, 50, 51, 52, 31, 63, 65] for some interesting related results.

1.4.1 Fundamental aspects of random set theory

This part of the theory is adequately covered in a recent tutorial paper [23], and in some earlier accounts (e.g. [56, 11]). We therefore confine ourselves to a brief review of the most significant results.

Strong limit theorems for random sets can be found in [3, 10]; also in [70, 71, 67, 66, 69, 68], among others. A "central limit theorem" for random sets can be found in [57], also in [56]. In loose terms, just as a Gaussian random variable appears as the limit of an average of independent random variables, the limit of an infinite union of independent random sets is a Boolean random set. This result by itself makes a strong case for the plausibility of the Boolean assumption. Infinite divisibility is another important property. A RS X is said to be infinitely divisible if, for any $n > 0$, X can be expressed as the union of n iid RS's X_i, $i = 1, \cdots, n$. A RS X is infinitely divisible iff its capacity functional is of a certain exponential type [45]. Thus, the union of two or more statistically independent infinitely divisible RS's is another infinitely divisible RS whose capacity functional is of the same exponential type as the capacity functional of its components. This is a powerful result, and it is crucial in proving certain limit theorems. The Boolean RS is infinitely divisible, and, in fact, this property essentially implies the "central limit theorem" [56]. Furthermore, it can be shown that the Boolean RS satisfies certain ergodic properties; i.e., spatial averages over one realization tend towards the corresponding expectations [45,

56].

Semi-Markovianity is the random set analog of the Markovian property [73]. For set-valued random variables there exists no natural total ordering, and thus no natural notion of causality. This problem can be worked around by introducing the class of semi-Markov random sets. In loose terms, a RS X is semi-Markov iff given that X misses a compact set K which separates two compact sets K_1, K_2, the random sets $X \cap K_1$, $X \cap K_2$ are conditionally independent [23].

Ever since his original contribution to the theory of random sets, Kendall has been pursuing an independent line of research, evolving around the notion of a shape-space, and an associated shape density [35, 36, 38]. The result of this investigation is an elegant statistical theory of shape [37]. While closely related to random set theory, Kendall's theory has a different flavor. It is largely devoted to the study of randomly generated triangles and convex polygons.

1.4.2 Statistical Inference

So far, the typical approach in the literature has been to focus on a particular set of properties and choose the model parameters to match the analytically obtained property values to the empirical estimates of these values. The empirical estimates are usually formed by taking spatial averages (it is assumed that the model satisfies some ergodic property), and the matching is done by some sort of ordinary or generalized least squares fit. Serra [56] describes such an approach for the case of a stationary Boolean model, with Poisson polyhedra as primary

grains. The analytical properties used are the miss probabilities for a few test sets, and the geometric covariogram for the primary grains. Diggle [13] discusses a similar approach for the case of a stationary Boolean RS with circular primary grains of random radius. The properties used are the distribution function of the distance from an arbitrary point in the plane to the nearest point occupied by the Boolean model, and the covariance function between points in the plane. Cressie and Laslett [11] consider the case of a stationary Boolean model with a.s. convex and isotropic primary grains. The properties used are again the miss probabilities for a few test sets. Schmitt [54] proposes a practically tractable formula for estimating the intensity of a stationary Boolean model, assuming only a.s. boundedness of the primary grain.

These approaches can be liberally described as "methods of moments". The matching of theoretical moments to sample values to estimate parameters is a well known trick of the trade, but it is only used when standard approaches, such as ML estimation, are intractable. The reason is that, in general, the statistical properties of estimators which are based on moment matching are unknown. In particular, standard measures of estimator performance, such as bias, efficiency, and consistency, are often impossible to obtain analytically.

A different approach has been proposed by Dupac [20]. He contemplated the use of what he calls "circular clumps", those primary grains whose borders are not covered (and so not deformed) by other primary grains. There exist two problems with this approach. First, the circular clumps are clearly spatially

dependent, and this dependency can not be easily incorporated into an inference procedure. Therefore, one is essentially forced to assume that the circular clumps are (at least approximately) independent. Second, since the smaller grains are more likely to survive without being hit by other grains, any size distribution estimate which is only based on circular clumps is biased. In much the same way, intensity estimates which are based on circular clump counts will be biased, especially if circular clumps do not occur "often enough", i.e. the degree of clumping in the data is high. These effects can be compensated, up to a certain extend, by incorporating the "grain survival probability" into the computations, but this cannot eliminate the problem, because of the inability to characterize the dependence between the clumps. Ayala et al., [4] pursued Dupac's idea for the case of a stationary Poisson germ process and circular primary grains with uniformly distributed radii. They proposed an almost-ML estimation procedure, which assumes that the clumps are independent. Their simulations indicate that there exists a consistent bias in the estimates, which can be partially attributed to the dependence between the circular clumps.

CHAPTER
TWO

UNIFORMLY BOUNDED DISCRETE RANDOM SETS

2.1 Introduction

The transition from continuous-domain random sets to discrete-domain random sets is a troublesome one [57]. In practice, one usually deals with finitely many samples of a portion of a binary image which is contained within a fixed window (i.e., *a realization of a uniformly bounded discrete random set*). Most automated image analysis systems operate on finite spaces. Given the current rate of advances in digital computers, it seems fair to say that this trend will only continue to grow in the foreseeable future.

There exist essentially three ways to introduce discrete random sets. Before we present our choice, let us discuss the alternatives. The first is through sampling (discretization) of continuous-domain random sets [45, 57]. This poses several technical problems. For example, sampling with a regular lattice introduces lattice-dependent artifacts (e.g., the sampling does not preserve the Euler-Poincare characteristic, [56]). In general, it is impossible to say anything

about several important features (e.g. convexity) of the underlying continuous-domain structure, based solely on its sampled realizations. The second approach is based on the theory of random fields [49]. A special class of random fields, namely the class of Markov random fields (MRF's), has been successfully used to model random texture. However, these models can not easily describe complicated geometrical structure, i.e. they generally fail to capture the morphological aspects of image data[1]. Random set theory, on the other hand, is closely related to Mathematical Morphology, a nonlinear image algebra which specifically addresses the problem of quantitative shape description. As a result of this connection, random set theory provides a unified framework which allows the modeling of both morphological ("syntactical") *and* statistical characteristics of images. Thus, the need to further develop the special class of discrete random sets along the lines of the more general continuous-domain random set theory becomes apparent.

We have chosen to define uniformly bounded discrete random sets directly on a finite lattice, and base subsequent developments on this definition. This axiomatic approach has many advantages. It avoids technicalities, and enables us to focus on problems which are important in practice. In particular, it allows us to talk about probability mass and ML, MAP inference. Certain important results of random set theory can be significantly strengthened in the case of

[1]Some preliminary work on modeling simple geometrical structure using MRF's has been recently reported in [8]. The basic idea is to incorporate geometrical constraints into the MRF clique structure in a suitable approximate sense. This idea may prove promising.

uniformly bounded discrete random sets. Of course, these benefits come at a certain price. In this case, we ignore the details of the underlying continuous physical structure which fall beneath our resolution. However, at any rate, this loss of detail is forced upon us by the limitations of the digital imaging system; we might as well accept it, and live with it. Our findings suggest that, from the point of view of applications, such an axiomatic approach is significantly more flexible[2]

Our programme is to develop an axiomatic formulation of uniformly bounded discrete random sets, strengthen a general characterization theorem of Choquet - Kendall - Matheron, and investigate several aspects of a theory of uniformly bounded discrete random sets.

2.2 Fundamentals

Definition 2.2.7 *Let B be a bounded subset of $\mathbf{Z}^2$. Assume that B contains the origin. Let $\Sigma(\Omega)$ denote the σ-algebra on Ω. Let $\Sigma(B)$ denote the power set (i.e. the set of all subsets) of B, and let $\Sigma(\Sigma(B))$ denote the power set of $\Sigma(B)$. A Uniformly Bounded Discrete Random Set, or, for brevity,* **Discrete Random Set (DRS)**, *X, on B, is a measurable mapping of a probability space $(\Omega, \Sigma(\Omega), P)$ into the measurable space $(\Sigma(B), \Sigma(\Sigma(B)))$. A DRS X, on B,*

[2]A similar idea has been concurrently and independently developed in [26, 25]. There exists a fundamental difference between the two formulations: we consider uniformly bounded discrete random sets, whereas Goutsias et al., consider discrete random sets on the infinite lattice $\mathbf{Z}^2$. The results evolve in different directions.

induces a unique probability measure, P_X, on $\Sigma(\Sigma(B))$.

Definition 2.2.8 *The functional*

$$T_X(K) = P_X(X \cap K \neq \emptyset),\ K \in \Sigma(B)$$

is called the **capacity functional** *of the DRS X.*

Definition 2.2.9 *The functional*

$$Q_X(K) = P_X(X \cap K = \emptyset) = 1 - T_X(K),\ K \in \Sigma(B)$$

is called the **generating functional** *of the DRS X.*

The generating functional plays for DRS's the role that the cumulative distribution function (cdf) plays for scalar discrete random variables. The following lemma will be useful.

Lemma 2.1 *(Variant of Moebius inversion for Boolean algebras. See [1] for basic Moebius inversion.) Let v be a function on $\Sigma(B)$. Then v can be represented as*

$$v(A) = \sum_{S \subseteq A^c} u(S) \qquad \text{"external decomposition"}$$

The function u is uniquely determined by v, namely

$$u(S) = \sum_{C \subseteq S} (-1)^{|C|} v(S^c \cup C)$$

where ^{c} denotes complement with respect to B.

Proof:

Uniqueness: Assume that the external decomposition formula holds. Look at the right hand side of the inversion formula.

$$\sum_{C\subseteq S}(-1)^{|C|}v(S^c\cup C)=\sum_{C\subseteq S}(-1)^{|C|}\sum_{D\subseteq S\cap C^c}u(D)=$$

$$\sum_{C\subseteq S}(-1)^{|C|}\sum_{D\subseteq S\setminus C}u(D)=\sum_{C\subseteq S}\sum_{D\subseteq S\setminus C}(-1)^{|C|}u(D)=$$

$$\sum_{D\subseteq S}\sum_{C\subseteq S\setminus D}(-1)^{|C|}u(D)=\sum_{D\subseteq S}u(D)\sum_{C\subseteq S\setminus D}(-1)^{|C|}=u(S)$$

Since

$$\sum_{C\subseteq S}(-1)^{|C|}=\begin{cases}0 & ,S\neq\emptyset\\ 1 & ,S=\emptyset\end{cases}$$

Existence: Assume that the inversion formula holds, and look at the right hand side of the external decomposition formula.

$$\sum_{S\subseteq A^c}u(S)=\sum_{S\subseteq A^c}\sum_{C\subseteq S}(-1)^{|C|}v(S^c\cup C)=$$

$$\sum_{S\subseteq A^c}\sum_{C\subseteq S}(-1)^{|C|}v((S\setminus C)^c)=\sum_{D\subseteq A^c}\sum_{C\subseteq A^c\setminus D}(-1)^{|C|}v(D^c)=$$

$$\sum_{D\subseteq A^c}v(D^c)\sum_{C\subseteq A^c\setminus D}(-1)^{|C|}=v((A^c)^c)=v(A)$$

As for the uniqueness part. □

Theorem 2.1 *Given $Q_X(K)$, $\forall K\in\Sigma(B)$, $P_X(A)$, $\forall A\in\Sigma(\Sigma(B))$ is uniquely determined, and, in fact, can be recovered via the measure reconstruction formulas*

$$P_X(A)=\sum_{K\in A}P_X(X=K)$$

with

$$P_X(X = K) = \sum_{K' \subseteq K} (-1)^{|K'|} Q_X(K^c \cup K')$$

Proof:

The functional Q_X can be expressed in terms of P_X as

$$Q_X(K) = \sum_{K' \subseteq K^c} P_X(X = K')$$

This observation, along with lemma 2.1, establish the validity of the theorem.□

The *uniqueness* part of this theorem (cf. theorem 1.1) is originally due to Choquet [9], and it has been independently introduced in the context of continuous-domain random set theory by Kendall [34] and Matheron [44, 45]. Related results can also be found in Ripley [49]. However, the measure reconstruction formulas are essentially only applicable within a uniformly bounded discrete random set formulation. In the case of (uncountably or countably) infinite observation sites, the uniqueness result relies heavily on Kolmogorov's extension theorem, which is non-constructive.

2.3 Examples of DRS models

When it comes to random set modeling, there is no such thing as a good book of "recipes". Random set models are very scarce. To quote Cressie and Laslett [11], "The choice of mathematical models available to the data analyst is often

governed by their tractability, rather than their applicability. When the data are sets, this leaning is even more profound." The Boolean model is important because it is one of the precious few random set models which are both tractable and applicable. In many applications, there exists no physical interpretation of the germ-grain construction; in others there exists strong evidence that the germs and the grains correspond to actual physical entities (e.g. modeling of bomb fields, where the germs correspond to the points of impact and the primary grains model random dispersion). In the former case, the Boolean model is simply used as a device to generate and analyze randomness; in the latter, it also provides some insight into the image data generation mechanism.

2.3.1 Mathematical Morphology

The theory of Mathematical Morphology has been developed mainly by Matheron [45], Serra [57, 21], and their collaborators, during the 70's and early 80's. Since then, Mathematical Morphology and its applications have become very popular. The theory is concerned with the quantitative analysis of shape with an emphasis on geometric structure. It is founded on certain elementary set-to-set mappings, namely set dilation/erosion[3], which are inherently non-linear. These mappings are defined in terms of a *structuring element*, a "small" primitive shape (set of points) which interacts with the input image to transform it, and, in the process, extract useful information about its geometrical and

[3]Here we follow the original definitions of Serra [57]. In his work the symbol $\oplus$ stands for Minkowski set addition, and the symbol $\ominus$ stands for Minkowski set subtraction.

topological structure. Let X_h denote the translate of X by the vector h, and

$$H^s = \{-h \mid h \in H\}$$

Definition 2.3.10 *The* **erosion**, $X \ominus H^s$, *of a set* $X \subset \mathbf{Z}^2$ *by a structuring element* H, *is defined as*

$$X \ominus H^s = \bigcap_{h \in H} X_{-h} = \left\{ z \in \mathbf{Z}^2 \mid H_z \subseteq X \right\}$$

Definition 2.3.11 *The* **dilation**, $X \oplus H^s$, *of a set* $X \subset \mathbf{Z}^2$ *by a structuring element* H, *is defined as*

$$X \oplus H^s = \bigcup_{h \in H} X_{-h} = \left\{ z \in \mathbf{Z}^2 \mid H_z \cap X \neq \emptyset \right\}$$

Erosion and dilation are *dual* operators, in the sense that $X \ominus H^s = (X^c \oplus H^s)^c$, where here ^{c} stands for complementation with respect to $\mathbf{Z}^2$. Two important composite Morphological operators are opening and closing.

Definition 2.3.12 *The* **opening**, $X \circ H$, *of a set* $X \subset \mathbf{Z}^2$ *by a structuring element* H, *is defined as*

$$X \circ H = (X \ominus H^s) \oplus H = \bigcup_{z \in \mathbf{Z}^2 \mid H_z \subseteq X} H_z$$

Definition 2.3.13 *The* **closing**, $X \bullet H$, *of a set* $X \subset \mathbf{Z}^2$ *by a structuring element* H, *is defined as*

$$X \bullet H = (X \oplus H^s) \ominus H$$

By duality of erosion/dilation it follows that opening and closing are dual operators. Both can be viewed as nonlinear smoothing operators. Opening and closing are *idempotent (stable)* operators in the sense that $(X \circ H) \circ H = X \circ H$, and $(X \bullet H) \bullet H = X \bullet H$. A set X is said to be (Morphologically) *open* (*closed*) with respect to the structuring element H iff $X \circ H = X$ ($X \bullet H = X$). We shall say that a set X is *smooth with respect to* H iff X can be expressed as a union of shifted replicas of H, i.e., iff:600 X can be written as $X = L \oplus H$, for some $L \subset \mathbf{Z}^2$. It can be proven [45] that X is open with respect to H, iff X is smooth with respect to H. X is closed with respect to H iff X^c is smooth with respect to H.

2.3.2 A DRS analog of the Boolean model

Let us now define the Boolean DRS. In doing so, we proceed by analogy with the continuous case. We first define the class of germ-grain DRS's. Then we refine this class to derive the Boolean DRS. Most of our work focuses on a particular class of Boolean DRS models, which is closely related to the Morphological analysis of shape, and the associated theory of shape-size distributions [44, 45]. We will study this class in detail.

Proposition 2.1 *Let* $\{X_i\}_{i=0}^{\infty}$ *be a sequence of (not necessarily independent) DRS's, on* B. *Then, for any finite* N, $\bigcup_{i=0}^{N} X_i$ *is a DRS on* B, *and, furthermore, the limit* $\bigcup_{i=0}^{\infty} X_i$ *always exists, and is a DRS on* B.

Proof: Observe that the space of realizations, $\Sigma(B)$, is itself a σ-algebra. The result follows easily. □

Definition 2.3.14 *Let Ψ be a DRS on B, and $\{G_1, G_2, \cdots\}$ be a set of nonempty, iid DRS's on B, characterized by the generating functional Q_G. Define*

$$X = \bigcup_{i=1,2,\ldots} G_i \oplus \{y_i\}$$

where $\Psi = \{y_1, y_2, \cdots\}$. Then X will be called a **germ-grain DRS**. *The points $\{y_1, y_2, \cdots\}$ will be called* the germs, *and the DRS's $\{G_1, G_2, \cdots\}$ will be called* the primary grains *of the DRS X.*

Remark: For brevity, we assume, from this point on, that the result of a $\oplus$ operation is automatically restricted to B. Also, ^{c} stands for complement with respect to B.

In general, we can not compute the generating functional, Q_X, of the germ-grain DRS X, in terms of the generating functional, Q_G, which characterizes the primary grains. This is a significant drawback. Nevertheless, this computation is possible for a restricted class of germ-grain DRS models.

Definition 2.3.15 *Let Ψ be a generalized Bernoulli lattice process (or, Bernoulli DRS, or, binary Bernoulli random field), on B, constructively defined in the following manner: each point $z \in B$ is contained in Ψ with probability $\lambda_s(z)$, independently of all others. Let $\{G_1, G_2, \cdots\}$ be a set of nonempty, iid DRS's*

on B, characterized by the generating functional Q_G. Define

$$X = \bigcup_{i=1,2,\ldots} G_i \oplus \{y_i\}$$

where $\Psi = \{y_1, y_2, \cdots\}$. Then X will be called a **Discrete Boolean Random Set (DBRS)** *, and will be denoted by (λ_s, Q_G)-DBRS. The function λ_s will be called the* intensity function (or, simply, the intensity) *of both the DBRS and the underlying Bernoulli lattice process.*

Proposition 2.2 *Let $\{X_1, X_2, \cdots, X_N\}$ be an independent sequence of nonempty DRS's, characterized by the generating functionals $Q_{X_1}, \cdots, Q_{X_N}$, respectively. Define*

$$Y = \bigcup_{i=1}^{N} X_i$$

Then

$$Q_Y(K) = \prod_{i=1}^{N} Q_{X_i}(K), \quad \forall K \in \Sigma(B)$$

Proof:

$$Q_Y(K) = Pr(Y \cap K = \emptyset) = Pr(X_1 \cap K = \emptyset, \cdots, X_N \cap K = \emptyset)$$

$$= Pr(X_1 \cap K = \emptyset) \cdots Pr(X_N \cap K = \emptyset) = \prod_{i=1}^{N} Q_{X_i}(K)$$

and the proof is complete. □

Proposition 2.3 *The generating functional of a (λ_s, Q_G)-DBRS X is given by*

$$Q_X(K) = \prod_{z \in B} \left[1 - \lambda_s(z) + \lambda_s(z) Q_{G \oplus \{z\}}(K)\right]$$

$$= \prod_{z \in B} [1 - \lambda_s(z) + \lambda_s(z) Q_G(K \oplus \{-z\})]$$

Proof:

Follows from independence, and proposition 2.2. □

Let H be a "small", "primitive" subset of B, which contains the origin, $\{\bar{0}\}$. In the terminology of Mathematical Morphology, H is a structuring element. We also assume that H is convex[4]. In the discrete case, the notion of size of a convex structuring element can be formalized via the operation of set dilation. Define

$$rH \triangleq \begin{cases} \{\bar{0}\} \oplus H \oplus H \oplus \cdots \oplus H, \quad (r \text{ times}) & , r = 1, 2, \ldots \\ \\ \{\bar{0}\} & , r = 0 \end{cases}$$

Definition 2.3.16 *Let Ψ be a generalized Bernoulli lattice process on B, of intensity function λ_s. Let $\{G_1, G_2, \cdots\}$ be a set of nonempty, convex iid DRS's on B, each given by $G_i = R_i H$, where $\{R_1, R_2, \cdots\}$ form an iid sequence of $\mathcal{Z}_+$-valued r.v.'s which is independent of Ψ, and each R_i is distributed according to a pmf $f_R(r)$, which is compactly supported on $\{0, 1, \ldots, \bar{R}\}$. Define*

$$X = \bigcup_{i=1,2,\ldots} G_i \oplus \{y_i\}$$

[4]In digital topology [39, 57, 25], the *convex hull* of a bounded set, $H \subset \mathbf{Z}^2$, is defined as the intersection of the convex hull of H in the topology of $\mathcal{R}^2$, with $\mathbf{Z}^2$. A bounded set, $H \subset \mathbf{Z}^2$, is *convex* iff it is identical to its convex hull.

where $\Psi = \{y_1, y_2, \cdots\}$. *Then* X *will be called a* **Discrete Radial Boolean Random Set (DRBRS)** , *with parameters* (λ_s, H, f_R), *and will be denoted by* (λ_s, H, f_R)*-DRBRS.*

We now proceed to compute the generating functional of a (λ_s, H, f_R) - DRBRS. Define

$$d^H(z, K) = min_{k \in K} ||z - k||_H$$

where

$$||z - k||_H = min\{n \geq 0 \mid (\{z\} \oplus nH) \cap \{k\} \neq \emptyset\}$$

Observe that for $z \in K$, $d^H(z, K) = 0$, since H contains the origin. With this notation in place, using proposition 2.3, and employing some simple geometric arguments (which amount to making sure that the primary grains which could hit K actually refrain from doing so), it can be shown that

$$Q_X(K) = \prod_{z \in K \oplus \check{R}H^s} \left[(1 - \lambda_s(z)) + \lambda_s(z) F_R(d^H(z, K) - 1)\right]$$

where

$$F_R(m) = \sum_{l=0}^{m} f_R(l)$$

and $F_R(-1) = 0$, by convention.

We can now use theorem 1 to compute $P_X(X = K)$ in terms of the model parameters.

$$P_X(X = K) = \sum_{K' \subseteq K} (-1)^{|K'|} \times$$
$$\prod_{z \in (K^c \cup K') \oplus \check{R}H^s} \left[(1 - \lambda_s(z)) + \lambda_s(z) F_R(d^H(z, K^c \cup K') - 1)\right]$$

2.4 Randomized superposition of DRS's

In this section we consider two important structural properties of DRS's, namely divisibility and superposition.

Definition 2.4.17 *Let* $\{X_1, X_2, \cdots, X_N\}$ *be an iid sequence of nonempty DRS's. The* **divisibility degree** *of a DRS* Y *is defined as*

$$dg(Y) = sup\left\{ N \mid Y = \bigcup_{i=1}^{N} X_i \right\}$$

Definition 2.4.18 *A DRS* Y *is* **divisible** *iff* $dg(Y) > 1$, *and* **indivisible** *iff* $dg(Y) = 1$. *A DRS* Y *is* **infinitely divisible** *iff* $dg(Y) = \infty$.

In general, and in contrast to its continuous-domain counterpart, the Boolean DRS *is not* infinitely divisible. However, based on the Boolean DRS (or any other DRS model, for that matter), we can derive infinitely divisible DRS's. This construction (which we call randomized superposition, for reasons that will soon be obvious) is of considerable theoretical and practical importance (cf. the discussion in section 1.3.1).

Proposition 2.4 *Let* N *be a non-negative integer-valued random variable, and let* $\{X_1, X_2, \cdots, X_N\}$ *be an iid sequence of nonempty DRS's, characterized by the generating functional* Q_X, *and independent of* N. *Define*

$$Y = \bigcup_{i=1}^{N} X_i$$

If N is a Binomial r.v.

$$Pr(N=n) = \binom{M}{n} p^n (1-p)^{M-n}, \quad n = 0, 1, \ldots, M, \quad p \in (0,1), \; M < \infty$$

then

$$Q_Y(K) = (1 - p + pQ_X(K))^M, \quad \forall K \in \Sigma(B)$$

If N is a Poisson r.v.

$$Pr(N=n) = \frac{\lambda^n}{n!} e^{-\lambda}, \quad n = 0, 1, \ldots, \quad 0 < \lambda < \infty$$

then

$$Q_Y(K) = e^{-\lambda(1-Q_X(K))}, \quad \forall K \in \Sigma(B)$$

Finally, if N is a geometrically-distributed r.v

$$Pr(N=n) = (1-p)p^n, \quad , n = 0, 1, \ldots, \quad p \in (0,1)$$

then

$$Q_Y(K) = \frac{1-p}{1 - pQ_X(K)}, \quad \forall K \in \Sigma(B)$$

Proof:

Clearly (cf. proposition 2.2),

$$Q_Y(K) = \sum_n [Q_X(K)]^n \, Pr(N=n)$$

If N is a binomial r.v.

$$Q_Y(K) = \sum_{n=0}^{M} [Q_X(K)]^n \binom{M}{n} p^n (1-p)^{M-n}$$

$$= (1-p)^M \sum_{n=0}^{M} \binom{M}{n} \left[\frac{p}{1-p} Q_X(K)\right]^n$$

Recall that, for any complex number λ,

$$\sum_{n=0}^{M} \binom{M}{n} \lambda^n = (1+\lambda)^M$$

Therefore,

$$Q_Y(K) = (1-p)^M \left(1 + \frac{p}{1-p} Q_X(K)\right)^M$$

$$= (1 - p + pQ_X(K))^M$$

If N is Poisson

$$Q_Y(K) = \sum_{n=0}^{\infty} [Q_X(K)]^n \frac{\lambda^n}{n!} e^{-\lambda} = e^{-\lambda} \sum_{n=0}^{\infty} \frac{[\lambda Q_X(K)]^n}{n!}$$

$$= e^{-\lambda} e^{\lambda Q_X(K)} = e^{-\lambda(1-Q_X(K))}$$

Finally, if N is geometrically distributed

$$Q_Y(K) = \sum_{n=0}^{\infty} [Q_X(K)]^n (1-p) p^n$$

$$= (1-p) \sum_{n=0}^{\infty} [pQ_X(K)]^n$$

$$= \frac{1-p}{1 - pQ_X(K)}$$

and the proof is complete. □

When N is a Poisson r.v., the resulting generating functional is of exponential type. In this case, it can be easily shown that Y is infinitely divisible. In fact, the generating functional of *any* infinitely divisible random set (be it continuous or discrete) is constrained to be of exponential type [45, 23].

Randomized superposition corresponds to a nonlinear deformation of the generating functional of the component DRS's, as it can be clearly seen in figure 2.1. Figure 2.1a presents a plot of the values of the generating functional of the DRS result of randomized superposition, when N is Binomialy distributed with parameters M, p, versus the values of the generating functional of the component DRS's, for $p = 0.3$, and for various values of M. Figure 2.1b presents the same plot, when N is distributed according to the Poisson pmf for various values of the parameter λ. Similarly, figure 2.1c corresponds to the case where N is distributed according to the geometric pmf with parameter p, for various values of p. Figure 2.1d compares the three deformations, for one particular choice of the parameters which forces the left endpoints to coincide.

Some remarks are in order. First, if we can sample from the distribution of the component DRS's, then we can also sample from the distribution of the resulting DRS. Thus, we can simulate the resulting model. Second, randomized superposition modifies the *ensemble properties* without affecting the *sample properties*; i.e. a sample of the resulting model will look like a "dense" sample of the original model. However, the statistical behavior of the former can be significantly different than that of the latter, as the plots in figure 2.1 demonstrate. Finally, by superimposing a random Poisson distributed number of iid DBRS realizations, we obtain an infinitely divisible DRS, whose individual realizations appear like realizations of a "dense" DBRS model. In general, it seems that, in the uniformly bounded DRS case, this "doubly stochastic" construction is

necessary in order to obtain an infinitely divisible analog of the Boolean random set model.

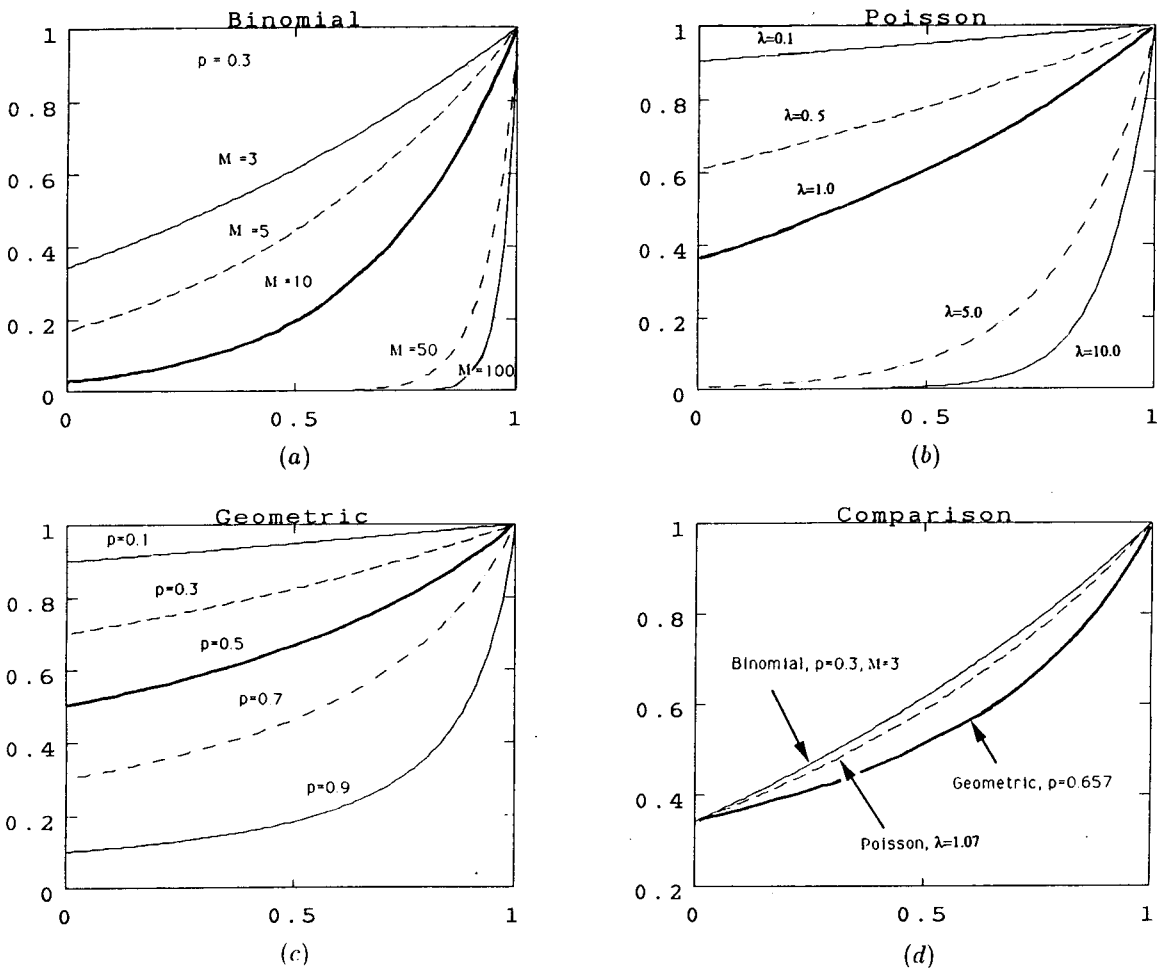


Figure 2.1: Randomized superposition corresponds to a nonlinear deformation of the generating functional of the component DRS's

CHAPTER
THREE

STATISTICAL INFERENCE OF THE DRBRS MODEL

3.1 Introduction

The cornerstone of statistical image analysis is the development of models which summarize the most important characteristics of images by means of a few parameters. Once a bagfull of discrete random set models has been put to our disposal, the next step is to fit the models to the data. This is, of course, obvious. Parametric models whose parameters cannot be fitted to the data are useless for the purposes of image analysis (although they can be of some use in image synthesis and computer graphics). Thus enters the need for statistical inference.

Statistical inference techniques are very scarce in random set theory. Most of the existing literature is concerned with parameter estimation for the Boolean model, based on some variant of the method of moments[1] [56, 13, 11, 57, 21, 54]. This can be partially attributed to the lack of a total set ordering, which makes

[1]The exception is an approximate ML approach which has been proposed in [20]. However, there exist problems with this approach, as it has been pointed out in [4].

for an extremely complicated "distribution function". The method of moments, while well-known and widely practiced, is largely considered as a last resort, when standard inference procedures (e.g., ML, MAP) can not be used. We will demonstrate that within a uniformly bounded discrete random set framework we can make some steps towards the ML, MAP inference of the Boolean model.

3.2 Intensity estimation in the case of constant radii and unknown constant intensity

In the previous chapter, we have seen that the likelihood function of a (λ_s, H, f_R)-DRBRS X is given by

$$P_X(X = K) = \sum_{K' \subseteq K} (-1)^{|K'|} \times$$

$$\prod_{z \in (K^c \cup K') \oplus \check{R}H^s} \left[(1 - \lambda_s(z)) + \lambda_s(z) F_R(d^H(z, K^c \cup K') - 1) \right]$$

Even though we have been able to write down an expression for the likelihood function, we are still faced with a complicated formula, which is difficult to work with. In particular, and largely due to the highly oscillatory Moebius kernel, $(-1)^{|K'|}$, it is not directly amenable to optimization, which immediately rules out direct ML parameter estimation. Furthermore, the computational complexity associated with a brute-force calculation of the likelihood is exponential in $|K|$. One would therefore be interested in obtaining tight bounds on $P_X(X = K)$. In order to be useful, these bounds must be reasonably well behaved, and relatively easy to compute. For the simple case of a DRBRS model of constant intensity,

$\lambda_s(z) = p = 1-q$, $\forall z \in B$, and primary grains of fixed size (one, by convention), the generating functional is simply given by

$$Q_X(K) = q^{|K \oplus H^s|}$$

We have the following result for this model.

Proposition 3.1 *For all* $q \in [0,1]$, *and all* realizable[2] $K \in \Sigma(B)$, $K \neq \emptyset, B$

$$L_q(K) \leq P_X(X = K) \leq U_q(K)$$

with

$$L_q(K) = q^{|K^c \oplus H^s|}(1-q)^{|(K^c \oplus H^s)^c|}$$

and

$$U_q(K) = \frac{1}{2} q^{|K^c|} \left[(1+q)^{|K|} + (1-q)^{|K|} \right] - 2^{|K|-1} q^{|K^c \oplus H^s| + |K \oplus H^s|}$$

Both bounds are polynomials *in* q, *they are equal to zero at the endpoints* $q = 0, 1$, *strictly positive for all* $q \in (0,1)$, *and* unimodal *in* $(0,1)$. *The mode of the lower bound is located at*

$$\underline{\hat{q}}(K) = \frac{|K^c \oplus H^s|}{|B|}$$

Proof:

Upper bound:

$$P_X(X = K) = \sum_{K' \subseteq K} (-1)^{|K'|} q^{|(K^c \cup K') \oplus H^s|}$$

[2]Meaning that K can be written as $K = L \oplus H$, for some $L \in \Sigma(B)$. If K can not be written this way, then it is not a realization of the DRBRS model under consideration, and, therefore, its probability is zero.

$$= \sum_{K' \subseteq K,\ |K'|=even} q^{|(K^c \cup K') \oplus H^s|} - \sum_{K' \subseteq K,\ |K'|=odd} q^{|(K^c \cup K') \oplus H^s|}$$

Observe that, by distributivity of dilation over union, and using the union bound

$$|(K^c \cup K') \oplus H^s| = |(K^c \oplus H^s) \cup (K' \oplus H^s)| \leq |K^c \oplus H^s| + |K' \oplus H^s|$$

Furthermore, since H is assumed to contain the origin

$$|(K^c \cup K') \oplus H^s| \geq |K^c \cup K'| = |K^c| + |K'|$$

Therefore, since q is a probability

$$\begin{aligned}
P_X(X = K) &\leq \sum_{K' \subseteq K,\ |K'|=even} q^{|K^c|+|K'|} - \sum_{K' \subseteq K,\ |K'|=odd} q^{|K^c \oplus H^s|+|K' \oplus H^s|} \\
&= q^{|K^c|} \sum_{K' \subseteq K,\ |K'|=even} q^{|K'|} - q^{|K^c \oplus H^s|} \sum_{K' \subseteq K,\ |K'|=odd} q^{|K' \oplus H^s|} \\
&\leq q^{|K^c|} \sum_{K' \subseteq K,\ |K'|=even} q^{|K'|} - q^{|K^c \oplus H^s|} \sum_{K' \subseteq K,\ |K'|=odd} q^{|K \oplus H^s|} \\
&= q^{|K^c|} \sum_{K' \subseteq K,\ |K'|=even} q^{|K'|} - q^{|K^c \oplus H^s|+|K \oplus H^s|} \sum_{K' \subseteq K,\ |K'|=odd} 1
\end{aligned}$$

Thus

$$P_X(X = K) \leq q^{|K^c|} \left(\sum_{i:even} \binom{|K|}{i} q^i \right) - q^{|K^c \oplus H^s|+|K \oplus H^s|} \left(\sum_{i:odd} \binom{|K|}{i} \right)$$

Using the fundamental identity

$$\sum_{i=0}^{|K|} \binom{|K|}{i} z^i = (1+z)^{|K|}, \quad \forall z \in \mathcal{C}$$

and successively setting $z = -1, 1$, we obtain

$$\sum_{i:odd} \binom{|K|}{i} = 2^{|K|-1}$$

Similarly, replacing z by qz and then setting $z = -1, 1$, we obtain

$$\sum_{i:even} \binom{|K|}{i} q^i = \frac{1}{2}\left[(1+q)^{|K|} + (1-q)^{|K|}\right]$$

From which, we finally obtain the expression for the upper bound.

We will need the following lemma.

Lemma 3.1 *(Descartes' rule of signs [47, pp. 36-43]) Let $p(x)$ be a polynomial of a real variable, with real coefficients.*

$$p(x) = \alpha_0 + \alpha_1 x + \alpha_2 x^2 + \cdots + \alpha_n x^n$$

Let C denote the number of changes of sign of the sequence of its coefficients (for each $m \geq 1$, if $\alpha_{m-1}\alpha_m < 0$, then (α_{m-1}, α_m) constitute a change of sign). Let Z be the number of positive *real zeros of $p(x)$ (a zero of multiplicity k is counted as k zeros). Then:*

$$C - Z \geq 0$$

and $C - Z$ is an even number.

Using once more the identity

$$\sum_{i=0}^{|K|} \binom{|K|}{i} z^i = (1+z)^{|K|}, \quad \forall z \in \mathcal{C}$$

it can be seen that $U_q(K)$ can be written as

$$U_q(K) = \sum_{i:even}^{|K|} \binom{|K|}{i} q^{|K^c|+i} - 2^{|K|-1} q^{|K^c \oplus H^s|+|K \oplus H^s|}$$

Since all the coefficients of this polynomial are strictly positive, except for the coefficient of the highest degree which is strictly negative, by employing

Descartes' rule of signs, we conclude that $U_q(K)$ has at most one zero in $(0,\infty)$. But $U_1(K) = 0$, and, therefore, this is the unique zero in $(0,\infty)$. Hence, $U_q(K) > 0,\ \forall q \in (0,1)$.

Next, consider the derivative of the upper bound, with respect to q. After some algebraic manipulation, it can be written as

$$\frac{d}{dq}U_q(K) = q^{|K^c|-1}\left[|K^c| + \sum_{i=2,\ i:even}^{|K|}\left[\binom{|K|}{i}(|K^c|+i)\right]q^i - \right.$$
$$\left. 2^{|K|-1}\left[|K^c \oplus H^s| + |K \oplus H^s|\right]q^{|K^c\oplus H^s|+|K\oplus H^s|-|K^c|}\right]$$

Again, since all the coefficients of this polynomial are strictly positive, except for the coefficient of the highest degree which is strictly negative, by employing Descartes' rule of signs, we conclude that $\frac{d}{dq}U_q(K)$ has at most one zero in $(0,\infty)$. But

$$\left.\frac{d}{dq}U_q(K)\right|_{q=0} = 0$$
$$\left.\frac{d}{dq}U_q(K)\right|_{q=0^+} > 0$$

and

$$\left.\frac{d}{dq}U_q(K)\right|_{q=1} = |K^c|2^{|K|-1} + |K|2^{|K|-2}$$
$$-2^{|K|-1}\left[|K^c \oplus H^s| + |K \oplus H^s|\right] < 0,\ \forall K \neq \emptyset$$

Therefore, by continuity, we conclude that $\frac{d}{dq}U_q(K)$ has at least one zero in $(0,1)$, which must also be unique. Hence, since its derivative has only one zero crossing in $(0,1)$, $U_q(K)$ must be *unimodal* in $(0,1)$.

Lower bound: It can be easily seen that *one* possible germ configuration which can give rise to the observation, K, is given by the set of points $(K^c \oplus H^s)^c$. In particular, let L denote the germ point process (which is itself a DRS). Then X can be written as $X = L \oplus H$. By simple geometric arguments $(X^c \oplus H^s)^c \oplus H = X$, and $L \subseteq (X^c \oplus H^s)^c$, i.e. $X^c \oplus H^s \subseteq L^c$. Hence

$$P_X(X = K) \geq q^{|K^c \oplus H^s|}(1-q)^{|(K^c \oplus H^s)^c|}$$

The lower bound is strictly positive for all $q \in (0,1)$ (by inspection). One can show that it is unimodal by looking at its derivative, and employing Descartes' rule of signs. After some manipulation,

$$\frac{d}{dq}L_q(K) = \sum_{i=0}^{|(K^c \oplus H^s)^c|} \binom{|(K^c \oplus H^s)^c|}{i} (-1)^i (|K^c \oplus H^s| + i) q^{|K^c \oplus H^s| + i - 1}$$

Since $K \neq \emptyset$, and K is realizable (meaning that it can be written as $K = L \oplus H$, for some $L \in \Sigma(B)$), it follows that $|(K^c \oplus H^s)^c| \geq 1$, $|K^c \oplus H^s| \geq 1$, in which case the corresponding sequence of coefficients of the polynomial $\frac{d}{dq}L_q(K)$ has exactly $|(K^c \oplus H^s)^c|$ sign changes, and, therefore, $C = |(K^c \oplus H^s)^c|$. Hence, the number of positive zeros is $Z \leq C = |(K^c \oplus H^s)^c|$. However, $\frac{d}{dq}L_q(K)$ can also be written as

$$\frac{d}{dq}L_q(K) = (1-q)^{|(K^c \oplus H^s)^c| - 1} \times$$

$$\left[|K^c \oplus H^s| q^{|K^c \oplus H^s| - 1}(1-q) - q^{|K^c \oplus H^s|} |(K^c \oplus H^s)^c| \right]$$

from which it is obvious that it has a zero of multiplicity $|(K^c \oplus H^s)^c| - 1$ at $q = 1$. Therefore, since $C - Z$ must be an even number, there remains

one more positive zero to be accounted for. By inspection of the last formula, $\frac{d}{dq}L_q(K) \neq 0$, *for* $q > 1$. Hence, this last zero is in $(0,1)$. Since the derivative has one and only one zero crossing in $(0,1)$, we have established unimodality of $L_q(K)$ in $(0,1)$. The mode location is obtained by simple differentiation of the logarithm of the lower bound. □

The measurements $|K|, |K \oplus H^s|, |K^c \oplus H^s|$, can be interpreted as "crude" statistical summaries of the observation. Bounds of this type can be used to estimate a *feasible region* of the true Maximum Likelihood (ML) estimate of q, given the observation K. Since these bounds are typically very high degree polynomials, their (unique) modes are very sharp, leading to very accurate localization of the true ML estimate of q. Specifically, since the bounds are *unimodal polynomials*, the true maximum likelihood estimate of q must be within the closed interval of q-values delimited by the two q-values at which the upper bound is equal to the peak of the lower bound. This situation is illustrated in figure 3.1. This technique can also be used for binary hypothesis testing between two values of q. If these two values are sufficiently far apart, the bounds will indicate that the corresponding likelihoods are in disjoint regions, in which case a decision procedure based on these bounds is as good as one based on the exact likelihood, but tremendously faster. We will return to hypothesis testing later on.

The mode, $\underline{\hat{q}}(K)$, of the lower bound, $L_q(K)$, *underestimates* q, i.e. $\underline{\hat{q}}(X)$ is a *biased* estimator of q on the basis of the observation, X. This can be seen as

follows.

$$E\underline{\hat{q}}(X) = \frac{E|X^c \oplus H^s|}{|B|}$$

Let L denote the germ point process. Then $X^c \oplus H^s \subseteq L^c$. Thus

$$|X^c \oplus H^s| \leq |L^c|$$

and

$$E\,|X^c \oplus H^s| \leq E\,|L^c|$$

So

$$E\underline{\hat{q}}(X) \leq \frac{E\,|L^c|}{|B|} = \frac{q|B|}{|B|} = q$$

A simple estimator of q on the basis of X is

$$\hat{\bar{q}}(X) = \frac{|B| - \mathcal{CC}(X)}{|B|}$$

where $\mathcal{CC}(X)$ is the number of connected components of X. Obviously, this estimator overestimates q, because the number of germs (points) of any particular realization of L is always greater than or equal to the number of connected components of the corresponding realization of $X = L \oplus H$. We also remark that the ML estimator of q on the basis of X is not guaranteed to be unbiased. In practice, for "typical" observations, all these estimates are "close" to each other. As an example, figure 3.2 depicts a realization of a DRBRS of constant intensity and fixed primary grain. For this example, $q = 0.999$, and the computed estimates are $\underline{\hat{q}} = 0.998$, $\hat{\bar{q}} = 0.99918$, whereas the feasible region is $[0.9978, 0.9992]$.

3.2.1 Noisy observations

Quite often we do not have the luxury of observing a noise-free realization of X. In practice, images are typically corrupted by sensor noise, sampling errors, and transmission errors. Therefore, it is of interest to investigate the robustness of estimation tools in a noisy environment. To do so, we need to assume a reasonable degradation mechanism, and get a grasp on the sensitivity of the relevant statistics as a function of a suitable parametrization of the noise.

One particular degradation model is the *independent union noise model.* It assumes that the observable DRS, Y, is the union of the "signal" DRS, X, with a "noise" DRS, N, which is independent of X. The statistics of N can be arbitrary. If N can be modeled as another DRBRS, of constant intensity and fixed primary grain (different than that of X), then we can work out similar bounds on the probability of the observation. However, in this case the upper bound is not unimodal. Alternatively, we can try to break up the solution into a signal estimation step, and a parameter estimation step. The signal estimation step provides an estimate, $\widehat{X}(Y)$, of the signal, X, on the basis of the observation, Y, whereas the parameter estimation step computes the necessary statistics on the estimate $\widehat{X}(Y)$. This approach is clearly suboptimal. However, if the estimate of X remains reasonably close to X, then we expect the overall procedure to be nearly optimal. Since our approach is essentially ML-based, it makes sense to use a ML estimator to perform the signal estimation step. Towards this end, we

have the following lemma[3].

Lemma 3.2 *Let $O_H(B)$ denote the collection of all H-open subsets of B. Assume that the signal DRS, X, on B, induces a probability mass function on $\Sigma(B)$ which has the following property*

$$P_X(X = K) = 0, \; iff \; K \in \Sigma(B) \backslash O_H(B)$$

where $\backslash$ stands for set difference. Furthermore, assume that the observable DRS is $Y = X \cup N$, where N is a homogeneous Bernoulli lattice process of intensity $r \in [0, 1)$ (i.e. each point $z \in B$ is included in N with probability r, independently of all other points), which is independent of X. Then $Y \circ H$ is the unique ML estimate of X on the basis of Y, regardless of the specific value of r.

Proof:

Let $\widehat{X}_{ML}(Y)$ denote the ML estimate of X on the basis of Y. Then, by definition,

$$\widehat{X}_{ML}(Y) = arg \; max_{K \in O_H(B)} \{Pr(Y \mid X = K)\}$$

$$= arg \; max_{K \in O_H(B), \; K \subseteq Y} \{Pr(Y \mid X = K)\}$$

$$= arg \; max_{K \in O_H(B), \; K \subseteq Y} \left\{ r^{|Y|-|K|} (1-r)^{|B|-|Y|} \right\}$$

$$= arg \; max_{K \in O_H(B), \; K \subseteq Y} \left\{ r^{-|K|} \right\}$$

$$= arg \; max_{K \in O_H(B), \; K \subseteq Y} \{|K|\}$$

[3]This is a variant of theorem 4.1.

So $\widehat{X}_{ML}(Y)$ is the largest H-open subset of Y, which is by definition the opening of Y by H, i.e.

$$\widehat{X}_{ML}(Y) = Y \circ H$$

and the proof is complete. □

Observe that the proof crucially depends on $|B|$ being finite. Now, it can be readily seen that (modulo some unavoidable edge effects) a DRBRS X of constant intensity and fixed primary grain, H, satisfies the condition of this lemma. Thus, if the observable DRS, Y, can be modeled as the union of the signal DRS, X, with an independent realization of a homogeneous Bernoulli lattice process (an assumption which is often reasonable in practice), then the ML estimate of X on the basis of Y is simply $Y \circ H$. It is worth noting that this ML signal estimation step does not assume knowledge of r, i.e. it is independent of the noise level. In practice, this estimate remains very close to X even when the intensity, r, of the noise process is large. Simulation experiments have demonstrated that the statistics $|\widehat{X}_{ML}(Y)|$, $|\widehat{X}_{ML}(Y) \oplus H^s|$, $|(\widehat{X}_{ML}(Y))^c \oplus H^s|$ are very robust, i.e. they remain very close to $|X|$, $|X \oplus H^s|$, $|X^c \oplus H^s|$, respectively, for up to 80% noise, i.e. for $r = 0.8$. The modes of the bounds themselves are rather insensitive to small perturbations of the statistics. For example, the mode of the lower bound

$$\underline{\hat{q}}(X) = \frac{|X^c \oplus H^s|}{|B|}$$

is robust under small perturbations of $|X^c \oplus H^s|$. Let us illustrate this ap-

proach. Consider figure 3.3. It depicts a realization of the observable DRS, Y, obtained by taking the union of the signal DRS, X, which is depicted in figure 3.2, with an independent realization of a homogeneous Bernoulli lattice process of intensity $r = 0.5$. The ML estimate of X on the basis of the realization of figure 3.3 is depicted in figure 3.4. As it can be seen, there is hardly any discernible difference between the realizations of figures 3.2 and 3.4. The statistics $|\widehat{X}_{ML}(Y)|$, $|\widehat{X}_{ML}(Y) \oplus H^s|$, $|(\widehat{X}_{ML}(Y))^c \oplus H^s|$ all differ by less than 0.4% from their nominal values, i.e. $|X|$, $|X \oplus H^s|$, $|X^c \oplus H^s|$, respectively.

3.2.2 Nonconstant radii

These bounds can be extended to the case of a DRBRS model of constant intensity, $\lambda_s(z) = p = 1 - q$, $\forall z \in B$, and primary grains of random size, by using the following approximation of the corresponding generating functional.

Lemma 3.3 *For a DRBRS X, of constant intensity, $p = 1-q$, and q sufficiently close to 1, the following approximation is valid*

$$Q_X(K) \cong q^{E|K \oplus RH^s|}$$

where the expectation is taken with respect to the pmf f_R of the radii.

Proof:

For convenience, let $S \triangleq RH$ denote the "generic" primary grain, centered at the origin. Then,

$$Q_X(K) = \prod_{z \in K \oplus \check{R}H^s} (1 - p\, T_S(K \oplus \{-z\}))$$

with

$$T_S(K) = 1 - F_R\left(d^H(\{(0,0)\}, K) - 1\right)$$

By definition of the distance metric, d^H, we can extend the product domain to the entire base frame, B.

$$Q_X(K) = \prod_{z \in B} (1 - p\, T_S(K \oplus \{-z\}))$$

Thus

$$logQ_X(K) = \sum_{z \in B} log\,(1 - p\, T_S(K \oplus \{-z\}))$$

Taking the derivative with respect to p, and evaluating at $p = 0$, we obtain

$$\frac{d}{dp} logQ_X(K)\Big|_{p=0} = -\sum_{z \in B} T_S(K \oplus \{-z\}) =$$

$$-\sum_{z \in B} Pr(S \cap (K \oplus \{-z\}) \neq \emptyset)$$

$$= -\sum_{z \in B} E\ 1(S \cap (K \oplus \{-z\}) \neq \emptyset)$$

$$= -E \sum_{z \in B} 1(S \cap (K \oplus \{-z\}) \neq \emptyset) = -E|K \oplus S^s|$$

Since

$$p = 0 \mapsto Q_X(K) = 1 \mapsto logQ_X(K) = 0$$

we have the following first order Taylor series approximation *of the logarithm of* $Q_X(K)$.

$$logQ_X(K) \cong \left(\frac{d}{dp} logQ_X(K)\Big|_{p=0}\right) p$$

i.e.

$$logQ_X(K) \cong -pE|K \oplus S^s|$$

$$Q_X(K) \cong e^{-pE|K\oplus S^s|} = \left(e^{-p}\right)^{E|K\oplus S^s|} = q^{E|K\oplus S^s|}$$

since, for p close to 0, $e^{-p} \cong (1-p) = q$. □

Using this approximation, which is asymptotically good as q goes to 1, and theorem 2.1, we can obtain the same upper bound, but this time on the *approximate (instead of the actual) probability.* In this case, H in the expression for the upper bound is replaced by $\bar{R}H$, where $\bar{R}$ is the maximum possible radius. A unimodal lower bound for this case can be obtained by employing the *Morphological Skeleton Transform.*

3.2.3 Morphological Skeletonization as a method of obtaining a consistent realization of the embedded marked point process

Another approach to the problem of estimating the probability of a given observation is suggested by looking at it from the viewpoint of shape analysis. The idea is that we can use certain Morphological shape description schemes to obtain one realization of the underlying marked point process (i.e. the germ points marked by their corresponding radii) which can give rise to the observed realization, K, of the DRBRS X. Then we can obtain a lower bound on $P_X(X = K)$ simply by computing the probability of this realization of the marked point process. If the grains of K are disconnected[4] and contained in B, then there exists *a unique* realization of the underlying marked point process that can give rise to K. In this case, the unique realization can be recovered, and the exact probabil-

[4]In the chessboard-block (or, 8-nearest neighbors) sense.

ity, $P_X(X = K)$, can be computed. This approach can lead to good estimation and hypothesis testing procedures if the data are "sufficiently sparse". As an example, let us consider the simple vs. simple hypothesis testing problem:

$$H_0\text{: } X \sim (\lambda_s^{(0)}, H, f_R^{(0)})\text{-DRBRS}$$

$$vs.\ H_1\text{: } X \sim (\lambda_s^{(1)}, H, f_R^{(1)})\text{-DRBRS}$$

In principle , given any observation $K \in \Sigma(B)$, the probability of this observation under each one of the two hypotheses can be computed using theorem 2.1, and the Bayesian rule of choice can be implemented. In practice, the computational cost associated with this brute-force method limits its applicability. We therefore pursue an alternative approach. The key idea is the following. Suppose that instead of the DRBRS realization, K, we were given the realization of the germ points $\{y_1, y_2, \cdots\}$ and the associated radii $\{R_1, R_2, \cdots\}$ that produced K. Let these data be represented by an ordered list of collections of sites $\{L_0, \cdots, L_{\bar{R}}\}$, corresponding to radii $\{0, \cdots, \bar{R}\}$ respectively. Note that, for one or more $n \in \left\{0, \cdots, \bar{R}\right\}$, L_n may be empty. The loglikelihood ratio test for these data is simply given by:

$$log\frac{Pr_1\{L_0, \cdots, L_{\bar{R}}\}}{Pr_0\{L_0, \cdots, L_{\bar{R}}\}} = \sum_{z \in B \mid z \notin \cup_{n=0}^{\bar{R}} L_n} log\left(\frac{1 - \lambda_s^{(1)}(z)}{1 - \lambda_s^{(0)}(z)}\right)$$

$$+ \sum_{n=0}^{\bar{R}} |L_n| log\left(\frac{f_R^{(1)}(n)}{f_R^{(0)}(n)}\right) + \sum_{n=0}^{\bar{R}} \sum_{z \in L_n} log\frac{\lambda_s^{(1)}(z)}{\lambda_s^{(0)}(z)} \underset{H_0}{\overset{H_1}{\gtrless}} t \quad (*)$$

where the optimal threshold, t, is a function of the prior probabilities of the two hypotheses, and the losses incurred when different kinds of decision errors are made. Therefore, we can easily classify the observation, according to Bayesian

decision theory. However, the recovery of these data from the observation K is an ill-posed problem.

Simply put, the *Morphological Skeleton*[5] [43, 24] of a binary shape, K, with respect to a structuring element, H, is the locus of the centers of all *maximal inscribable replicas* of H in K. A *replica* of H is a scaled and shifted version of H. A replica of H is *maximal* in K *iff* it can not be properly contained in any other replica of H which can be inscribed in K. The *Morphological Skeleton Function (MSF)* of K with respect to H is the function whose support is the Morphological Skeleton of K with respect to H, and its value at each skeleton point is equal to the radius of the corresponding maximal inscribable replica of H. The Morphological Skeleton is explicitly given by

$$SK(K) = \bigcup_{n=0}^{N} S_n(K) = \bigcup_{n=0}^{N} [(K \ominus nH^s) - (K \ominus nH^s) \circ H]$$

where

$$N = max\,\{n \mid K \ominus nH^s \neq \emptyset\}$$

The set $S_n(K)$ is the locus of centers of maximal inscribable replicas of size n, and it is called the n^{th} skeleton subset of K. Given all the skeleton subsets, K can be reconstructed via

$$K = \bigcup_{n=0}^{N} S_n(K) \oplus nH \quad (**)$$

[5]Many other related notions of a skeleton exist. However, the given definition is sufficient for our purposes.

Given all the skeleton subsets, the MSF is uniquely determined. Conversely, given the MSF, all the skeleton subsets are uniquely determined.

From equation (**), it is clear that the MSF provides *one* realization of the germ points (the support set of the MSF), along with their associated radii (the values of the MSF), which can give rise to K. We propose the use of the log-likelihood ratio test (*) applied to these data (i.e. $L_n = S_n,\ n = 0, \cdots, N$, and $L_n = \emptyset,\ N < n \leq \bar{R}$) as a decision rule for the simple hypothesis testing problem under consideration[6]. If the grains of K are disconnected[7] (a situation which arises with high probability if the intensity of the germ process is uniformly low and $\bar{R}$ is small), and contained in B, then the true (unique) realization of the underlying marked point process is actually recovered[8], and the proposed decision rule is exact Maximum Likelihood. The overall procedure can be efficiently implemented (in polynomial time), thanks to the existence of fast Morphological Skeletonization algorithms [43]. Figure 3.5 depicts a realization of a DRBRS and its skeleton. Simulation results have been very encouraging, even when the primary grains overlap substantially. These simulations suggested that, for the purposes of hypothesis testing between two DRBRS models of different intensities, the size of the skeleton is an important statistic, in the sense of possessing high discriminatory power. This prompted us to investigate whether it is possi-

[6]This idea has been concurrently and independently developed in [26], as a means of performing shape-size analysis and synthesis of a different class of DRS models.

[7]In the chessboard-block, or, 8-nearest neighbor sense.

[8]The reason is that the n^{th} skeleton subset of a union of disconnected sets with respect to a convex structuring element is the union of the n^{th} skeleton subsets of the disconnected sets, and the MSF of $nH \oplus \{z\}$ with respect to H is equal to n at z and zero elsewhere.

ble to make Maximum Likelihood decisions between DRBRS models of different intensities (but otherwise identical), based solely on the size of the skeleton. As it turns out, this is a move in the right direction. In fact, the important statistic is the size of a superset of the skeleton. This is the subject of the following theorem.

Theorem 3.1 *Consider the simple vs. simple hypothesis testing problem:*

$$H_0\text{: } X \sim (p_0, H, f_R)\text{-}\mathbf{DRBRS}$$

$$\textit{vs. } H_1\text{: } X \sim (p_1, H, f_R)\text{-}\mathbf{DRBRS}$$

where p_0, p_1 are constants, both in $(0,1)$, $p_1 > p_0$, and $f_R(r)$ (the common size distribution) is zero outside $\left\{\underline{R}, \cdots, \bar{R}\right\}$, where $\underline{R} \geq 0$. Define

$$\gamma(X) \triangleq \frac{|(X^c \oplus \underline{R}H^s)^c|}{|B|}$$

Let K be the observation, and let $P_0(X = K)$, $P_1(X = K)$ denote the probability of the observation under the null and alternative hypothesis, respectively. If

$$\gamma(K) < l(p_0, p_1) \triangleq \frac{log(1-p_1) - log(1-p_0)}{log(p_0(1-p_1)) - log(p_1(1-p_0))}$$

then $P_0(X = K) > P_1(X = K)$.

Proof:

Let L be a realization of the germ points which can give rise to the observation, K. The probability of this realization under p_0 is $Pr_0(L) = p_0^{|L|}(1-p_0)^{|B|-|L|}$, whereas under p_1 it is $Pr_1(L) = p_1^{|L|}(1-p_1)^{|B|-|L|}$. It is easy to see that

$$Pr_0(L) > Pr_1(L) \Longleftrightarrow \frac{|L|}{|B|} < l(p_0, p_1)$$

But, any L which can give rise to K necessarily satisfies

$$L \subseteq (K^c \oplus \underline{R}H^s)^c$$

thus

$$\frac{|L|}{|B|} \leq \gamma(K) < l(p_0, p_1), \text{ by assumption}$$

Therefore, $Pr_0(L) > Pr_1(L)$, uniformly over all L which can give rise to K. Hence, since the two models have the same primary grain and size distribution, we conclude that $P_0(X = K) > P_1(X = K)$. □

Some remarks are in order. This theorem states that if $\gamma(K) < l(p_0, p_1)$, then we can "safely" decide in favor of the null hypothesis, H_0, in the sense that our decision coincides with the ML decision. That is, if $\gamma(K) < l(p_0, p_1)$, then we make a computationally cheap decision, which also happens to be statistically sound. This technique may have potential for application in the automated screening of cell samples, where the alternative hypothesis corresponds to an abnormally high average number of cells per unit area. Then, most of the observed samples can be classified with minimal effort, whereas the few samples which do not meet the criterion of theorem 2 can be examined in greater detail, by either a machine, or a human expert.

By symmetry, if the size of the smallest L which can give rise to the observation K satisfies $\frac{|L|}{|B|} > l(p_0, p_1)$, then we can conclude that $P_1(X = K) > P_0(X = K)$, and we can "safely" decide in favor of the alternative hypothesis, H_1. However, there exists no known efficient (polynomial time) algorithm which

can determine the size of the smallest L which can give rise to K. The only way we know how to do this is by exhaustive search, whose complexity is exponential in $|K|$. In this case we might as well compute the exact likelihood of the observation under each one of the two hypotheses, and compare.

The behavior of l as a function of p_0, p_1 is of considerable interest, because it determines the rate at which the above theorem can be used to simplify ML decisions. It can be shown that l is *roughly* halfway between p_0 and p_1, i.e.,

$$l(p_0, p_1) \cong \frac{p_0 + p_1}{2}$$

Let $E_0\gamma(X)$ denote the expectation of $\gamma(X)$ under p_0. It can be proven, along the lines of proof of proposition 1, that $E_0\gamma(X) \geq p_0$. In practice, we can estimate both $E_0\gamma(X)$ and the standard deviation of $\gamma(X)$ under p_0, from a set of training data. Denote these by $\hat{\gamma}_0$, and $\hat{\sigma}_0$, respectively. Then, as a rule of thumb, if $p_1 > 2(\hat{\gamma}_0 + \hat{\sigma}_0) - p_0$, we will be able to classify most of the "normal" samples optimally, and with minimal effort.

Returning to the skeleton idea, we remark that it can also be used for parameter estimation purposes. Under a strong separability condition, we can construct strongly consistent and efficient estimators of the associated model parameters[9].

Proposition 3.2 *Let* $\{X_i\}_{i=1}^N$ *be a sequence of* N *iid* (λ_s, H, f_R)*-DRBRS's, on*

[9]See the appendix for a brief overview of some relevant aspects of ML parameter estimation.

B, such that for each X_i, $i = 1, \ldots, N$

$$X_i = \bigcup_{j=1,2,\ldots} G_{ij} \oplus \{y_{ij}\}$$

the translated primary grains $\{G_{ij} \oplus \{y_{ij}\}\}_{j=1,2,\ldots}$ *are almost surely disconnected*[10] *(in the 8-nearest neighbor sense) and contained in* B, *but* $\lambda_s(z)$ *is not identically equal to zero for all* $z \in B$. *For brevity, let* $X^{(N)}$ *denote the sequence* $\{X_i\}_{i=1}^N$, *and* 1_K *be the indicator function of the set* K. *The sequence of empirical estimators of* $\lambda_s(z)$, *and* $f_R(n)$ *on the basis of* X^N, *defined for each* $N \in \mathcal{Z}_+^*$ *by*

$$\widehat{\lambda_s(z)}(X^{(N)}) \triangleq \frac{1}{N} \sum_{i=1}^{N} 1_{SK(X_i)}(z), \quad \forall z \in B$$

$$\widehat{f_R(n)}(X^{(N)}) \triangleq \begin{cases} \dfrac{\sum_{i=1}^{N} |S_n(X_i)|}{\sum_{i=1}^{N} |SK(X_i)|} & , n = 0, \ldots, \bar{R} - 1 \\ 1 - \sum_{m=0}^{\bar{R}-1} \widehat{f_R(m)}(X^{(N)}) & , n = \bar{R} \end{cases}$$

respectively, is a strongly consistent sequence of Minimum Variance Unbiased Estimators (MVUE's) of the intensity field $\lambda_s(z)$, $z \in B$, *and the size distribution* $f_R(n)$, $n = 0, 1, \cdots, \bar{R}$ *on the basis of* $X^{(N)}$. *Furthermore, it is a sequence of Maximum Likelihood Estimators (MLE's) of the intensity field and the size distribution on the basis of* $X^{(N)}$.

Proof:

If the translated primary grains, $\{G_{ij} \oplus \{y_{ij}\}\}_{j=1,2,\ldots}$ of X_i are a.s. disconnected

[10]This is an example of a so-called *hard-core* model [65]

and contained in B, then the MSF of X_i is a.s. the same as the underlying marked point process. The reason is that the n^{th} skeleton subset of a union of disconnected sets with respect to a convex structuring element is the union of the n^{th} skeleton subsets of the disconnected sets, and the MSF of $nH \oplus \{z\}$ with respect to H is equal to n at z and zero elsewhere. Given the sequence of underlying independent marked point process realizations, by spatial independence, the problem of estimating the intensity function of the point field reduces to the problem of individually estimating the value of the intensity function at each point, based on the corresponding sequence of N iid binary-valued observations indicating the presence or absence of a germ at the given location. Invoking lemma 7.1 of the appendix, the result is established for the proposed sequence of intensity field estimators. By independence of the marks, the problem of estimating the size distribution reduces to the problem of estimating the marginal pmf of an iid sequence of M-ary valued r.v.'s (here M is equal to $\bar{R}+1$). Invoking lemma 7.2 of the appendix, the result is established for the proposed sequence of estimators of the primary grain size distribution. □

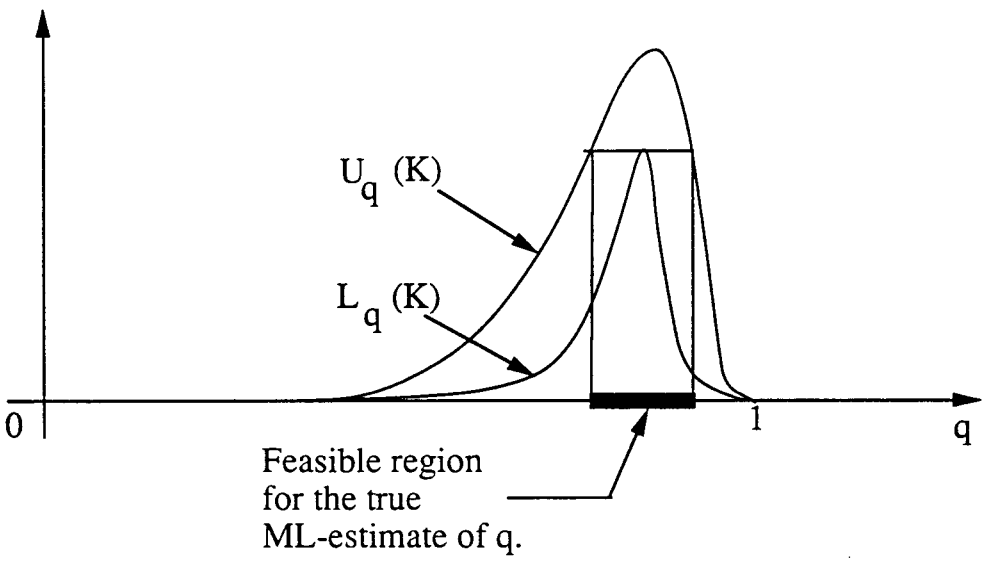


Figure 3.1: The concept of the feasible region.

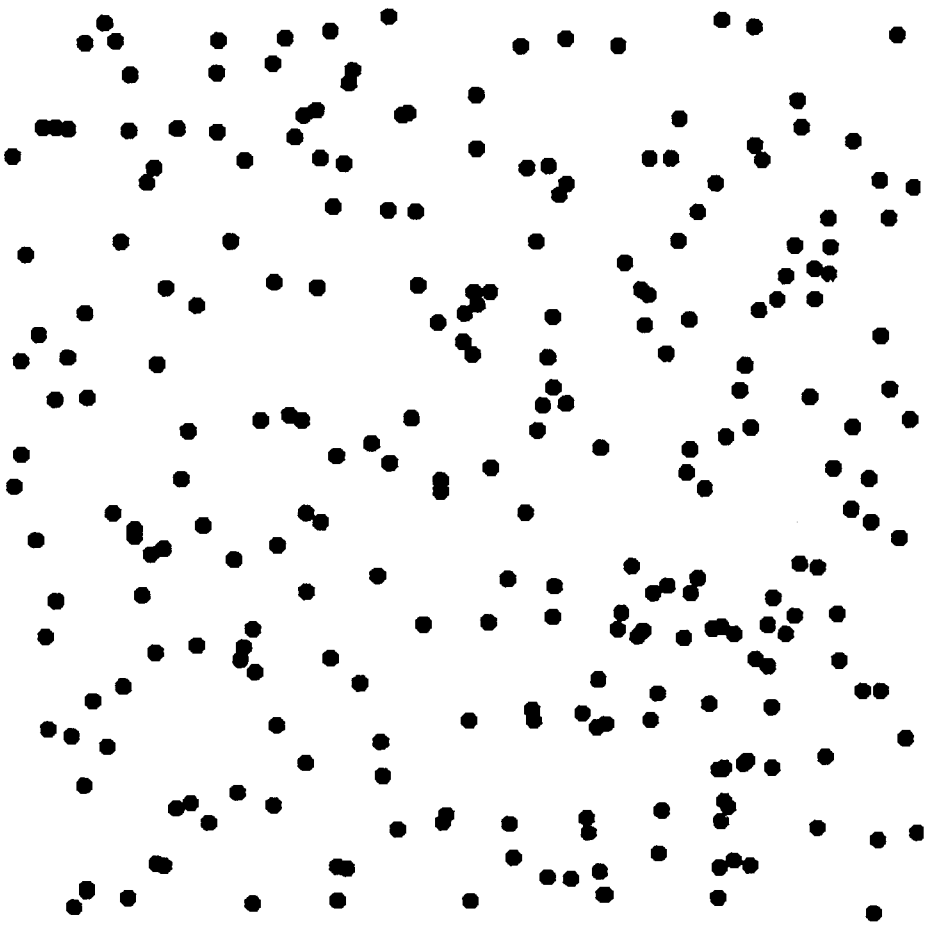


Figure 3.2: Realization of a Boolean model of constant intensity and fixed primary grain.

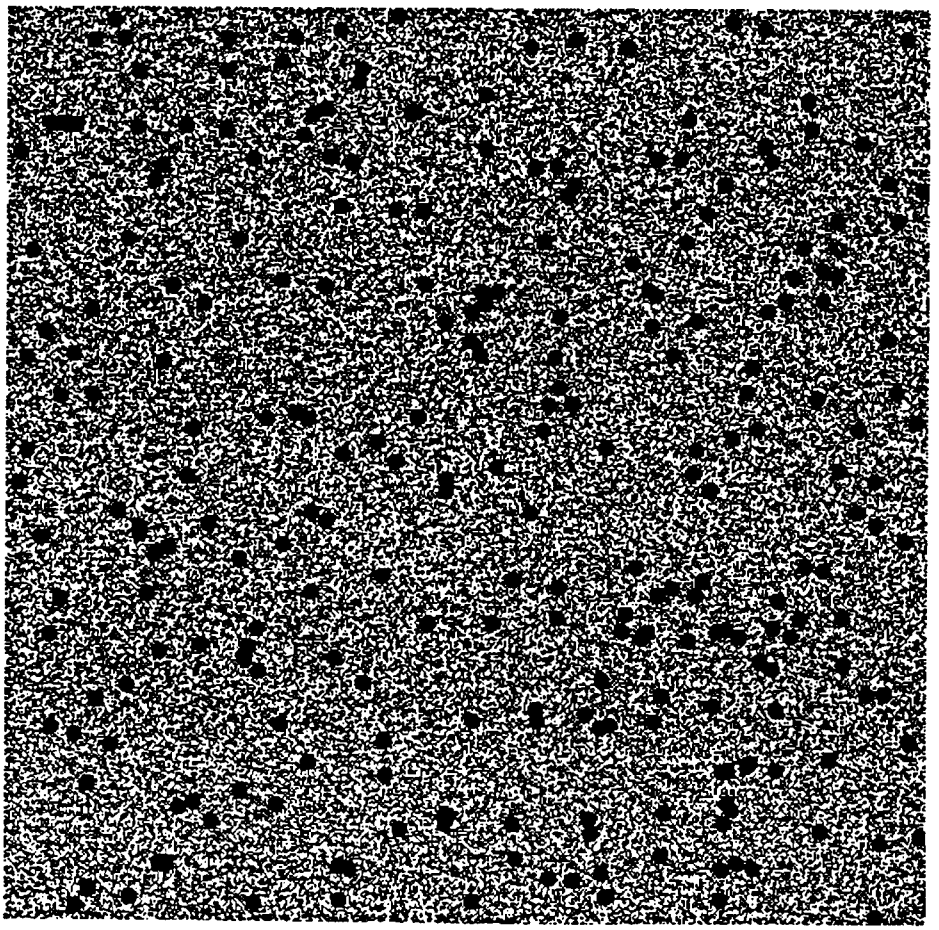


Figure 3.3: The realization of figure 3.2, corrupted by iid union noise of intensity 0.5

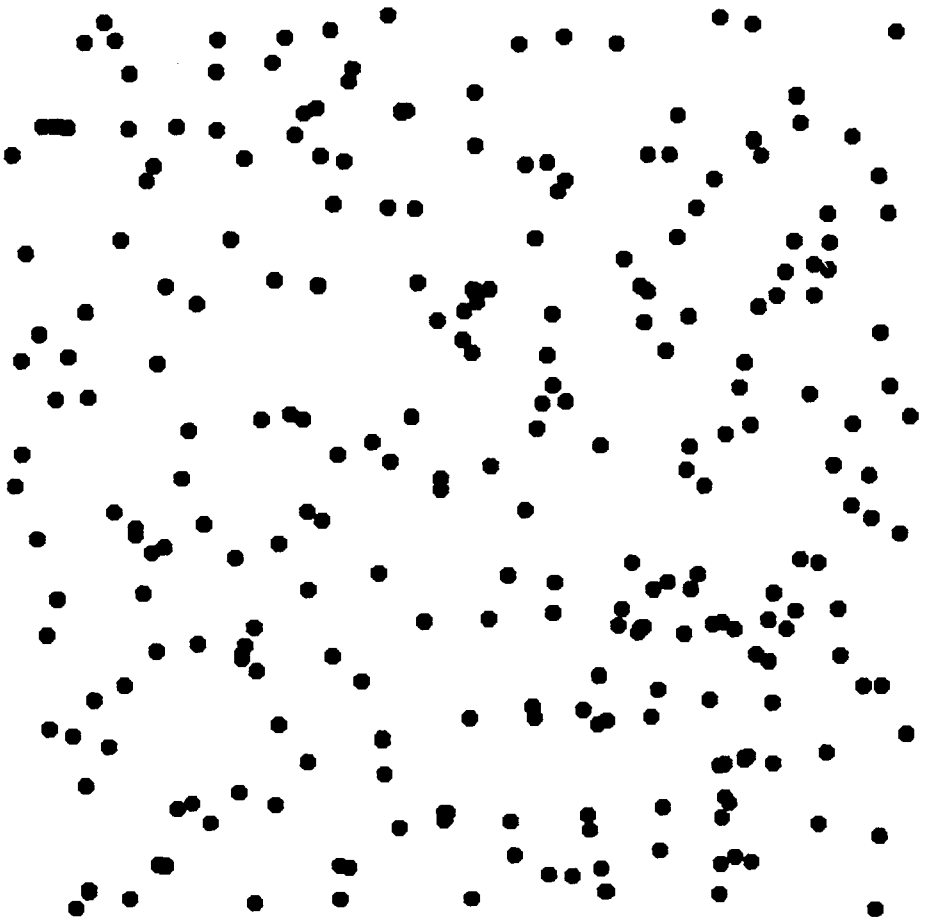


Figure 3.4: The ML estimate of the signal of figure 3.2, on the basis of the observation depicted in figure 3.3.

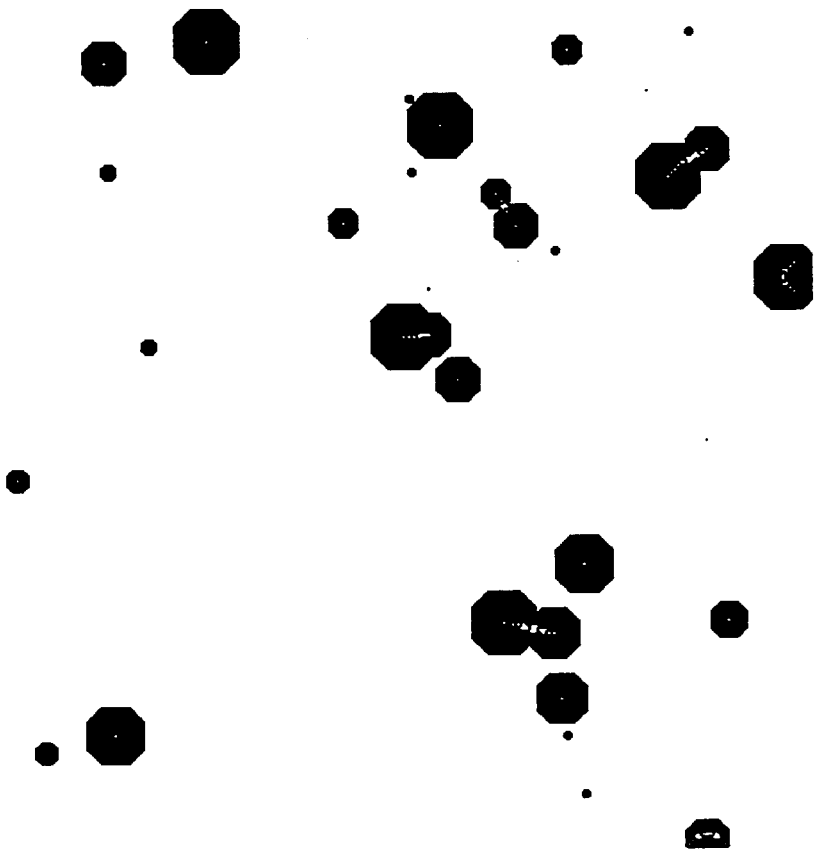


Figure 3.5: Realization of a DRBRS and its skeleton. The skeleton points are the white points highlighted within the primary grains.

CHAPTER
FOUR

OPTIMAL FILTERING

4.1 Introduction

An important problem in digital image processing and analysis is the development of optimal filtering procedures which attempt to restore a binary image ("signal") from its degraded version [55, 17]. Here, the degradation mechanism usually models the combined effect of two distinct types of distortion, namely, image object obscurations because of clutter, and sensor/channel noise. It is typically assumed that the degraded image can be accurately modeled as the union of the uncorrupted binary image with an independent realization of the noise process, which is a binary image itself [27]. This degradation model is known as the union noise model. Other models exist, such as the intersection noise model, and the union/intersection noise model, which are defined in the obvious fashion. The assumption of independence is crucial for the theoretical analysis of optimal filters, and it is plausible in many practical situations. These models are rather general, in that they can be tailored to describe most popular

types of signal-independent noise, e.g. salt-and-pepper noise (also known as Binary Symmetric Channel (BSC) transmission noise), burst channel errors, noise with a geometric structure [27], occlusion, etc.

This part of our research has been largely motivated by the works of Haralick, Dougherty, and Katz [27], and Schonfeld and Goutsias [55]. Their approach is model-based, in that they assume specific probabilistic/geometrical models that govern the behavior of both signal and noise "patterns", i.e. the elementary geometrical primitives from which the signal and noise images are constructed. Haralick, Dougherty, and Katz assume that the signal and noise patterns are "non-interfering" with one another, meaning that each signal or noise pattern is disconnected from all remaining signal and noise patterns. Schonfeld and Goutsias make a stronger assumption concerning the separability of noise patterns. These assumptions are reasonable if the image is sparse, i.e. the signal and noise patterns are most likely to remain uncluttered. Haralick, Dougherty, and Katz adopt the area of the symmetric set difference between the ideal image and its reconstruction as their choice of distance metric, and work with a union noise model to derive the optimal (in the sense of minimizing the expected distance between the signal and its reconstruction) value of a "size" parameter which determines the optimal filter within a restricted family of Morphological Opening Filters [45, 57, 16]. In their work, the signal and noise patterns are all assumed to be of the same basic shape, and only their size varies. Schonfeld and Goutsias consider Morphological Alternating Sequential Filters (ASF's) [45, 57, 16],

and work with the union/intersection noise model. They adopt an implicit least mean difference "uniform" optimality criterion (i.e., the best filter, within a family of filters, is defined to be the one which minimizes the average (over the family) area of the symmetric set difference between the *output* of the given filter, and those of all remaining filters in the family, for a given class of inputs). They derive the "optimal" ASF by means of minimizing an upper bound on their cost function. Related work can also be found in a series of papers by Dougherty et al. [17, 19, 18].

The work reported in this dissertation focuses on a different viewpoint. As it turns out, by restricting our attention to suitable classes of filtering operations, and uniformly bounded discrete random sets, we can obtain optimal filtering results, *under considerably milder assumptions on the signal and noise patterns*; i.e., results that are applicable for all signal and noise models, under the assumption of mutual independence of the signal and the noise. Specifically, one need not assume that signal and noise patterns are "non-interfering". Furthermore, it is possible to obtain simple, closed characterizations of the optimal filter. The resulting formulas are intuitively appealing, and directly amenable to design and implementation.

Some motivating comments are in order. A fair question to ask is whether it is necessary, for the purposes of filtering, to model binary image data as random sets. We feel that it is, for two reasons. First, this enables us to formulate the optimal filtering problem within a rigorous statistical framework.

Second, random set theory is closely related to Mathematical Morphology. Thus, random set theory allows the simultaneous modeling of two important aspects of binary images: geometric structure, and statistical behavior. In and by itself, neither one of the two can provide a complete summary of the images under consideration. In the terminology of nonlinear filtering, our unified approach allows the joint optimization of both the *syntactic* and the *statistical* properties of a filter structure.

Another important question is how much common ground exists between the optimal filtering problem for discrete random sets, and standard optimal filtering theory for the case of real or vector-valued random variables. To what extent can we translate well established results (e.g., the Orthogonality Principle), in the discrete random set setting? The answer is that the analogy is rather superficial. The major difference is that our problem does not have the nice vector space structure which underscores classical optimal filtering theory. We will explain this in detail.

The rest of this chapter is organized as follows. Section 4.2 contains a formalization of the optimal filtering problem, including a discussion of our choice of distance metric. Some connections with classical optimal filtering theory are also made, and the fundamental differences are pointed out. Section 4.3 contains results on optimal mask filtering, which is motivated by the set-theoretic analysis of optimal filters. Section 4.4 takes on a Morphological filtering approach, which is motivated by the widespread use of Morphological filters, their well-

known shape-preservation properties, and a fresh statistical insight into some "folk theorems" of applied Morphological filtering. Some experimental results are presented in section 4.4.5.

4.2 Formulation of the Optimal Filtering Problem

Let X, N, and Y be DRS's on B. X models the "signal", whereas N models the noise. Let $g : \Sigma(B) \times \Sigma(B) \mapsto \Sigma(B)$ be a mapping that models the degradation (measurability is automatically satisfied here, since the domain of g is finite). The observed DRS is $Y = g(X, N)$. Let $d : \Sigma(B) \times \Sigma(B) \mapsto \mathbf{Z}_+$ be a distance metric between subsets of B. In this context, the optimal filtering problem is to find a mapping $f : \Sigma(B) \mapsto \Sigma(B)$ such that the expected cost (expected error)

$$E(e) \triangleq Ed(X, \widehat{X}), \quad \widehat{X} = f(Y) = f(g(X, N))$$

is minimized, over all possible choices of the mapping ("filter") f. This problem is in general intractable. The main difficulty is the lack of structure on the search space. The family of all mappings $f : \Sigma(B) \mapsto \Sigma(B)$ is just too big. It is common practice to *impose* structure on the search space, i.e. constrain f to lie in $\mathcal{F}$, a suitably chosen subcollection of *admissible* mappings (family of filters), and optimize within this subcollection. The resulting filter is the best among its peers, but it is not guaranteed to be globally optimal.

We adopt the following distance metric (area of the symmetric set difference)[1]

$$d(X, \widehat{X}) = |(X \backslash \widehat{X}) \cup (\widehat{X} \backslash X)|$$

$$= |(X \backslash \widehat{X})| + |(\widehat{X} \backslash X)|$$

$$= |(X \cup \widehat{X}) \backslash (X \cap \widehat{X})|$$

$$= |(X \cup \widehat{X})| - |(X \cap \widehat{X})|$$

where $|\ |$ stands for set cardinality, $\backslash$ stands for set difference, i.e. $X \backslash Y = X \cap Y^c$, and ^{c} stands for complementation with respect to the lattice, B. This distance metric is essentially the *Hamming distance* [48] when $X, \widehat{X}$ are viewed as vectors in $\{0,1\}^{|B|}$. Since the component variables are binary, it can also be interpreted as the square of the L_2 distance of vectors in $\{0,1\}^{|B|}$, i.e., with some abuse of notation,

$$d(X, \widehat{X}) = (X - \widehat{X})^T (X - \widehat{X})$$

where on the left hand side symbols are interpreted as sets, while on the right hand side symbols are interpreted as column vectors in $\{0,1\}^{|B|}$, and ^{T} stands for transpose. In this setting, the *sufficiency part* of the Orthogonality Principle (OP) [48] applies. It states that a sufficient condition for the existence of an $f^* \in \mathcal{F}$ such that

$$E\left[(X - f^*(Y))^T (X - f^*(Y))\right] \leq E\left[(X - f(Y))^T (X - f(Y))\right], \quad \forall f \in \mathcal{F}$$

[1]This distance metric has been previously used; e.g., see [55, 27].

is that

$$E\left[(X - f^*(Y))^T(f^*(Y) - f(Y))\right] = 0, \quad \forall f \in \mathcal{F}$$

However, unlike the case of vectors in R^n, where $\mathcal{F}$ is a vector space over the field of reals (known as the space of *square integrable* estimators), here $\mathcal{F}$ is not a vector space. The proof of the necessity part of the OP strongly depends on $\mathcal{F}$ having a vector space structure. When $\mathcal{F}$ does not have a vector space structure, the key notion is that of *conditional expectation.* Here, however, we run into trouble defining what we mean by conditional expectation of a DRS X given a DRS Y, let alone computing it. Fortunately, it turns out that it is often possible to write down an expression for $Ed(X, \widehat{X})$, and optimize over $\mathcal{F}$ by brute force.

Technically speaking, $d(X, \widehat{X})$ can be considered as a quadratic distance measure when we view $X, \widehat{X}$ as vectors in $\{0,1\}^{|B|}$. From a set-theoretic point of view, $d(X, \widehat{X})$ is clearly not a quadratic distance measure, since it penalizes errors in a linear fashion. However, the squared area of the symmetric set difference (which is a quadratic distance measure in the set-theoretic sense) does not yield useful optimality conditions. This is partly due to the lack of a meaningful and tractable definition for the *expectation of a DRS* X. In the continuous case there exists such a definition. The expectation of a random compact set X can be defined via the expectation of *random selections*, i.e. random vectors which are a.s. contained in X. The expectation of X is defined as the union of expectations of all its random selections. Random selections exist, and the resulting notion of

expectation of a random compact set is well defined. This definition is adapted for the development of strong limit theorems [70, 71, 3, 67, 66, 69, 68, 10]. However, it is not convenient for our purposes. Consider a uniformly bounded DRS X defined on a finite lattice $B \subset \mathbf{Z}^2$. A random selection from X is a random vector taking values in B; however, its expectation is not necessarily a point in B. Thus, the resulting expectation of X will not necessarily be a subset of B. Furthermore, in contrast to the expectation of a random variable or random vector, this notion of expectation of a random compact set (measurable mapping) depends not only on the induced distribution over the space of realizations, but also on the mapping itself! This surprising fact has interesting ramifications [71]. However, in our case, it introduces unnecessary complications. Finally, here we are not interested in asymptotics when the lattice goes to infinity; instead, we focus on filtering. For this we need an alternative definition of expectation. From a quadratic estimation-theoretic point of view[2] a proper *formal* definition of the expectation of a DRS X, would be as follows.

$$EX \triangleq arg\ min_{W \in \Sigma(B)} E\left[d(X,W)\right]^2$$

However, there exist several flaws with this formal definition.

Proposition 4.1

$$arg\ min_{W \in \Sigma(B)} E\left[d(X,W)\right]^2 =$$

[2]Under certain conditions, the expectation of a scalar random variable, x, satisfies $Ex = arg\ min_{t \in \mathbf{R}} E|x-t|^2$.

$$arg\ min_{W\in\Sigma(B)}\Big\{|W|^2+2|W|\sum_{z\in W^c}Pr(z\in X)$$

$$-2|W|\sum_{z\in W}Pr(z\in X)-4\sum_{z\in W^c}\sum_{y\in W}Pr(z\in X\ ,\ y\in X)\Big\}$$

Proof:

Let $C(z)$ be a Boolean proposition, which, for each point $z\in B$, is either true, or false. Define the *indicator function*

$$1(C(z))\triangleq\begin{cases}1 & ,\ if\ C(z)\ is\ true\ at\ z\\ 0 & ,\ otherwise\end{cases}$$

Let *supp* $1(C(z))$ stand for the *support set* of the indicator function $1(C(z))$, i.e. the subset of B where $C(z)$ is true.

$$E\left[d(X,W)\right]^2=E\left[|X\cup W|-|X\cap W|\right]^2=E\left[|X|+|W|-2|X\cap W|\right]^2$$

$$=E|X|^2+|W|^2+2|W|E|X|+4E|X\cap W|^2-4E\left[|X||X\cap W|\right]-4|W|E|X\cap W|$$

Now,

$$E|X|=E\sum_{z\in B}1(z\in X)=\sum_{z\in B}E\ 1(z\in X)=\sum_{z\in B}Pr(z\in X)$$

Similarly,

$$E|X\cap W|=\sum_{z\in W}Pr(z\in X)$$

Also,

$$E|X|^2=E|X||X|=E\sum_{z\in B}1(z\in X)\sum_{y\in B}1(y\in X)$$

$$=E\sum_{z\in B}\sum_{y\in B}1(z\in X)1(y\in X)=E\sum_{z\in B}\sum_{y\in B}1(z\in X,\ y\in X)$$

$$= \sum_{z \in B} \sum_{y \in B} E\ 1(z \in X,\ y \in X) = \sum_{z \in B} \sum_{y \in B} Pr(z \in X,\ y \in X)$$

Similarly,

$$E|X \cap W|^2 = \sum_{z \in W} \sum_{y \in W} Pr(z \in X,\ y \in X)$$

and, finally

$$E\left[|X||X \cap W|\right] = \sum_{z \in B} \sum_{y \in W} Pr(z \in X,\ y \in X)$$

Collecting the terms which depend on W, and rearranging, we obtain the desired formula. □

If we assume that $Pr(z \in X) = p$, $\forall z \in B$, and $Pr(z \in X\ ,\ y \in X) = Pr(z \in X)Pr(y \in X) = p^2$, $\forall\ z,y\ s.t.\ z \neq y$, and $p < 0.5$, then $EX = \emptyset$, regardless of the specific value of p. If $p = 0.5$, then *any* $W \in \Sigma(B)$ will do. However, the single most important problem is that, given a specification of the first and second-order statistics of X, it is not clear how to find an explicit solution to the above minimization problem. On the other hand, the *median of a DRS* X, formally defined as[3]

$$MX \triangleq arg\ min_{W \in \Sigma(B)} Ed(X,W)$$

is much easier to deal with. Although the solution to this latter minimization problem is not (in general) unique, it can be forced to be unique by means of a simple regularization.

[3]Under certain conditions, the median of a scalar random variable, x, satisfies $Mx = arg\ min_{t \in \mathbf{R}} E|x - t|$.

Proposition 4.2

$$arg\ min_{W\in\Sigma(B)} Ed(X,W) =$$

$$arg\ min_{W\in\Sigma(B)} \left\{ \sum_{z\in W^c} T_X(\{z\}) + \sum_{z\in W} (1 - T_X(\{z\})) \right\}$$

Proof:

$$Ed(X,W) = E\left[|X \cup W| - |X \cap W|\right] = E\left[|X| + |W| - 2|X \cap W|\right]$$

$$= E|X| + E|W| - 2E|X \cap W| = E|X| + |W| - 2E|X \cap W|$$

Now,

$$E|X| = E \sum_{z\in B} 1(z \in X) = \sum_{z\in B} E\ 1(z \in X) = \sum_{z\in B} Pr(z \in X)$$

$$= \sum_{z\in B} Pr(X \cap \{z\} \neq \emptyset) = \sum_{z\in B} T_X(\{z\})$$

Similarly,

$$E|X \cap W| = \sum_{z\in W} T_X(\{z\})$$

So that

$$Ed(X,W) = \sum_{z\in W^c} T_X(\{z\}) - \sum_{z\in W} T_X(\{z\}) + |W|$$

Note that $|W|$ can be written as

$$|W| = \sum_{z\in W} 1$$

and the result follows. □

Thus,

$$M_R X \triangleq supp\ 1(1 - T_X(\{z\}) < T_X(\{z\}))$$

is the unique minimum cardinality solution to the minimization problem

$$min_{W \in \Sigma(B)} Ed(X, W)$$

These considerations essentially dictate our choice of distance metric. In terms of the degradation, we assume that N is independent of X, and that the mapping g is given by

$$g(X, N) = X \cup N \quad (union\ noise\ model)$$

or,

$$g(X, N) = X \cap N \quad (intersection\ noise\ model)$$

Although we will be mainly concerned with either union *or* intersection noise, on one occasion we will allow g to be a mapping from $\Sigma(B) \times \Sigma(B) \times \Sigma(B)$ to $\Sigma(B)$

$$g(X, N_1, N_2) = (X \cap N_1) \cup N_2 \quad (union/intersection\ noise\ model)$$

where X, N_1, N_2 will be assumed to be mutually independent.

4.3 Optimal Mask Filters

In the case of union noise, a suitable class of filters is

$$f(Y) = f_W(Y) = Y \cap W = (X \cup N) \cap W,\ for\ some\ W \in \Sigma(B)$$

Similarly, in the case of intersection noise, we can consider the following class of filters

$$f(Y) = f^W(Y) = Y \cup W = (X \cap N) \cup W,\ for\ some\ W \in \Sigma(B)$$

Finally, in the case of union/intersection noise, we can consider

$$f(Y) = f_{W_2}^{W_1}(Y) = (Y \cap W_2) \cup W_1 =$$

$$((((X \cap N_1) \cup N_2) \cap W_2) \cup W_1), \text{ for some } W_1, W_2, \text{ both in } \Sigma(B)$$

These filters arise naturally in this setting; we will call them *mask filters*. For example, in the case of union noise the optimal filter should retain a subset of the observation points and reject the rest; this should be done in some sort of statistically optimal fashion. This is achieved by intersecting the observation with a suitably chosen mask. In the simplest case the mask is fixed; in the adaptive case it is allowed to depend on the observation. Note that, under the given degradation models, unconstrained adaptive mask filtering is the most general filtering structure that we can consider. We will show that explicit optimization is possible, under some restrictions on the adaptation strategy.

Let us first consider fixed-mask filtering. Here, we only work with the union/intersection noise model. The other two noise models are special cases. We have the following proposition.

Proposition 4.3 *Under the expected symmetric set difference measure, an optimal fixed pair of masks, (W_1, W_2), is given by*

$$W_2 = supp\ 1\,(T_X(\{z\}) > max(T_1(\{z\}), T_2(\{z\})))$$

$$W_1 = supp\ 1\,(T_2(\{z\}) \leq min(T_X(\{z\}), T_1(\{z\})))$$

whereas, the associated minimum expected cost, achieved by such an optimal pair of masks, is

$$E(e^*) = \sum_{z \in B} min\,(T_X(\{z\}), T_1(\{z\}), T_2(\{z\}))$$

with

$$T_1(\{z\}) = T_X(\{z\})\,(1 - T_{N_1}(\{z\}))\,(1 - T_{N_2}(\{z\})) + (1 - T_X(\{z\}))\,T_{N_2}(\{z\})$$

and

$$T_2(\{z\}) = T_X(\{z\})\,(1 - T_{N_1}(\{z\})) + (1 - T_X(\{z\}))$$

Proof:

Without loss of generality, we may assume that $W_1 \subseteq W_2$, since it makes no sense removing points from the observation, only to reinstate them at the next filtering step. After some manipulation,

$$E(e) = Ed(X, \widehat{X}) = E\,|X \cap (N_1^c \cup W_2^c) \cap [(N_2^c \cap W_1^c) \cup W_2^c]|$$

$$+E\,|(X^c \cap N_2 \cap W_2) \cup (X^c \cap W_1)|$$

A crucial observation here is that

$$E|X| = E \sum_{z \in B} 1(z \in X) = \sum_{z \in B} E\ 1(z \in X)$$

$$= \sum_{z \in B} Pr(z \in X) = \sum_{z \in B} Pr(X \cap \{z\} \neq \emptyset) = \sum_{z \in B} T_X(\{z\})$$

Consider the first term of the expected error

$$E\,|X \cap (N_1^c \cup W_2^c) \cap [(N_2^c \cap W_1^c) \cup W_2^c]|$$

$$= \sum_{z \in B} Pr\,(z \in X \cap (N_1^c \cup W_2^c) \cap [(N_2^c \cap W_1^c) \cup W_2^c])$$

$$(by\ indep.\ of\ X, N_1, N_2)$$

$$= \sum_{z \in B} Pr\,(z \in X)\, Pr\,(z \in N_1^c \cup W_2^c)\, Pr\,(z \in (N_2^c \cap W_1^c) \cup W_2^c)$$

$$= \sum_{z \in B} T_X(\{z\})\,[1(z \in W_2^c) + 1(z \in W_2)\,(1 - T_{N_1}(\{z\}))] \times$$

$$[1(z \in W_2^c) + 1(z \in W_2)1(z \in W_1^c)\,(1 - T_{N_2}(\{z\}))]$$

Next, consider the second term of the expected error

$$E\,|(X^c \cap N_2 \cap W_2) \cup (X^c \cap W_1)|$$

$$= \sum_{z \in B} Pr\,(z \in (X^c \cap N_2 \cap W_2) \cup (X^c \cap W_1))$$

$$= \sum_{z \in B} Pr\,(z \in X^c \cap [(N_2 \cap W_2) \cup W_1])$$

$$= \sum_{z \in B} Pr\,(z \in X^c, z \in (N_2 \cap W_2) \cup W_1)$$

$$(by\ independence\ of\ X, N_2)$$

$$= \sum_{z \in B} Pr\,(z \in X^c)\, Pr\,(z \in (N_2 \cap W_2) \cup W_1)$$

$$= \sum_{z \in B} (1 - T_X(\{z\}))\,[1(z \in W_1) + 1(z \in W_1^c)1(z \in W_2)T_{N_2}(\{z\})]$$

Therefore, the overall expression for the expected cost becomes

$$E(e) = Ed(X, \widehat{X})$$

$$= \sum_{z \in B} \{T_X(\{z\})\,[1(z \in W_2^c) + 1(z \in W_2)\,(1 - T_{N_1}(\{z\}))] \times$$

$$[1(z \in W_2^c) + 1(z \in W_2)1(z \in W_1^c)\,(1 - T_{N_2}(\{z\}))] +$$

$$(1 - T_X(\{z\}))\,[1(z \in W_1) + 1(z \in W_1^c)1(z \in W_2)T_{N_2}(\{z\})]\}$$

Consider the term in curly braces. As we have mentioned before, $W_1 \subseteq W_2$. Therefore, for each $z \in B$, we have the following three choices

$$(i)\ z \in W_1^c,\ z \in W_2^c,\ or\ (ii)\ z \in W_1^c,\ z \in W_2,\ or\ (iii)\ z \in W_1,\ z \in W_2$$

In case (i) the term in curly braces is equal to $T_X(\{z\})$, in case (ii) it is equal to $T_1(\{z\})$, and in case (iii) it is equal to $T_2(\{z\})$. The result follows. □

If the first-order statistics (pixel hitting probabilities) of the signal and noise DRS's are spatially invariant, then, obviously, the optimal pair of masks is either $(\emptyset, B)$, or $(\emptyset, \emptyset)$, or (B, B). In this case, fixed-mask filtering is not appropriate. It is exactly when the signal and/or the noise statistics are highly nonstationary (meaning not even first-order stationary) that this filtering approach makes sense. In such a highly nonstationary environment, traditional shift-invariant neighborhood filtering operations (e.g. local mean, median, order statistics) typically fail to provide adequate filtering, and their optimization is very difficult. On the other hand, the optimization of the masks is based on simple statistics, which can be efficiently estimated from training data. A potentially big gain in quality of restoration rests exactly with proper use of the nonstationary nature of the signal and/or the noise.

An obvious drawback of fixed-mask filtering is that it does not exploit the autocorrelation structure of the signal and the noise. Furthermore, it is nonadaptive. Whenever higher-order statistics are available, we would like to use

them. We would also like to allow for an adaptation of the mask using information extracted from the given input. Adaptive mask filtering fits both bills. The trade-off is an increase in design/implementation complexity.

Let us first consider the case of union noise. Assume that we are given the degraded data $Y = X \cup N = K$. One adaptation strategy is to incorporate this information into the cost function. This is done by considering the *conditional* expectation of $d(X, \widehat{X})$, conditioned on the given information. However, this does not lead to a closed-form solution for the optimal filter. The reason is that the minimization of this conditional expectation requires the explicit computation of a pseudo-convolution of likelihoods on the lattice of realizations. This computation is in general intractable. Instead, we can condition on part of the available information. This corresponds to minimizing the expected error over a wider collection of events than what is necessary (and optimal). The trade-off is in terms of tractability. If we condition on the event $X \cup N \subseteq K$, i.e. $(X \cup N) \cap K^c = \emptyset$, then we can work out closed-form expressions for the optimal filter and the associated minimum error. In what follows E denotes conditional expectation, conditioned on $(X \cup N) \cap K^c = \emptyset$.

Proposition 4.4 *Given that $X \cup N \subseteq K$, an optimal intersection mask, W, for filtering out the noise component, N, is given by*

$$W = K \cap supp\ 1\left([1 - T_X(K^c \cup \{z\})]\,[T_N(K^c \cup \{z\}) - T_N(K^c)] \leq \right.$$

$$\left. [T_X(K^c \cup \{z\}) - T_X(K^c)]\,[1 - T_N(K^c)]\right)$$

The corresponding minimum cost achieved by such an optimal choice of W is given by[4]

$$E(e^*) = \frac{1}{(1 - T_X(K^c))(1 - T_N(K^c))} \times$$

$$\sum_{z \in K} min \{[1 - T_X(K^c \cup \{z\})] [T_N(K^c \cup \{z\}) - T_N(K^c)],$$

$$[T_X(K^c \cup \{z\}) - T_X(K^c)] [1 - T_N(K^c)]\}$$

Proof:

The total expected error is

$$E|X \cap W^c| + E|N \cap W \cap X^c|$$

Now,

$$E|X \cap W^c| = E \sum_{z \in B} 1(z \in X \cap W^c)$$

$$= \sum_{z \in B} E1(z \in X \cap W^c) = \sum_{z \in B} Pr(z \in X \cap W^c \mid (X \cup N) \cap K^c = \emptyset)$$

$$= \sum_{z \in K} Pr(z \in X \cap W^c \mid (X \cup N) \cap K^c = \emptyset)$$

$$= \sum_{z \in K} \frac{Pr(z \in X \cap W^c, (X \cup N) \cap K^c = \emptyset)}{Pr((X \cup N) \cap K^c = \emptyset)}$$

Observe that

$$Pr((X \cup N) \cap K^c = \emptyset) = Pr(X \cap K^c = \emptyset, N \cap K^c = \emptyset)$$

$$(by\ independence\ of\ X, N) \ = (1 - T_X(K^c))(1 - T_N(K^c))$$

[4]We assume that $Pr(X \cup N \subseteq K) > 0$. Note that this, in turn, implies that $T_X(K^c) < 1$, and $T_N(K^c) < 1$.

also

$$Pr\,(z \in X \cap W^c, (X \cup N) \cap K^c = \emptyset)$$

$$= Pr\,(X \cap W^c \cap \{z\} \neq \emptyset, X \cap K^c = \emptyset, N \cap K^c = \emptyset)$$

$$(by\ independence\ of\ X, N)$$

$$= Pr\,(X \cap (W^c \cap \{z\}) \neq \emptyset, X \cap K^c = \emptyset)\, Pr\,(N \cap K^c = \emptyset)$$

$$= (T_X(K^c \cup (W^c \cap \{z\})) - T_X(K^c))\,(1 - T_N(K^c))$$

Therefore, the first term of the expected cost becomes

$$\sum_{z \in K} \frac{T_X(K^c \cup (W^c \cap \{z\})) - T_X(K^c)}{1 - T_X(K^c)}$$

For the second term of the expected cost

$$E|N \cap W \cap X^c| = \sum_{z \in B} E1\,(z \in N \cap W \cap X^c)$$

$$= \sum_{z \in B} Pr\,(z \in N \cap W \cap X^c \mid (X \cup N) \cap K^c = \emptyset)$$

$$= \sum_{z \in K} Pr\,(z \in N \cap W \cap X^c \mid (X \cup N) \cap K^c = \emptyset)$$

$$= \sum_{z \in K} \frac{Pr\,(z \in N \cap W \cap X^c, (X \cup N) \cap K^c = \emptyset)}{Pr\,((X \cup N) \cap K^c = \emptyset)}$$

We have seen that the denominator is equal to

$$(1 - T_X(K^c))(1 - T_N(K^c))$$

Whereas the nominator

$$Pr\,(z \in N \cap W \cap X^c, (X \cup N) \cap K^c = \emptyset)$$

$$= Pr\left(\left(\{z\} \cap W\right) \cap N \cap X^c \neq \emptyset, (X \cup N) \cap K^c = \emptyset\right)$$

$$= Pr\left(\left(\{z\} \cap W\right) \in N, \left(\{z\} \cap W\right) \in X^c, X \cap K^c = \emptyset, N \cap K^c = \emptyset\right)$$

$$= Pr\left(X \cap K^c = \emptyset, \left(\{z\} \cap W\right) \in X^c\right) Pr\left(N \cap K^c = \emptyset, \left(\{z\} \cap W\right) \in N\right)$$

$$= Pr\left(X \cap K^c = \emptyset, X \cap \left(\{z\} \cap W\right) = \emptyset\right) Pr\left(N \cap K^c = \emptyset, N \cap \left(\{z\} \cap W\right) \neq \emptyset\right)$$

$$= Pr\left(X \cap \left(K^c \cup \left(\{z\} \cap W\right)\right) = \emptyset\right) Pr\left(N \cap K^c = \emptyset, N \cap \left(\{z\} \cap W\right) \neq \emptyset\right)$$

$$= \left[1 - T_X(K^c \cup \left(\{z\} \cap W\right))\right] \left[T_N(K^c \cup \left(\{z\} \cap W\right)) - T_N(K^c)\right]$$

Therefore, the second term of the expected cost becomes

$$\sum_{z \in K} \frac{\left[1 - T_X(K^c \cup \left(\{z\} \cap W\right))\right] \left[T_N(K^c \cup \left(\{z\} \cap W\right)) - T_N(K^c)\right]}{(1 - T_X(K^c))(1 - T_N(K^c))}$$

and the overall expected cost becomes

$$E(e) = \frac{1}{(1 - T_X(K^c))(1 - T_N(K^c))} \times$$

$$\sum_{z \in K} \{\left[1 - T_X(K^c \cup \left(\{z\} \cap W\right))\right] \left[T_N(K^c \cup \left(\{z\} \cap W\right)) - T_N(K^c)\right] +$$

$$\left[T_X(K^c \cup \left(\{z\} \cap W^c\right)) - T_X(K^c)\right] \left[1 - T_N(K^c)\right]\}$$

From which the claimed results follow by inspection. □

Observe how information about the higher-order statistics of the signal and the noise is incorporated into the filter structure, by means of the capacity functionals of the signal and the noise. Note that the minimum cost achieved by an optimal choice of W is not necessarily increasing in K; in the expression for the minimum cost, we can show that the sum is increasing in K, but the normalizing factor, $1/((1-T_X(K^c))(1-T_N(K^c)))$, is decreasing in K. Thus, the

tightest possible K (i.e., the observation itself) is not necessarily the best choice. However, we have experimented with this particular choice with satisfactory results.

A simple simulation experiment is presented in figures 4.1,4.2. Figure 4.1 depicts a realization of a DRBRS model ("the signal") of constant intensity, and deterministic primary grain, corrupted by iid union noise (an independent realization of a Bernoulli lattice process of constant intensity). Figure 4.2 depicts the restored image, obtained by applying the optimal adaptive mask filter of proposition 4.4 to the noisy observation (K was taken to be the observation itself). To the trained eye, the restored image appears to be the Morphological opening of the observed image, using the primary grain of the signal as structuring element. This is, indeed, the case. Let $1-q_X$, $1-q_N$ be the intensities of the DRBRS, and the iid union noise, respectively. We will show that if $q_N \geq 2-q_X^{-1}$ (i.e., the signal is "strong" relative to the noise), and K is taken to be the observation itself, then the optimal adaptive mask filter of proposition 4.4 is exactly the Morphological opening of the observation, using the primary grain of the signal as structuring element. The capacity functional of the DRBRS X is simply given by

$$T_X(K') = 1 - q_X^{|K' \oplus H^s|}$$

where H is the (deterministic) primary grain of the DRBRS X. The capacity

functional of the noise is given by

$$T_N(K') = 1 - q_N^{|K'|}$$

After some manipulations, the expression for the optimal W becomes

$$W = \left\{ z \in K \mid 2 - q_N \leq q_X^{-|H^s|+|(K^c \oplus H^s) \cap (H^s)_z|} \right\}$$

which means that we "pass" all $z \in K$ which satisfy this condition and reject the rest. It can be readily seen that if

$$2 - q_N \leq q_X^{-1} \; (\Longleftrightarrow q_N \geq 2 - q_X^{-1})$$

then

$$2 - q_N \leq q_X^{-n}, \quad n = -1, \cdots, -|H^s|$$

and, therefore, we pass all $z \in K$ except for those which satisfy

$$|(K^c \oplus H^s) \cap (H^s)_z| = |H^s| \Longleftrightarrow (H^s)_z \subseteq K^c \oplus H^s \Longleftrightarrow$$

(modulo some edge effects)

$$z \in (K^c \oplus H^s) \ominus H = K^c \bullet H = (K \circ H)^c$$

Hence, modulo some edge effects, we pass $K \backslash (K \circ H)^c = K \cap (K \circ H)$, i.e., since $K \circ H \subseteq K$, the filtered image is $K \circ H$, where K is the observation, and H is the primary grain of the signal. This identification is important for two reasons. First, we have seen (cf. lemma 3.2) that, under our suppositions for the signal and the noise, the Morphological opening is the Maximum Likelihood estimator

of the signal, based on the noisy observation. Second, the Morphological opening is generally considered a good choice for the given filtering problem (e.g. see [27, 17, 16]).

In a similar manner, we can obtain the following generalization. Consider the case of two DRBRS's X, N, of constant intensities and deterministic primary grains, $1 - q_X$, $1 - q_N$, and H_X, H_N, respectively. Then

$$T_X(K') = 1 - q_X^{|K' \oplus H_X^s|}$$

and

$$T_N(K') = 1 - q_N^{|K' \oplus H_N^s|}$$

After some manipulations, the expression for the optimal mask, W, in proposition 4.4, becomes

$$W = \left\{ z \in K \mid 2 - q_N^{|H_N^s| - |(K^c \oplus H_N^s) \cap (H_N^s)_z|} \leq q_X^{-|H_X^s| + |(K^c \oplus H_X^s) \cap (H_X^s)_z|} \right\}$$

from which it is easy to see that if

$$2 - q_N^{|H_N^s|} \leq q_X^{-1} \ (\Longleftrightarrow q_N^{|H_N^s|} \geq 2 - q_X^{-1})$$

then, modulo edge effects, and taking K to be the observation itself, the optimal adaptive mask filter of proposition 4.4 reduces to the Morphological opening of the observation by H_X, the (deterministic) primary grain of the signal DRBRS X. This is not anymore guaranteed to be the ML estimator of X on the basis of the observation; However, it is widely believed to be a good estimator (e.g. see

[27, 17, 16]). For example, if $|H_N| < |H_X|$, then opening by H_X will eliminate all instances of isolated noisy patterns.

An example is given in figures 4.3, 4.4. Figure 4.3 depicts a realization of the observable DRS, $X \cup N$, where H_X, H_N were taken to be a discrete hexagon of size 12, and a discrete square of size 10, respectively. Figure 4.4 depicts the opening of the DRS realization of figure 4.3, using H_X as structuring element.

This somewhat surprising identification of the optimal adaptive mask filter of proposition 4.4 with the Morphological opening filter is rather interesting. We have started with the objective of optimizing the statistical behavior of a mask filter structure, and ended up with a Morphological filter, which is the intuitively "obvious" choice from the viewpoint of syntactical optimization. This reflects the ability of the statistical optimization procedure to pick up the morphological structure of the signal and the noise, and, in effect, take both statistical and syntactical properties into consideration. This is the first example of such a joint optimization. We will see more of it as we move on. Note that this identification also provides some independent corroborating evidence of the usefulness of optimal adaptive mask filters.

The case of intersection noise can be addressed by appealing to duality. One can simply take the complement of all the sets and operations involved, and apply the result which has been obtained for the case of union noise. This is clear, because

$$d(X_1, X_2) = d(X_1^c, X_2^c)$$

and

$$((X \cap N) \cup W)^c = (X^c \cup N^c) \cap W^c$$

Hence, by conditioning on the event $X^c \cup N^c \subseteq K^c$, i.e. $(X^c \cup N^c) \cap K = \emptyset$, we obtain the following result:

Proposition 4.5 *Given that $X^c \cup N^c \subseteq K^c$ an optimal union ("fill") mask, W, for filtering out the intersection noise component, N, is specified by*

$$W^c = K^c \cap supp\ 1\left([1 - T_{X^c}(K \cup \{z\})]\,[T_{N^c}(K \cup \{z\}) - T_{N^c}(K)] \leq\right.$$

$$\left.[T_{X^c}(K \cup \{z\}) - T_{X^c}(K)]\,[1 - T_{N^c}(K)]\right)$$

The corresponding minimum cost achieved by such an optimal choice of W is given by[5]

$$E(e^*) = \frac{1}{(1 - T_{X^c}(K))(1 - T_{N^c}(K))} \times$$

$$\sum_{z \in K^c} min\left\{[1 - T_{X^c}(K \cup \{z\})]\,[T_{N^c}(K \cup \{z\}) - T_{N^c}(K)]\,,\right.$$

$$\left.[T_{X^c}(K \cup \{z\}) - T_{X^c}(K)]\,[1 - T_{N^c}(K)]\right\}$$

4.4 Optimal Morphological Filters

Morphological filtering is probably one of the most successful applications of the theory of Mathematical Morphology. For convenience, let us recall some important properties of Morphological operators (see section 2.3.1 for definitions). Opening and closing are *idempotent (stable)* operators in the sense that

[5]Again, we assume that $Pr(X^c \cup N^c \subseteq K^c) > 0$, and this implies that $T_{X^c}(K) < 1$, and $T_{N^c}(K) < 1$. Also, similar remarks hold here regarding the best choice of K^c.

$(Y \circ W) \circ W = Y \circ W$, and $(Y \bullet W) \bullet W = Y \bullet W$. A set Y is said to be (Morphologically) *open* (*closed*) with respect to the structuring element W iff $Y \circ W = Y$ $(Y \bullet W = Y)$. We shall say that a set Y is *smooth with respect to* W iff Y can be expressed as a union of shifted replicas of W. Y is open with respect to W, iff Y is smooth with respect to W. Y is closed with respect to W iff Y^c is smooth with respect to W.

4.4.1 Some results on constrained optimality, or, why Morphology is popular.

Complex Morphological filters can be constructed by composing more elementary operators. For example, the family of Alternating Sequential Filters (ASF's) is constructed by alternating openings and closings with structuring elements of increasing size[6]. One good reason for the widespread use of Morphological filters is their excellent shape-preservation (syntactic) properties. Important characterizations (e.g., root signal structure) are well developed and understood [64], and this has helped build valuable intuition in the image processing community. Consequently, the empirical design of these filters has been greatly facilitated, and the resulting filters perform surprisingly well in a variety of noisy environments. However, with few exceptions [55], very little has been done in terms of "generic" DRS-theoretic optimization of Morphological filters.

Morphological filters are very flexible, mainly because of the freedom to

[6]See [58] for a recent survey of Morphological filtering.

choose the structuring element(s), to meet specified criteria. Among other things, Morphological filters have been widely used to filter out certain kinds of impulsive noise, such as the so-called salt-and-pepper noise, in both binary and gray scale images [55, 16, 19, 18, 14, 15, 64]. For example, it is widely believed that opening is suitable under a union noise model, while closing is suitable under an intersection noise model. ASF's are deemed appropriate under a combined union/intersection noise model. Indeed, these filters are used extensively, and they deliver adequate filtering in a variety of noisy environments. The natural question, then, is whether we can provide some sort of theoretical justification for their use. As it turns out, these filters are indeed optimal under a reasonable worst-case scenario. In particular, if we assume that the signal, X, is sufficiently smooth, and the noise is iid, then these operators provide the Maximum A Posteriori (MAP) estimate of X, on the basis of the observation Y. For the rest of this subsection, we assume that structuring elements contain the origin. We have the following results.

Theorem 4.1 *Let $O_W(B)$ denote the collection of all W-open subsets of B. Assume that the signal DRS, X, on B, induces the following probability mass function on $\Sigma(B)$:*

$$P_X(X = K) = \begin{cases} \frac{1}{|O_W(B)|} & , \; if \; K \in O_W(B) \\ \\ 0 & , \; otherwise \end{cases}$$

where $|\ |$ stands for set cardinality. Furthermore, assume that the observable

DRS is $Y = X \cup N$, *where* N *is a homogeneous Bernoulli lattice process of intensity* $r \in [0,1)$ *(i.e. each point* $z \in B$ *is included in* N *with probability* r, *independently of all other points), which is independent of* X. *Then* $Y \circ W$ *is the unique MAP estimate of* X *on the basis of* Y, *regardless of the specific value of* r.

Proof:

Let $\widehat{X}_{MAP}(Y)$ denote the MAP estimate of X on the basis of Y. Then, by definition,

$$\widehat{X}_{MAP}(Y) = arg\ max_{K \in \Sigma(B)} \{Pr(X = K \mid Y)\}$$

Using Bayes' rule,

$$\widehat{X}_{MAP}(Y) = arg\ max_{K \in \Sigma(B)} \{Pr(Y \mid X = K) P_X(X = K)\}$$

$$= arg\ max_{K \in O_W(B)} \left\{ Pr(Y \mid X = K) \frac{1}{|O_W(B)|} \right\}$$

$$= arg\ max_{K \in O_W(B)} \{Pr(Y \mid X = K)\}$$

$$= arg\ max_{K \in O_W(B),\ K \subseteq Y} \{Pr(Y \mid X = K)\}$$

$$= arg\ max_{K \in O_W(B),\ K \subseteq Y} \left\{ r^{|Y|-|K|} (1-r)^{|B|-|Y|} \right\}$$

$$= arg\ max_{K \in O_W(B),\ K \subseteq Y} \left\{ r^{-|K|} \right\}$$

$$= arg\ max_{K \in O_W(B),\ K \subseteq Y} \{|K|\}$$

So $\widehat{X}_{MAP}(Y)$ is the largest W-open subset of Y, which is by definition the opening of Y by W, i.e.

$$\widehat{X}_{MAP}(Y) = Y \circ W$$

and the proof is complete. □

A little reflection on the above result is in order. First, observe that the proof *crucially* depends on $|B|$ being finite. Indeed, this theorem, as well as the three theorems that follow, do not make sense when the lattice extends to infinity. Thus, a uniformly bounded discrete random set approach offers a fresh statistical perspective of Morphological filtering, one which is not apparent within other formulations. The suppositions of the theorem indeed correspond to a worst-case statistical scenario: if all that is known about the signal is that it is almost surely (a.s.) smooth (open) with respect to W, then it is reasonable to model this knowledge using a uniform distribution over the set of all W-open subsets of B, to reflect the fact that the signal exhibits no other (known) probabilistic structure. Also, iid noise is the worst kind of noise, in the sense of maximizing the Shannon entropy of the noise DRS N. Both these suppositions are plausible in practice, and this explains why the opening filter is successful under a union noise model. It is worth noticing that the MAP estimate does not depend on the noise level, r. Similarly, we have the following theorem.

Theorem 4.2 *Let $C_W(B)$ denote the collection of all W-closed subsets of B. Assume that the signal DRS, X, on B, induces the following probability mass*

function on $\Sigma(B)$*:*

$$P_X(X = K) = \begin{cases} \frac{1}{|C_W(B)|} & , \; if \; K \in C_W(B) \\ \\ 0 & , \; otherwise \end{cases}$$

Furthermore, assume that the observable DRS is $Y = X \cap N$*, where* N *is a homogeneous Bernoulli lattice process of intensity* $r \in [0, 1)$*, which is independent of* X*. Then* $Y \bullet W$ *is the unique MAP estimate of* X *on the basis of* Y*, regardless of the specific value of* r*.*

Proof:

By definition,

$$\widehat{X}_{MAP}(Y) = arg \; max_{K \in \Sigma(B)} \{Pr(X = K \mid Y)\}$$

$$= arg \; max_{K \in \Sigma(B)} \{Pr(Y \mid X = K) P_X(X = K)\}$$

$$= arg \; max_{K \in C_W(B)} \left\{ Pr(Y \mid X = K) \frac{1}{|C_W(B)|} \right\}$$

$$= arg \; max_{K \in C_W(B)} \{Pr(Y \mid X = K)\}$$

$$= arg \; max_{K \in C_W(B), \; K \supseteq Y} \{Pr(Y \mid X = K)\}$$

$$= arg \; max_{K \in C_W(B), \; K \supseteq Y} \left\{ r^{|Y|} (1 - r)^{|K| - |Y|} \right\}$$

$$= arg \; max_{K \in C_W(B), \; K \supseteq Y} \left\{ (1 - r)^{|K|} \right\}$$

$$= arg \; min_{K \in C_W(B), \; K \supseteq Y} \{|K|\}$$

So $\widehat{X}_{MAP}(Y)$ is the smallest W-closed superset of Y, which is by definition the closing of Y by W, i.e.

$$\widehat{X}_{MAP}(Y) = Y \bullet W$$

and the proof is complete. □

The following two theorems are straightforward extensions of the above theorems. We state them here without proof.

Theorem 4.3 *Let $O_{W_1,\cdots W_M}(B)$ denote the collection of all subsets K of B which can be written as*

$$K = \bigcup_{i=1,\cdots,M} K_i, \quad K_i \in O_{W_i}(B),\ i = 1,\cdots,M$$

Assume that the signal DRS, X, on B, induces the following probability mass function on $\Sigma(B)$:

$$P_X(X = K) = \begin{cases} \frac{1}{|O_{W_1,\cdots,W_M}(B)|} & , \ if\ K \in O_{W_1,\cdots,W_M}(B) \\ \\ 0 & , \ otherwise \end{cases}$$

Furthermore, assume that the observable DRS is $Y = X \cup N$, where N is a homogeneous Bernoulli lattice process of intensity $r \in [0,1)$, which is independent of X. Then

$$\widehat{X}_{MAP}(Y) = \bigcup_{i=1,\cdots,M} Y \circ W_i$$

Theorem 4.4 *Let $C_{W_1,\cdots W_M}(B)$ denote the collection of all subsets K of B*

which can be written as

$$K = \bigcap_{i=1,\cdots,M} K_i, \quad K_i \in C_{W_i}(B),\ i = 1,\cdots,M$$

Assume that the signal DRS, X, on B, induces the following probability mass function on $\Sigma(B)$:

$$P_X(X = K) = \begin{cases} \frac{1}{|C_{W_1,\cdots,W_M}(B)|} & , \ if\ K \in C_{W_1,\cdots,W_M}(B) \\ \\ 0 & , \ otherwise \end{cases}$$

Furthermore, assume that the observable DRS is $Y = X \cap N$, where N is a homogeneous Bernoulli lattice process of intensity $r \in [0,1)$, which is independent of X. Then

$$\widehat{X}_{MAP}(Y) = \bigcap_{i=1,\cdots,M} Y \bullet W_i$$

A natural question which arises here is what happens if we loose the uniform probability structure over the collection of smooth realizations. The answer is that the MAP estimate will typically be intractable. However, if every smooth realization has a positive probability, then we can still claim that the proposed estimate in any of the above theorems is the Maximum Likelihood (ML) estimate of X on the basis of Y. For example, a (λ_s, H, f_R)-DRBRS X, with $f_R(R = 0) = 0$, and primary grains which are properly contained in B, induces a pmf which satisfies $P_X(X = K) = 0$, $iff\ K \in \Sigma(B) \backslash O_H(B)$, but $P_X(X = K)$ is not uniform over $O_H(B)$. Nevertheless, we can still claim that, under the remaining assumptions of theorem 4.1, $Y \circ H$ is the ML estimate of X on the basis of Y.

In general, if we assume that X satisfies some arbitrary (not necessarily Morphological) smoothness conditions, i.e. $X \in \mathcal{S}$, a class of smooth subsets of B, and that X is uniformly distributed over $\mathcal{S}$, then under an iid symmetric (Binary Symmetric Channel, BSC) noise model of pixel inversion probability $r < 0.5$, it is easy to see that

$$\widehat{X}_{MAP}(Y) = arg\ min_{K \in \mathcal{S}} d(Y, K)$$

where $d(Y, K)$ is the area of the symmetric set difference distance between Y and K. In other words, $\widehat{X}_{MAP}(Y)$ is the "projection" of the data Y onto $\mathcal{S}$. However, it is not clear how to compute this projection under general smoothness conditions. Furthermore, quite often the noise is not iid, and the signal is nonsmooth, or only approximately smooth. The lack of a rigorous DRS-theoretic optimization approach for this general case has been evident in the literature. Our programme is to develop such an approach. Specifically, for each degradation model, we will construct a suitable class of Morphological operators, argue about its merits, and derive results which explicitly characterize the optimal choice of structuring element(s) in terms of the fundamental functionals of random set theory, namely the generating functional of the signal, and the generating functional of the noise.

4.4.2 Optimal increasing, shift-invariant filters with a basis constraint.

A surprising result, originally due to Matheron [45], and subsequently improved upon, and used by Maragos [42], and Dougherty et al., and Giardina [16, 17, 19, 18], is that a very large class of (linear and non-linear) shift-invariant operations can be decomposed into a union of erosions by suitable structuring elements.

Let $E = \mathbf{Z}^2$, and let $\Sigma(E)$ denote the power set of E. Let $\Psi : \Sigma(E) \mapsto \Sigma(E)$. Recall that Ψ is *increasing* **iff** $X_1 \subseteq X_2 \Rightarrow \Psi(X_1) \subseteq \Psi(X_2)$, $\forall X_1 \in \Sigma(E)$, $X_2 \in \Sigma(E)$. We now reproduce some key theorems, taken from [45, 42].

Theorem 4.5 *[45] For any* shift-invariant and increasing *mapping* $\Psi : \Sigma(E) \mapsto \Sigma(E)$, *and for all* $X \in \Sigma(E)$,

$$\Psi(X) = \bigcup_{W \in Ker(\Psi)} X \ominus W^s$$

where the kernel of Ψ, $Ker(\Psi)$, *is defined as*

$$Ker(\Psi) \triangleq \{W \in \Sigma(E) \mid \bar{0} \in \Psi(W)\}$$

Theorem 4.6 *[42] For any* shift-invariant and increasing *mapping* $\Psi : \Sigma(E) \mapsto \Sigma(E)$, *and for all* $X \in \Sigma(E)$,

$$\Psi(X) = \bigcup_{W \in Bas(\Psi)} X \ominus W^s$$

where the erosion basis *of* Ψ, $Bas(\Psi)$, *is defined as*

$$Bas(\Psi) \triangleq \left\{W \in Ker(\Psi) \mid W' \in Ker(\Psi) \text{ and } W' \subseteq W \Rightarrow W' = W\right\}$$

As a result of the latter theorem, the number of structuring elements that are needed for the decomposition is greatly reduced. Dougherty et al. [17, 19, 18], have made extensive use of this result to reduce the complexity associated with the design and implementation of optimal mean-square Morphological filters. By duality, there exists an equivalent decomposition of any shift-invariant and increasing mapping as an intersection of dilations [14] over a *dilation basis.*

Strictly speaking, these theorems cannot be used with bounded domains. However, modulo some modifications which account for edge effects, they can be utilized. Then, the question of finding the optimal shift-invariant and increasing filter reduces to the problem of optimal basis design. This reduction is a significant one; the former problem is highly unstructured, whereas the latter admits a natural hierarchical decomposition, in terms of a basis size constraint. In other words, we can consider a sequence of problems characterized by an increasing basis size. The performance of the optimal constrained filter is necessarily a non-decreasing function of the size of the basis. The upper bound on the size of the basis is usually determined by design and implementation considerations. However, under a basis size constraint, we are faced with an additional problem: should we choose the expansion in terms of an erosion basis, or in terms of a dilation basis? We will argue for the following point: *under an intersection noise model, contrary to our intuition, we should think of the optimal filter as a union of erosions, whereas under a union noise model we should think of the optimal filter as an intersection of dilations.* In both cases, we can work

out theoretical formulas for the cost function.

4.4.3 Optimizing a single structuring element

In the case of union noise, the simplest non-trivial expansion in terms of a dilation basis involves two structuring elements, one of which is constrained to be the origin. This is because we want the overall operation to be anti-extensive, i.e. the output must be contained in the input. Dilation by the origin simply yields the input itself. Therefore, the simplest non-trivial class of constrained dilation basis filters for union noise can be written as follows:

$$f(Y) = f_W(Y) = (Y \oplus W^s) \cap Y$$

$$= [(X \cup N) \oplus W^s] \cap (X \cup N), \ \textit{for some structuring element}, \ W$$

Similarly, in the case of intersection noise, the simplest non-trivial expansion in terms of an erosion basis involves two structuring elements, one of which is constrained to be the origin (because we want the overall operation to be extensive, i.e. the output must contain the input). Again, since erosion by the origin yields the input itself, the simplest non-trivial class of constrained erosion basis filters for intersection noise can be written as follows:

$$f(Y) = f^W(Y) = (Y \ominus W^s) \cup Y$$

$$= [(X \cap N) \ominus W^s] \cup (X \cap N), \ \textit{for some structuring element}, \ W$$

Some motivation is necessary at this point. Let us first consider the case of intersection noise. Intuitively, since the noise removes points from the signal,

we should use some sort of "fill-in" operation to cancel the effect of noise. By definition,

$$Y \ominus W^s = \{z \mid W_z \subseteq Y\}$$

If the structuring element, W, is appropriately chosen (in particular it must not contain the origin), then the erosion operation is a fill-in operation, i.e. it fills gaps in the "body" of the observation. However, it also introduces new gaps, which is an undesired side effect. Nevertheless, we can easily get rid of these "spurious" new gaps, by simply taking the union of the resulting eroded set with the input set (i.e. the observation, Y) itself. Some structuring elements that can be used in this mode are depicted in figure 4.5 (a cross indicates the location of the origin). An example is given in figure 4.6. Figure 4.6a depicts a test image, while figure 4.6b depicts a degraded version of the test image, obtained by intersecting it with the set of points which make up a realization of a Bernoulli random field, of intensity 0.9. Figure 4.6c depicts the estimate, $\widehat{X} = [(X \cap N) \ominus W^s] \cup (X \cap N)$ where X is the original test image depicted in figure 4.6a, N is the set of points of the Bernoulli field, $X \cap N$ is the observation depicted in figure 4.6b, and W is the leftmost of the structuring elements which appear in figure 4.5. For this example, the structuring element was not optimized.

In loose terms, if a structuring element does not contain *a neighborhood* of the origin, then it can be used in a gap-filling mode. The larger this neighborhood, the wider the gaps that can be (partially) filled by an erosion with the given

structuring element.

Similarly, by duality, if the structuring element is appropriately chosen (in particular, it must not contain the origin), dilation can remove points from the observation, and, therefore, it can be appropriate under a union noise model. After performing a dilation with a suitably chosen structuring element, we take the intersection of the resulting set with the input (observation) set, to eliminate points that have been introduced by the dilation operation.

This mode of use of the two basic Morphological operations may seem strange at first, since, for example, and partially because of its name, most people think of erosion as a shrink-type operation. However, one should keep in mind that this is only true if the erosion structuring element contains the origin. In fact, most people would consider using the operations in a reverse fashion: dilation for the case of intersection noise[7], and erosion for the case of union noise. The reason for our "unconventional" approach is that this way we can take advantage of certain distributivity properties, and obtain closed-form characterizations of the optimal filters.

Let us first consider intersection noise. Here[8],

$$g(X, N) = X \cap N$$

[7]See [28] for an account of such an approach, when the intersection mask, N, is a deterministic, regularly spaced grid, which undersamples the observation.

[8]This operation can be viewed as *random sampling* the DRS X. In this context, our results characterize the optimal (within a class) Morphological reconstruction filter, for DRS's that have undergone random sampling.

and

$$\widehat{X} = f(Y) = f^W(Y) = (Y \ominus W^s) \cup Y$$

$$= [(X \cap N) \ominus W^s] \cup (X \cap N), \textit{ for some structuring element, } W \in \mathcal{W}$$

where $\mathcal{W}$ is the collection of structuring elements over which we intend to optimize. We need to make a small modification to our fidelity criterion, in order to account for incomplete data close to the border of B. Towards this end, define

$$B \backslash \partial B = B \cap \left(\bigcap_{W \in \mathcal{W}} B \ominus W^s \right)$$

$B \backslash \partial B$ is exactly the set of points $z \in B$ with the property that $W_z \subseteq B$, $\forall W \in \mathcal{W}$. Then we only consider the total expected error restricted to $B \backslash \partial B$. We also assume that estimates of X are only valid within $B \backslash \partial B$. For brevity, we use the same symbol to denote a DRS and its restriction to $B \backslash \partial B$. The meaning is clear from context. We have the following proposition.

Proposition 4.6 *Under the assumption of mutual independence of the signal and noise DRS's, X, N, the value of the expected error, $E(e) = Ed(X, \widehat{X})$, incurred when X is estimated by $\widehat{X} = [(X \cap N) \ominus W^s] \cup (X \cap N)$, is given by*

$$E(e) = \sum_{z \in B \backslash \partial B} \{ Q_{X^c}(\{z\}) (1 - Q_{N^c}(\{z\}))$$

$$+ Q_{N^c}(W_z) (Q_{X^c}(W_z) - Q_{X^c}(\{z\} \cup W_z))$$

$$+ Q_{X^c}(\{z\} \cup W_z) (Q_{N^c}(\{z\} \cup W_z) - Q_{N^c}(W_z)) \}$$

Proof:

Observe that, by distributivity of erosion over intersection

$$(X \cap N) \ominus W^s = (X \ominus W^s) \cap (N \ominus W^s)$$

This property is crucial for the proof. During the course of the proof, we will need the following elementary result. Define the functional

$$\Gamma_{X,n}(K_0; K_1, \cdots, K_n) \triangleq Pr(X \cap K_0 = \emptyset \ , \ X \cap K_1 \neq \emptyset, \cdots, X \cap K_n \neq \emptyset)$$

By definition, $\Gamma_{X,0}(K) = Q_X(K)$. Using Bayes' rule, one can easily show that this functional satisfies the following recursion, known as the *inclusion - exclusion principle.*

$$\Gamma_{X,n}(K_0; K_1, \cdots, K_n) = \Gamma_{X,n-1}(K_0; K_1, \cdots, K_{n-1})$$

$$-\Gamma_{X,n-1}(K_0 \cup K_n; K_1, \cdots, K_{n-1})$$

We are now ready to proceed with the proof of the proposition. The total expected error is

$$E(e) = E|X \backslash \widehat{X}| + E|\widehat{X} \backslash X|$$

Let us first consider the second term.

$$E|\widehat{X} \backslash X| = E|\widehat{X} \cap X^c|$$

$$= E\,|([(X \cap N) \ominus W^s] \cup (X \cap N)) \cap X^c|$$

$$= E\,|([(X \cap N) \ominus W^s] \cap X^c) \cup (X \cap N \cap X^c)|$$

$$= E\,|[(X \cap N) \ominus W^s] \cap X^c|$$

Now, since

$$(X \cap N) \ominus W^s = (X \ominus W^s) \cap (N \ominus W^s)$$

The last expression is equal to

$$E|(X \ominus W^s) \cap (N \ominus W^s) \cap X^c|$$

$$= E \sum_{z \in B \setminus \partial B} 1(z \in (X \ominus W^s) \cap (N \ominus W^s) \cap X^c)$$

$$= \sum_{z \in B \setminus \partial B} E\ 1(z \in (X \ominus W^s) \cap (N \ominus W^s) \cap X^c)$$

$$= \sum_{z \in B \setminus \partial B} Pr(z \in (X \ominus W^s) \cap (N \ominus W^s) \cap X^c)$$

$$= \sum_{z \in B \setminus \partial B} Pr(z \in (X \ominus W^s) \cap X^c)\ ,\ z \in N \ominus W^s)$$

$$= \sum_{z \in B \setminus \partial B} Pr(z \in (X \ominus W^s) \cap X^c) Pr(z \in N \ominus W^s)$$

$$= \sum_{z \in B \setminus \partial B} Pr(W_z \subseteq X\ ,\ z \in X^c) Pr(W_z \subseteq N)$$

$$= \sum_{z \in B \setminus \partial B} Pr(X^c \cap W_z = \emptyset\ ,\ X^c \cap \{z\} \neq \emptyset) Pr(N^c \cap W_z = \emptyset)$$

The first term of the total expected error

$$E|X \setminus \widehat{X}| = E|X \cap (\widehat{X})^c|$$

$$= E\,|X \cap ([(X \cap N) \ominus W^s] \cup (X \cap N))^c|$$

$$= E\,|X \cap [(X \cap N) \ominus W^s]^c \cap (X \cap N)^c|$$

$$= E\,|[X \cap (X \cap N)^c] \cap [(X \cap N) \ominus W^s]^c|$$

$$= E\,|[X \cap (X^c \cup N^c)] \cap [(X \cap N) \ominus W^s]^c|$$

$$= E\,|[(X \cap X^c) \cup (X \cap N^c)] \cap [(X \cap N) \ominus W^s]^c|$$

$$= E\,|X \cap N^c \cap [(X \cap N) \ominus W^s]^c|$$

$$= E\,|X \cap N^c \cap [(X \ominus W^s) \cap (N \ominus W^s)]^c|$$

$$= E\,|X \cap N^c \cap [(X \ominus W^s)^c \cup (N \ominus W^s)^c]|$$

$$= E\,|(X \cap N^c \cap (X \ominus W^s)^c) \cup (X \cap N^c \cap (N \ominus W^s)^c)|$$

$$= E\,|X \cap N^c \cap (X \ominus W^s)^c| + E\,|X \cap N^c \cap (N \ominus W^s)^c|$$

$$-E\,|X \cap N^c \cap (X \ominus W^s)^c \cap (N \ominus W^s)^c|$$

Now,

$$E\,|X \cap N^c \cap (X \ominus W^s)^c| = E\,|(X \cap (X \ominus W^s)^c) \cap N^c|$$

$$= \sum_{z \in B \setminus \partial B} Pr(z \in X\ ,\ \neg(W_z \subseteq X))Pr(z \in N^c)$$

$$= \sum_{z \in B \setminus \partial B} Pr(X^c \cap \{z\} = \emptyset\ ,\ X^c \cap W_z \neq \emptyset)Pr(N^c \cap \{z\} \neq \emptyset)$$

where $\neg$ denotes logical negation. Also,

$$E\,|X \cap N^c \cap (N \ominus W^s)^c|$$

$$= \sum_{z \in B \setminus \partial B} Pr(z \in X)Pr(z \in N^c\ ,\ \neg(W_z \subseteq N))$$

$$= \sum_{z \in B \setminus \partial B} Pr(X^c \cap \{z\} = \emptyset)Pr(N^c \cap \{z\} \neq \emptyset\ ,\ N^c \cap W_z \neq \emptyset)$$

And

$$E\,|X \cap N^c \cap (X \ominus W^s)^c \cap (N \ominus W^s)^c|$$

$$= E\,|(X \cap (X \ominus W^s)^c) \cap (N^c \cap (N \ominus W^s)^c)|$$

$$= \sum_{z \in B \setminus \partial B} Pr(z \in X \ , \ \neg(W_z \subseteq X)) Pr(z \in N^c \ , \ \neg(W_z \subseteq N))$$

$$= \sum_{z \in B \setminus \partial B} Pr(X^c \cap \{z\} = \emptyset \ , \ X^c \cap W_z \neq \emptyset) Pr(N^c \cap \{z\} \neq \emptyset \ , \ N^c \cap W_z \neq \emptyset)$$

Therefore, putting everything together,

$$E(e) = E|X \setminus \widehat{X}| + E|\widehat{X} \setminus X|$$

$$= \sum_{z \in B \setminus \partial B} \{ Pr(X^c \cap \{z\} = \emptyset \ , \ X^c \cap W_z \neq \emptyset) Pr(N^c \cap \{z\} \neq \emptyset)$$

$$+ Pr(X^c \cap \{z\} = \emptyset) Pr(N^c \cap \{z\} \neq \emptyset \ , \ N^c \cap W_z \neq \emptyset)$$

$$- Pr(X^c \cap \{z\} = \emptyset \ , \ X^c \cap W_z \neq \emptyset) Pr(N^c \cap \{z\} \neq \emptyset \ , \ N^c \cap W_z \neq \emptyset)$$

$$+ Pr(X^c \cap W_z = \emptyset \ , \ X^c \cap \{z\} \neq \emptyset) Pr(N^c \cap W_z = \emptyset) \}$$

$$= \sum_{z \in B \setminus \partial B} \{ \Gamma_{X^c,1}(\{z\}; W_z) (1 - Q_{N^c}(\{z\})) + Q_{X^c}(\{z\}) \Gamma_{N^c,2}(\emptyset; \{z\}, W_z)$$

$$- \Gamma_{X^c,1}(\{z\}; W_z) \Gamma_{N^c,2}(\emptyset; \{z\}, W_z) + \Gamma_{X^c,1}(W_z; \{z\}) Q_{N^c}(W_z) \}$$

$$= \sum_{z \in B \setminus \partial B} \{ (Q_{X^c}(\{z\}) - Q_{X^c}(\{z\} \cup W_z)) (1 - Q_{N^c}(\{z\}))$$

$$+ Q_{X^c}(\{z\}) (1 - Q_{N^c}(\{z\}) - Q_{N^c}(W_z) + Q_{N^c}(\{z\} \cup W_z))$$

$$- (Q_{X^c}(\{z\}) - Q_{X^c}(\{z\} \cup W_z)) (1 - Q_{N^c}(\{z\}) - Q_{N^c}(W_z) + Q_{N^c}(\{z\} \cup W_z))$$

$$+ (Q_{X^c}(W_z) - Q_{X^c}(\{z\} \cup W_z)) Q_{N^c}(W_z) \}$$

From which, after some manipulations, we obtain

$$E(e) = \sum_{z \in B \setminus \partial B} \{ Q_{X^c}(\{z\}) (1 - Q_{N^c}(\{z\})) \}$$

$$+Q_{N^c}(W_z)\,(Q_{X^c}(W_z) - Q_{X^c}(\{z\} \cup W_z))$$

$$+Q_{X^c}(\{z\} \cup W_z)\,(Q_{N^c}(\{z\} \cup W_z) - Q_{N^c}(W_z))\}$$

and the result is established. □

Observe that the total expected error is equal to the sum of the probabilities of individual pixel errors. If we make the natural[9] assumption that both X, and N, are obtained by sampling *stationary* random sets [45], then all the functionals in the above sum are independent of the location, $\{z\}$, and we obtain the following result.

Corollary 4.1 *Under the condition of mutual independence of the signal and noise DRS's, X, N, assuming that X, N, are obtained by sampling stationary RS's, and that X is estimated by $\widehat{X} = [(X \cap N) \ominus W^s] \cup (X \cap N)$, the optimal choice of the structuring element W is the one which minimizes the* probability of pixel error

$$P_{pixel\ error}(W) = Q_{X^c}(\{\bar{0}\})\,(1 - Q_{N^c}(\{\bar{0}\}))$$

$$+Q_{N^c}(W)\,(Q_{X^c}(W) - Q_{X^c}(\{\bar{0}\} \cup W))$$

$$-Q_{X^c}(\{\bar{0}\} \cup W)\,(Q_{N^c}(W) - Q_{N^c}(\{\bar{0}\} \cup W))$$

[9]Since we are using a *shift-invariant* filtering operation. If the functionals in the above sum depend on the location, then, clearly, the optimal filter will not (in general) be shift-invariant; i.e., the optimal structuring element will be different for different locations within the image.

Let us examine the individual terms of this sum. The first term,

$$Q_{X^c}(\{\bar{0}\})\,(1 - Q_{N^c}(\{\bar{0}\}))$$

of the probability of pixel error, $P_{pixel\ error}(W)$, is exactly the probability of pixel error between the signal X, and the observation $X \cap N$ (this can be seen by setting $W = \{\bar{0}\}$, which corresponds to no filtering of the observation). This first term is independent of W, and, therefore, it is not under our control. The remaining two terms of the sum are both non-negative functions of W (it can be easily shown that the generating functional of an arbitrary DRS is constrained to be decreasing). When considered as a function of W, this sum clearly brings out the interplay between "signal power" and "noise power", and how it determines the structuring element that achieves the optimal trade-off between eliminating gaps introduced by noise, and retaining gaps that are present in the signal itself.

Some notes on the applicability of this result are in order. If the generating functionals $Q_{X^c}(\cdot)$, $Q_{N^c}(\cdot)$ (or, equivalently, the capacity functionals $T_{X^c}(\cdot)$, $T_{N^c}(\cdot)$), are given, then optimization of W over a relatively small collection of allowable $W's$ is straightforward. In general, for large collections of candidate structuring elements, some sort of suboptimal search must be pursued, to avoid a potentially difficult exhaustive search. See [41] for an "expert" structuring element library design approach. We shall return to this point later on. At any rate, even if the generating functionals are not available (which is the case in most applications), all the quantities which are relevant to our optimization

problem can be efficiently and accurately estimated from running (sample) averages, by virtue of stationarity and the law of large numbers. For example, $Q_{X^c}(W)$ can be estimated by "sliding"the structuring element W across a realization of X^c and counting the number of times that the two have an empty intersection, and similarly for the others.

Let us now turn to union noise. Here,

$$g(X, N) = X \cup N$$

and

$$\widehat{X} = f(Y) = f_W(Y) = (Y \oplus W^s) \cap Y$$

$$= [(X \cup N) \oplus W^s] \cap (X \cup N), \ \textit{for some structuring element}, \ W \in \mathcal{W}$$

As expected, we can once more resort to duality. In particular, since

$$(\widehat{X})^c = ([(X \cup N) \oplus W^s] \cap (X \cup N))^c$$

$$= [(X \cup N) \oplus W^s]^c \cup (X \cup N)^c$$

$$= [(X \cup N)^c \ominus W^s] \cup (X^c \cap N^c)$$

$$= [(X^c \cap N^c) \ominus W^s] \cup (X^c \cap N^c)$$

and

$$d(X_1, X_2) = d(X_1^c, X_2^c)$$

we can easily reduce this case to the case of the previous subsection, by replacing X, N by their complements, X^c, N^c. Thus we have the following result.

Proposition 4.7 *Under the assumption of mutual independence of the signal and noise DRS's, X, N, the value of the expected error, $E(e) = Ed(X, \widehat{X})$, incurred when X is estimated by $\widehat{X} = [(X \cup N) \oplus W^s] \cap (X \cup N)$, is given by*

$$E(e) = \sum_{z \in B \setminus \partial B} \{Q_X(\{z\})(1 - Q_N(\{z\}))$$

$$+Q_N(W_z)(Q_X(W_z) - Q_X(\{z\} \cup W_z))$$

$$+Q_X(\{z\} \cup W_z)(Q_N(\{z\} \cup W_z) - Q_N(W_z))\}$$

Again, if we make the assumption that both X, and N, are obtained by sampling stationary RS's, then we obtain the following result.

Corollary 4.2 *Under the condition of mutual independence of the signal and noise DRS's, X, N, assuming that X, N, are obtained by sampling stationary RS's, and that X is estimated by $\widehat{X} = [(X \cup N) \oplus W^s] \cap (X \cup N)$, the optimal choice of the structuring element W is the one which minimizes the* probability of pixel error

$$P_{pixel\ error}(W) = Q_X(\{\bar{0}\})(1 - Q_N(\{\bar{0}\}))$$

$$+Q_N(W)(Q_X(W) - Q_X(\{\bar{0}\} \cup W))$$

$$-Q_X(\{\bar{0}\} \cup W)(Q_N(W) - Q_N(\{\bar{0}\} \cup W))$$

As in the case of intersection noise, similar remarks hold here regarding the interpretation of the individual terms of the sum. Again, if the generating functionals $Q_X(\cdot)$, $Q_N(\cdot)$, are given, then optimization over a small collection of

candidate structuring elements is straightforward. If these functionals are not available, their values can be estimated from running averages, as before.

Let us now show how one can reduce the complexity of the search for the optimal structuring element, by assuming that the signal DRS, X, is smooth (i.e. Morphologically open) with respect to some structuring element. We will need the following.

Definition 4.4.19 *A DRS X is H-open if*[10]

$$P_X(X = K) = P_{X \circ H}(X \circ H = K), \ \forall K \in \Sigma(B)$$

Lemma 4.1 *X is H-open iff $Q_X(K) = Q_{X \ominus H^s}(K \oplus H^s)$, $\forall K \in \Sigma(B)$.*

Proof:

First observe that, for any DRS X, and $\forall K \in \Sigma(B)$,

$$Q_{X \oplus H}(K) = P_X((X \oplus H) \cap K = \emptyset)$$

$$= P_X(X \cap (K \oplus H^s) = \emptyset) = Q_X(K \oplus H^s)$$

Thus, assuming X is H-open,

$$Q_{X \ominus H^s}(K \oplus H^s) = Q_{(X \ominus H^s) \oplus H}(K) = Q_{X \circ H}(K)$$

[10]Observe that this definition asserts that X is H-open iff $X \circ H = X$ in the sense of distributions. However, this implies that

$$P_X(X \circ H \neq X) = \sum_{K \circ H \neq K} P_X(X = K) = \sum_{K \circ H \neq K} P_{X \circ H}(X \circ H = K) = 0$$

i.e. $P_X(X \circ H = X) = 1$, which implies $X \circ H = X$, $P_X - a.s.$ Thus we do not make any distinction.

$$= \sum_{K' \subseteq K^c} P_{X \circ H}(X \circ H = K') = \sum_{K' \subseteq K^c} P_X(X = K') = Q_X(K)$$

Conversely, assume that $Q_X(K) = Q_{X \ominus H^s}(K \oplus H^s)$, $\forall K \in \Sigma(B)$. Then, since $Q_{X \ominus H^s}(K \oplus H^s) = Q_{(X \ominus H^s) \oplus H}(K) = Q_{X \circ H}(K)$, it follows that $Q_X(K) = Q_{X \circ H}(K)$, $\forall K \in \Sigma(B)$. By theorem 2.1,

$$P_{X \circ H}(X \circ H = K) = \sum_{K' \subseteq K} (-1)^{|K'|} Q_{X \circ H}(K^c \cup K')$$

$$= \sum_{K' \subseteq K} (-1)^{|K'|} Q_X(K^c \cup K') = P_X(X = K), \ \forall K \in \Sigma(B)$$

and the proof is complete. □

So now let us assume that the signal DRS, X, is H-open, where H is convex and contains the origin. Let $\mathcal{W}$ denote the collection of candidate structuring elements over which we intend to optimize. Consider the second term of the sum for the probability of pixel error. Using the above lemma,

$$Q_X(W) - Q_X(\{\bar{0}\} \cup W)$$

$$= Q_{X \ominus H^s}(W \oplus H^s) - Q_{X \ominus H^s}((\{\bar{0}\} \cup W) \oplus H^s)$$

By distributivity of dilation over union,

$$Q_{X \ominus H^s}((\{\bar{0}\} \cup W) \oplus H^s)$$

$$= Q_{X \ominus H^s}((\{\bar{0}\} \oplus H^s) \cup (W \oplus H^s))$$

$$= Q_{X \ominus H^s}(H^s \cup (W \oplus H^s))$$

Thus, under the condition

$$H^s \subseteq W \oplus H^s, \quad \forall\, W \in \mathcal{W}$$

the second term of the sum for the probability of pixel error is zero. In loose terms, this condition amounts to requiring H to be "large enough" relative to the structuring elements in $\mathcal{W}$. Since the signal is usually associated with the more prominent patterns in the image, this requirement is not very restrictive. For example, if $\mathcal{W}$ is the collection of structuring elements depicted in figure 4.5, and H is a 3×3 square of pixels, which is centered at the origin, then the above condition is satisfied. Therefore, the optimal $W \in \mathcal{W}$ should maximize the third term of the sum, namely

$$G(W) \triangleq Q_X(\{\bar{0}\} \cup W)\,(Q_N(W) - Q_N(\{\bar{0}\} \cup W))$$

Now,

$$(Q_N(W) - Q_N(\{\bar{0}\} \cup W)) = \Gamma_{N,1}(W; \{\bar{0}\})$$

$$= P_N(N \cap W = \emptyset,\ N \cap \{\bar{0}\} \neq \emptyset)$$

which is clearly decreasing in W. Furthermore, $Q_X(\{\bar{0}\} \cup W)$ is decreasing in W. Thus, $G(W)$ is decreasing in W, i.e.

$$W_1 \subseteq W_2 \Rightarrow G(W_2) \leq G(W_1)$$

Hence, since we seek to maximize $G(W)$, we can eliminate from consideration all structuring elements in $\mathcal{W}$ which properly contain other structuring elements

in $\mathcal{W}$, i.e., it suffices to optimize over the subcollection

$$\widetilde{\mathcal{W}} = \left\{W \in \mathcal{W} \mid W' \in \mathcal{W} \text{ and } W' \subseteq W \Rightarrow W' = W\right\}$$

Thus, we have proven the following.

Corollary 4.3 *Under the condition of mutual independence of the signal and noise DRS's, X, N, assuming that X, N, are obtained by sampling stationary RS's, X is H-open, where H is convex, containing the origin, and such that $H^s \subseteq W \oplus H^s$, $\forall W \in \mathcal{W}$, and that X is estimated by $\widehat{X} = [(X \cup N) \oplus W^s] \cap (X \cup N)$, the optimal structuring element is*

$$W^* = arg\ min_{W \in \mathcal{W}} P_{pixel\ error}(W) = arg\ max_{W \in \widetilde{\mathcal{W}}} G(W)$$

This elimination can translate to a significant reduction in search complexity. For example, if $\mathcal{W}$ is the collection of structuring elements depicted in figure 4.5, and H is a 3×3 square of pixels, which is centered at the origin, then it suffices to optimize over the two leftmost structuring elements. By duality, a similar reduction can be achieved under an intersection noise model, if we assume that X^c is H-open, i.e. that X is H-closed.

4.4.4 Multiple structuring elements.

In certain situations, particularly when the noise level is high, a single erosion followed by a union, even if optimal, may not suffice to properly reconstruct the signal. In this case, it is beneficial to consider larger bases, i.e., filters

with multiple structuring elements. The structuring elements must be *jointly* optimized, to eliminate a wider class of error patterns. Using exactly the same algebraic methods, as in the proof of proposition 4.6, and with some patience, we can obtain similar optimality results for the case of multiple structuring elements. For example, we have the following result.

Proposition 4.8 *Under the assumption of mutual independence of the signal and noise DRS's, X, N, the value of the expected error, $E(e) = Ed(X, \widehat{X})$, incurred when X is estimated by*

$$\widehat{X} = [(X \cap N) \ominus W_1^s] \cup [(X \cap N) \ominus W_2^s] \cup (X \cap N)$$

is given by

$$E(e) = \sum_{z \in B \setminus \partial B} \{Q_{X^c}(\{z\})\,(1 - Q_{N^c}(\{z\}))$$

$$+Q_{X^c}(W_{1_z})Q_{N^c}(W_{1_z}) + Q_{X^c}(W_{2_z})Q_{N^c}(W_{2_z})$$

$$+Q_{X^c}(\{z\} \cup W_{1_z})Q_{N^c}(\{z\} \cup W_{1_z}) + Q_{X^c}(\{z\} \cup W_{2_z})Q_{N^c}(\{z\} \cup W_{2_z})$$

$$-2Q_{X^c}(\{z\} \cup W_{1_z})Q_{N^c}(W_{1_z}) - 2Q_{X^c}(\{z\} \cup W_{2_z})Q_{N^c}(W_{2_z})$$

$$-Q_{X^c}(W_{1_z} \cup W_{2_z})Q_{N^c}(W_{1_z} \cup W_{2_z}) - Q_{X^c}(\{z\} \cup W_{1_z} \cup W_{2_z})Q_{N^c}(\{z\} \cup W_{1_z} \cup W_{2_z})$$

$$+2Q_{X^c}(\{z\} \cup W_{1_z} \cup W_{2_z})Q_{N^c}(W_{1_z} \cup W_{2_z})\}$$

Proof:

$$Ed(X, \widehat{X}) = E|\widehat{X} \setminus X| + E|X \setminus \widehat{X}|$$

$$E|\widehat{X}\backslash X| = E|\widehat{X} \cap X^c|$$

$$= E\,|([(X \cap N) \ominus W_1^s] \cap X^c) \cup ([(X \cap N) \ominus W_2^s] \cap X^c) \cup (X \cap N \cap X^c)|$$

$$= E\,|([(X \cap N) \ominus W_1^s] \cap X^c) \cup ([(X \cap N) \ominus W_2^s] \cap X^c)|$$

$$= E\,|[(X \ominus W_1^s) \cap (N \ominus W_1^s) \cap X^c] \cup [(X \ominus W_2^s) \cap (N \ominus W_2^s) \cap X^c]|$$

$$= E\,|[(X \ominus W_1^s) \cap (N \ominus W_1^s) \cap X^c]| + E\,|[(X \ominus W_2^s) \cap (N \ominus W_2^s) \cap X^c]|$$

$$-E\,|[(X \ominus W_1^s) \cap (X \ominus W_2^s) \cap X^c] \cap [(N \ominus W_1^s) \cap (N \ominus W_2^s)]|$$

Consider

$$E\,|[(X \ominus W_1^s) \cap (N \ominus W_1^s) \cap X^c]|$$

$$= \sum_{z \in B \backslash \partial B} Pr(W_{1_z} \subseteq X,\ z \in X^c) Pr(W_{1_z} \subseteq N)$$

$$= \sum_{z \in B \backslash \partial B} Pr(X^c \cap W_{1_z} = \emptyset,\ X^c \cap \{z\} \neq \emptyset) Pr(N^c \cap W_{1_z} = \emptyset)$$

$$= \sum_{z \in B \backslash \partial B} \Gamma_{X^c,1}(W_{1_z}; \{z\}) Q_{N^c}(W_{1_z})$$

$$= \sum_{z \in B \backslash \partial B} [Q_{X^c}(W_{1_z}) - Q_{X^c}(\{z\} \cup W_{1_z})]\, Q_{N^c}(W_{1_z})$$

By symmetry, it follows that

$$E\,|[(X \ominus W_2^s) \cap (N \ominus W_2^s) \cap X^c]|$$

$$= \sum_{z \in B \backslash \partial B} [Q_{X^c}(W_{2_z}) - Q_{X^c}(\{z\} \cup W_{2_z})]\, Q_{N^c}(W_{2_z})$$

Next, consider

$$E\,|[(X \ominus W_1^s) \cap (X \ominus W_2^s) \cap X^c] \cap [(N \ominus W_1^s) \cap (N \ominus W_2^s)]|$$

$$= \sum_{z \in B \backslash \partial B} Pr(W_{1_z} \subseteq X,\ W_{2_z} \subseteq X,\ z \in X^c) Pr(W_{1_z} \subseteq N,\ W_{2_z} \subseteq N)$$

$$= \sum_{z \in B \setminus \partial B} Pr((W_{1_z} \cup W_{2_z}) \subseteq X,\ z \in X^c) Pr((W_{1_z} \cup W_{2_z}) \subseteq N)$$

$$= \sum_{z \in B \setminus \partial B} (Q_{X^c}(W_{1_z} \cup W_{2_z}) - Q_{X^c}(\{z\} \cup W_{1_z} \cup W_{2_z}))\, Q_{N^c}(W_{1_z} \cup W_{2_z})$$

Thus

$$E|\widehat{X} \setminus X| = \sum_{z \in B \setminus \partial B} \{(Q_{X^c}(W_{1_z}) - Q_{X^c}(\{z\} \cup W_{1_z}))\, Q_{N^c}(W_{1_z})$$

$$+ (Q_{X^c}(W_{2_z}) - Q_{X^c}(\{z\} \cup W_{2_z}))\, Q_{N^c}(W_{2_z})$$

$$- (Q_{X^c}(W_{1_z} \cup W_{2_z}) - Q_{X^c}(\{z\} \cup W_{1_z} \cup W_{2_z}))\, Q_{N^c}(W_{1_z} \cup W_{2_z})\}$$

Let us now consider

$$E|X \setminus \widehat{X}| = E|X \cap (\widehat{X})^c|$$

$$= |B \setminus \partial B| - E\left|\left(X \cap (\widehat{X})^c\right)^c\right| = |B \setminus \partial B| - E\left|X^c \cup \widehat{X}\right|$$

$$= |B \setminus \partial B| - E\,|X^c \cup [(X \cap N) \ominus W_1^s] \cup [(X \cap N) \ominus W_2^s] \cup (X \cap N)|$$

$$= |B \setminus \partial B| - E\,|(N \cup X^c) \cup [(X \cap N) \ominus W_1^s] \cup [(X \cap N) \ominus W_2^s]|$$

$$= |B \setminus \partial B| - E\,|(N \cup X^c) \cup [(X \ominus W_1^s) \cap (N \ominus W_1^s)] \cup [(X \ominus W_2^s) \cap (N \ominus W_2^s)]|$$

Observe that, for any three sets A, C, D:

$$|A \cup C \cup D| = |A| + |C| + |D|$$

$$-|A \cap C| - |A \cap D| - |C \cap D| + |A \cap C \cap D|$$

Thus

$$E|X \setminus \widehat{X}| = |B \setminus \partial B| - E|N \cup X^c|$$

$$-E|(X \ominus W_1^s) \cap (N \ominus W_1^s)|$$

$$-E|(X \ominus W_2^s) \cap (N \ominus W_2^s)|$$

$$+E|(N \cup X^c) \cap (X \ominus W_1^s) \cap (N \ominus W_1^s)|$$

$$+E|(N \cup X^c) \cap (X \ominus W_2^s) \cap (N \ominus W_2^s)|$$

$$+E|(X \ominus W_1^s) \cap (N \ominus W_1^s) \cap (X \ominus W_2^s) \cap (N \ominus W_2^s)|$$

$$-E|(N \cup X^c) \cap (X \ominus W_1^s) \cap (N \ominus W_1^s) \cap (X \ominus W_2^s) \cap (N \ominus W_2^s)|$$

We need to compute each term.

$$E|N \cup X^c| = |B \backslash \partial B| - E|(N \cup X^c)^c| = |B \backslash \partial B| - E|X \cap N^c|$$

$$= |B \backslash \partial B| - \sum_{z \in B \backslash \partial B} Pr(z \in X) Pr(z \in N^c)$$

$$= |B \backslash \partial B| - \sum_{z \in B \backslash \partial B} Q_{X^c}(\{z\})(1 - Q_{N^c}(\{z\}))$$

$$E|(X \ominus W_1^s) \cap (N \ominus W_1^s)|$$

$$= \sum_{z \in b \backslash \partial B} Pr(W_{1_z} \subseteq X) Pr(W_{1_z} \subseteq N)$$

$$= \sum_{z \in b \backslash \partial B} Pr(X^c \cap W_{1_z} = \emptyset) Pr(N^c \cap W_{1_z} = \emptyset)$$

$$= \sum_{z \in b \backslash \partial B} Q_{X^c}(W_{1_z}) Q_{N^c}(W_{1_z})$$

And so, by symmetry,

$$E|(X \ominus W_2^s) \cap (N \ominus W_2^s)|$$

$$= \sum_{z \in b \backslash \partial B} Q_{X^c}(W_{2_z}) Q_{N^c}(W_{2_z})$$

$$E\,|(N \cup X^c) \cap (X \ominus W_1^s) \cap (N \ominus W_1^s)|$$

$$= E\,|[N \cap (X \ominus W_1^s) \cap (N \ominus W_1^s 0] \cup [X^c \cap (X \ominus W_1^s) \cap (N \ominus W_1^s 0]|$$

$$= E\,|N \cap (X \ominus W_1^s) \cap (N \ominus W_1^s)|$$

$$+E\,|X^c \cap (X \ominus W_1^s) \cap (N \ominus W_1^s)|$$

$$-E\,|N \cap X^c \cap (X \ominus W_1^s) \cap (N \ominus W_1^s)|$$

$$= \sum_{z \in B \setminus \partial B} \{Pr(W_{1_z} \subseteq X) Pr(W_{1_z} \subseteq N,\ z \in N)$$

$$+Pr(W_{1_z} \subseteq X,\ z \in X^c) Pr(W_{1_z} \subseteq N)$$

$$-Pr(W_{1_z} \subseteq X,\ z \in X^c) Pr(W_{1_z} \subseteq N,\ z \in N)\}$$

$$= \sum_{z \in B \setminus \partial B} \{Pr(X^c \cap W_{1_z} = \emptyset) Pr(N^c \cap W_{1_z} = \emptyset,\ N^c \cap \{z\} = \emptyset)$$

$$+Pr(X^c \cap W_{1_z} = \emptyset,\ X^c \cap \{z\} \neq \emptyset) Pr(N^c \cap W_{1_z} = \emptyset)$$

$$-Pr(X^c \cap W_{1_z} = \emptyset,\ X^c \cap \{z\} \neq \emptyset) Pr(N^c \cap W_{1_z} = \emptyset,\ N^c \cap \{z\} = \emptyset)\}$$

$$= \sum_{z \in B \setminus \partial B} \{Q_{X^c}(W_{1_z}) Q_{N^c}(\{z\} \cup W_{1_z})$$

$$+\Gamma_{X^c,1}(W_{1_z}; \{z\}) Q_{N^c}(W_{1_z}) - \Gamma_{X^c,1}(W_{1_z}; \{z\}) Q_{N^c}(\{z\} \cup W_{1_z})\}$$

$$= \sum_{z \in B \setminus \partial B} \{Q_{X^c}(W_{1_z}) Q_{N^c}(\{z\} \cup W_{1_z})$$

$$+ \left(Q_{X^c}(W_{1_z}) - Q_{X^c}(\{z\} \cup W_{1_z})\right) Q_{N^c}(W_{1_z})$$

$$- \left(Q_{X^c}(W_{1_z}) - Q_{X^c}(\{z\} \cup W_{1_z})\right) Q_{N^c}(\{z\} \cup W_{1_z})\}$$

$$= \sum_{z \in B \setminus \partial B} \{Q_{X^c}(W_{1_z}) Q_{N^c}(\{z\} \cup W_{1_z})$$

$$+(Q_{X^c}(W_{1_z}) - Q_{X^c}(\{z\} \cup W_{1_z}))(Q_{N^c}(W_{1_z}) - Q_{N^c}(\{z\} \cup W_{1_z}))\}$$

Thus, by symmetry

$$E\,|(N \cup X^c) \cap (X \ominus W_2^s) \cap (N \ominus W_2^s)|$$

$$= \sum_{z \in B \setminus \partial B} \{Q_{X^c}(W_{2_z}) Q_{N^c}(\{z\} \cup W_{2_z})$$

$$+(Q_{X^c}(W_{2_z}) - Q_{X^c}(\{z\} \cup W_{2_z}))(Q_{N^c}(W_{2_z}) - Q_{N^c}(\{z\} \cup W_{2_z}))\}$$

Also

$$E\,|(X \ominus W_1^s) \cap (N \ominus W_1^s) \cap (X \ominus W_2^s) \cap (N \ominus W_2^s)|$$

$$= \sum_{z \in B \setminus \partial B} Pr(W_{1_z} \subseteq X,\ W_{2_z} \subseteq X) Pr(W_{1_z} \subseteq N,\ W_{2_z} \subseteq N)$$

$$= \sum_{z \in B \setminus \partial B} Pr(X^c \cap W_{1_z} = \emptyset,\ X^c \cap W_{2_z} = \emptyset) Pr(N^c \cap W_{1_z} = \emptyset,\ N^c \cap W_{2_z} = \emptyset)$$

$$= \sum_{z \in B \setminus \partial B} Pr(X^c \cap (W_{1_z} \cup W_{2_z}) = \emptyset) Pr(N^c \cap (W_{1_z} \cup W_{2_z}) = \emptyset)$$

$$= \sum_{z \in B \setminus \partial B} Q_{X^c}(W_{1_z} \cup W_{2_z}) Q_{N^c}(W_{1_z} \cup W_{2_z})$$

Finally, let us compute

$$E\,|(N \cup X^c) \cap (X \ominus W_1^s) \cap (N \ominus W_1^s) \cap (X \ominus W_2^s) \cap (N \ominus W_2^s)|$$

$$= E\,|[N \cap (X \ominus W_1^s) \cap (N \ominus W_1^s) \cap (X \ominus W_2^s) \cap (N \ominus W_2^s)]$$

$$\cup [X^c \cap (X \ominus W_1^s) \cap (N \ominus W_1^s) \cap (X \ominus W_2^s) \cap (N \ominus W_2^s)]|$$

$$= E\,|N \cap (X \ominus W_1^s) \cap (N \ominus W_1^s) \cap (X \ominus W_2^s) \cap (N \ominus W_2^s)|$$

$$+E\,|X^c \cap (X \ominus W_1^s) \cap (N \ominus W_1^s) \cap (X \ominus W_2^s) \cap (N \ominus W_2^s)|$$

$$-E\,|N \cap X^c \cap (X \ominus W_1^s) \cap (N \ominus W_1^s) \cap (X \ominus W_2^s) \cap (N \ominus W_2^s)|$$

$$= \sum_{z \in B \setminus \partial B} \{Q_{X^c}(W_{1_z} \cup W_{2_z}) Q_{N^c}(\{z\} \cup W_{1_z} \cup W_{2_z})$$

$$+\Gamma_{X^c,1}(W_{1_z} \cup W_{2_z}; \{z\}) Q_{N^c}(W_{1_z} \cup W_{2_z})$$

$$-\Gamma_{X^c,1}(W_{1_z} \cup W_{2_z}; \{z\}) Q_{N^c}(\{z\} \cup W_{1_z} \cup W_{2_z})\}$$

$$= \sum_{z \in B \setminus \partial B} \{Q_{X^c}(W_{1_z} \cup W_{2_z}) Q_{N^c}(\{z\} \cup W_{1_z} \cup W_{2_z})$$

$$+ \left(Q_{X^c}(W_{1_z} \cup W_{2_z}) - Q_{X^c}(\{z\} \cup W_{1_z} \cup W_{2_z})\right) Q_{N^c}(W_{1_z} \cup W_{2_z})$$

$$- \left(Q_{X^c}(W_{1_z} \cup W_{2_z}) - Q_{X^c}(\{z\} \cup W_{1_z} \cup W_{2_z})\right) Q_{N^c}(\{z\} \cup W_{1_z} \cup W_{2_z})\}$$

$$= \sum_{z \in B \setminus \partial B} \{Q_{X^c}(W_{1_z} \cup W_{2_z}) Q_{N^c}(\{z\} \cup W_{1_z} \cup W_{2_z})$$

$$+ \left(Q_{X^c}(W_{1_z} \cup W_{2_z}) - Q_{X^c}(\{z\} \cup W_{1_z} \cup W_{2_z})\right)$$

$$- \left(Q_{N^c}(W_{1_z} \cup W_{2_z}) - Q_{N^c}(\{z\} \cup W_{1_z} \cup W_{2_z})\right)\}$$

So now we have expressions for all the components of $E|X \setminus \widehat{X}|$, plus an expression for $E|\widehat{X} \setminus X|$, all in terms of the generating functional of X^c, as well as that of N^c. By adding them up, we obtain an expression for the total expected error, which, after some simplification, yields

$$E(e) = \sum_{z \in B \setminus \partial B} \{Q_{X^c}(\{z\}) \left(1 - Q_{N^c}(\{z\})\right)$$

$$+Q_{X^c}(W_{1_z}) Q_{N^c}(W_{1_z}) + Q_{X^c}(W_{2_z}) Q_{N^c}(W_{2_z})$$

$$+Q_{X^c}(\{z\} \cup W_{1_z}) Q_{N^c}(\{z\} \cup W_{1_z}) + Q_{X^c}(\{z\} \cup W_{2_z}) Q_{N^c}(\{z\} \cup W_{2_z})$$

$$-2Q_{X^c}(\{z\} \cup W_{1_z}) Q_{N^c}(W_{1_z}) - 2Q_{X^c}(\{z\} \cup W_{2_z}) Q_{N^c}(W_{2_z})$$

$$-Q_{X^c}(W_{1_z} \cup W_{2_z}) Q_{N^c}(W_{1_z} \cup W_{2_z}) - Q_{X^c}(\{z\} \cup W_{1_z} \cup W_{2_z}) Q_{N^c}(\{z\} \cup W_{1_z} \cup W_{2_z})$$

$$+2Q_{X^c}(\{z\} \cup W_{1_z} \cup W_{2_z})Q_{N^c}(W_{1_z} \cup W_{2_z})\}$$

which is the desired formula. □

Again, by assuming that X, N are obtained by sampling stationary RS's, we can obtain a characterization of the optimal pair of structuring elements in terms of the probability of pixel error. Under appropriate smoothness conditions, we can reduce the complexity of the search for the optimal pair in a manner similar to the one of the previous subsection. The details are straightforward, but cumbersome. Obviously, by duality, similar results can be obtained for the case of union noise, as well as for more than two structuring elements.

4.4.5 Experimental Results

In order to corroborate our theoretical results, we have designed a series of simulation experiments. One such experiment is described here in detail. The results of another experiment, involving a real-life image, are also presented. These experiments are solely intended to serve as "proof of concept". No claims are made regarding the relative merit of our approach as measured against other approaches in the literature. A thorough comparative study of different filter structures is analytically impossible, and it therefore requires extensive simulation, which is beyond our present scope. Our purpose here is to demonstrate that our theoretical results actually make sense in practice.

Let us make the assumptions of corollary 4.1. For the purposes of simulation, we need models for the signal and noise. We assume that the signal, X, is a

DRBRS of constant intensity, and that the noise, N, is given by the set of points of a Bernoulli random field, of constant intensity $p = 0.9$. These models are only used to generate realizations of the signal, the noise, and the observation. The entire simulation is data driven, and all relevant probabilities are estimated using running averages. This approach is "honest", and close to real world problems.

We also have to choose a collection of structuring elements, over which we will optimize. We consider the collection depicted in figure 4.5, and label the structuring elements $W_1, \cdots, W_4$, from left to right.

Figure 4.7 depicts a realization of the signal DRS, X, while figure 4.8 depicts a realization of the noise DRS N. These are solely used to estimate the relevant probabilities. The results for the signal and the noise are tabulated in tables 4.1 and 4.2, respectively. The results for the estimated probability of pixel error are tabulated in table 4.3. These have been computed using tables 4.1,4.2, and the formula of corollary 4.1. In table 4.3, the leftmost entry is the estimated probability of pixel error between the signal, X, and the observation, $X \cap N$, i.e., when no filtering takes place (this corresponds to $W = \{\bar{0}\}$). It is given here for comparison purposes. Clearly, the optimal structuring element is W_2, with W_1 running a close second (this is justified by the symmetry in the data). The worst structuring element is W_4.

Figure 4.9 depicts another (independent) realization of X, while figure 4.10 depicts a realization of the observation, $Y = X \cap N$, obtained by intersecting the DRS realization of figure 4.9 with an independent realization of N. Figure

4.11 depicts the restored image, $\widehat{X} = (Y \ominus W_2^s) \cup Y$, where Y is the DRS realization of figure 4.10. This is the best possible restoration within the given family of structuring elements. For comparison purposes, figure 4.12 depicts the restored image, $\widetilde{X} = (Y \ominus W_4^s) \cup Y$, where Y is the DRS realization of figure 4.10. This is the worst (non-trivial) restoration within the given family of structuring elements. Close inspection of these figures reveals several interesting phenomena. In particular, even though W_2 does a better job than W_4 in filling up gaps introduced by noise, it also bridges together shape components which were originally disconnected. This is evident in the upper right-hand part of the figures. Nevertheless, this source of error is counter-balanced by the relative effectiveness of W_2 in terms of noise elimination. As a result, the overall quality of restoration achieved by W_2 is still visibly better. However, under a low-noise scenario, this situation can be reversed, i.e., retaining the connectivity structure of the signal will become more important, and, eventually, it will supersede noise elimination as the dominant factor. In this case, W_4 will provide superior performance.

These simulation results are encouraging; they clearly support the theory and satisfy our intuition. Furthermore, considering the fact that the optimal filter only requires two set translations and two set unions (two translations and one union are needed to implement the erosion with W_2), the quality of restoration seems good. Even better results can be achieved using multiple structuring elements.

In real life, image statistics are often spatially varying. In most cases, it is possible to model such images as piecewise statistically invariant. This approach is taken quite often when Markov random field (MRF) models are used. In this setting, one can either segment the image in disjoint regions with approximately invariant statistics *prior* to the filtering step, or rely on the filtering algorithm to perform simultaneous signal estimation *and* segmentation. Either way, this is not a trivial task, and these algorithms are generally not amenable to in-depth statistical analysis. Therefore, it is of interest to test our algorithms on images which violate our assumptions. For this purpose, consider figure 4.13. It depicts a binary version of the well-known "Lena" picture. This picture can be modeled as piecewise statistically invariant. However, let us bypass the segmentation step and blindly apply our algorithm. Figure 4.14 depicts a version of Lena which has been degraded by a combination of burst and memoryless transmission errors. Specifically, we assume that the image is scanned row-wise, and individual bits are transmitted over a channel which is memoryless most of the time, but occasionally switches to a burst error channel. We assume that the noise only affects the white part of the image, i.e. a union noise model. The signal and noise statistics were estimated from the original picture and an independent realization of the noise. Then, based on the "dual" of proposition 4.8, as it applies to the case of sampling stationary RS's, the optimal two-fold dilation filter was sought within the class of filters with structuring elements in the collection of figure 4.5. The optimal pair of structuring elements was found

to be (W_2, W_4). The restored image is depicted in figure 4.15. Given that we have violated the stationarity assumption, the overall result is surprisingly good.

	$W = \{\bar{0}\}$	W = W_1	$W = W_2$	$W = W_3$	$W = W_4$
$Q_{X^c}(W)$	0.494698	0.438927	0.438970	0.426667	0.383403
$Q_{X^c}(\{\bar{0}\} \cup W)$		0.437678	0.437682	0.426082	0.383403

Table 4.1: Estimated values of the generating functional $Q_{X^c}(\cdot)$.

	$W = \{\bar{0}\}$	W = W_1	$W = W_2$	$W = W_3$	$W = W_4$
$Q_{N^c}(W)$	0.900431	0.810915	0.810519	0.657578	0.432145
$Q_{N^c}(\{\bar{0}\} \cup W)$		0.730154	0.729550	0.591795	0.389004

Table 4.2: Estimated values of the generating functional $Q_{N^c}(\cdot)$.

	$W = \{\bar{0}\}$	W = W_1	$W = W_2$	$W = W_3$	$W = W_4$
$P_{pixel\ error}(W)$	0.0493	0.01491	0.01490	0.0217	0.0328

Table 4.3: Estimated values of the probability of pixel error.

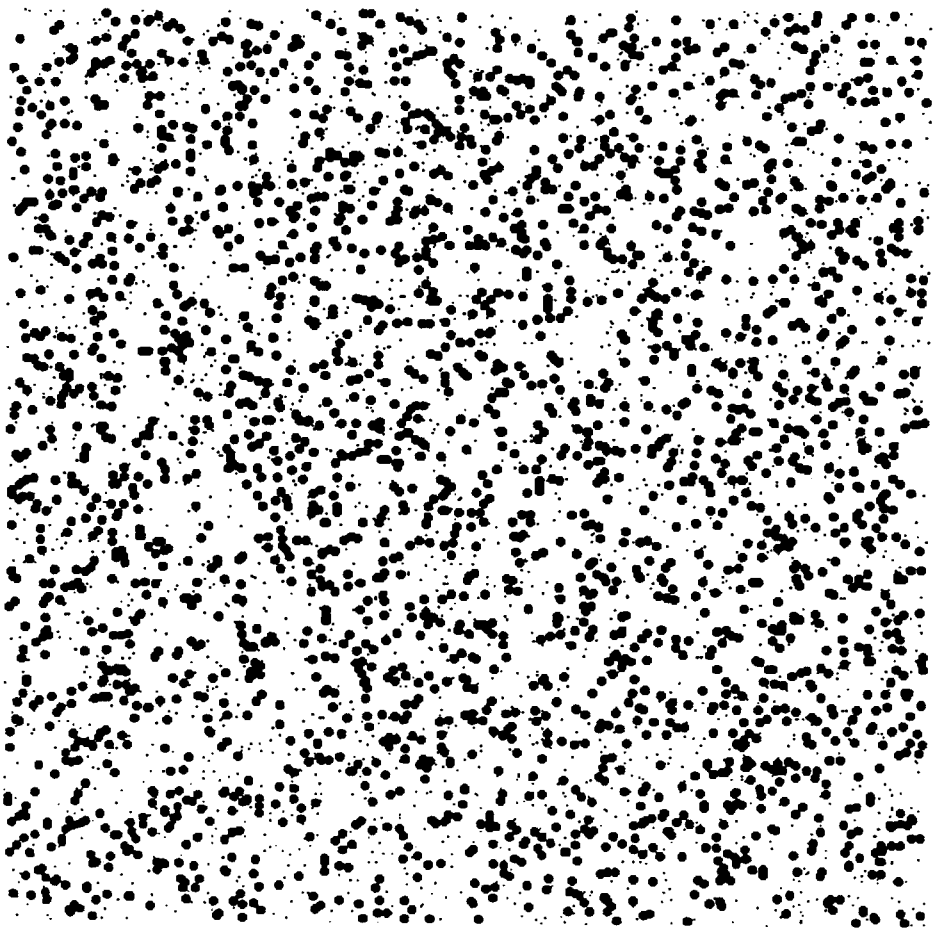


Figure 4.1: A realization of a DRBRS corrupted by iid union noise

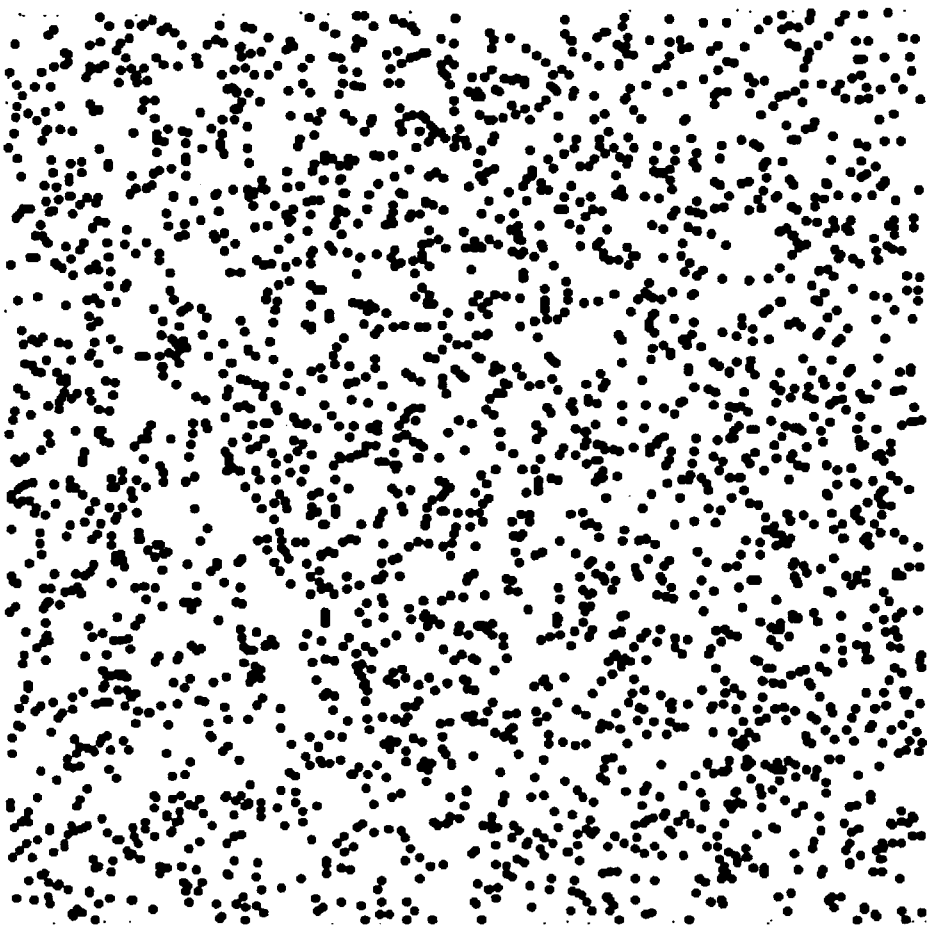


Figure 4.2: Restored image, obtained by using the optimal adaptive mask filter.

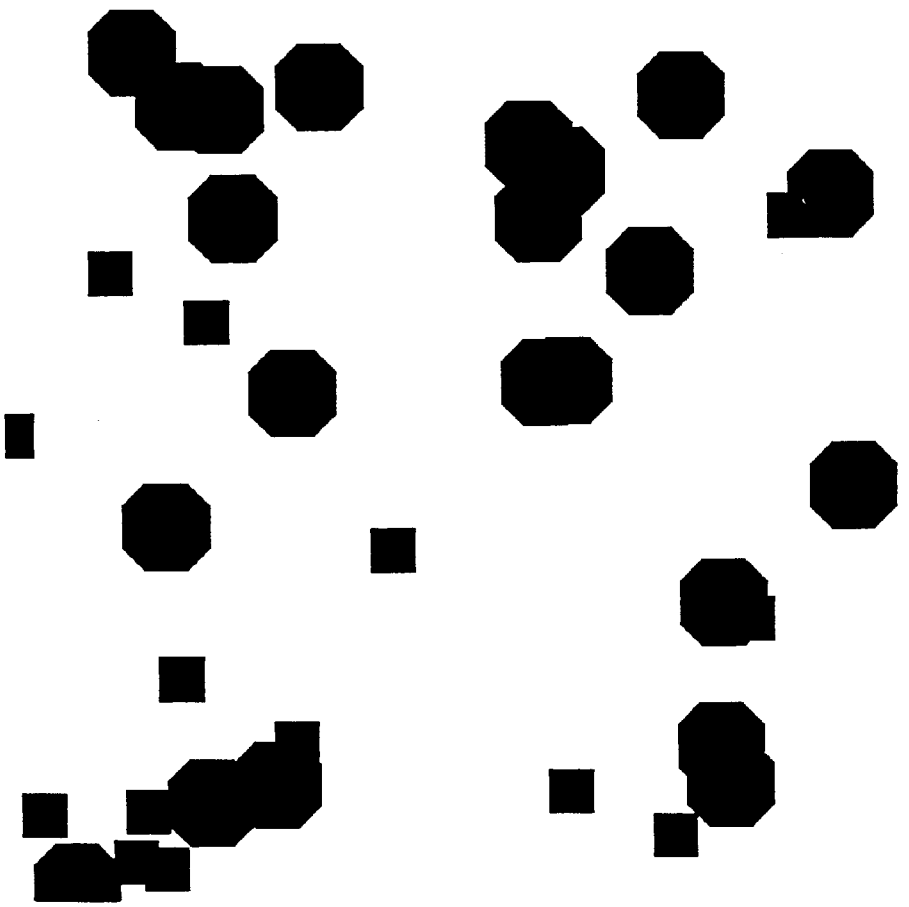


Figure 4.3: A realization of the union of two DRBRS models

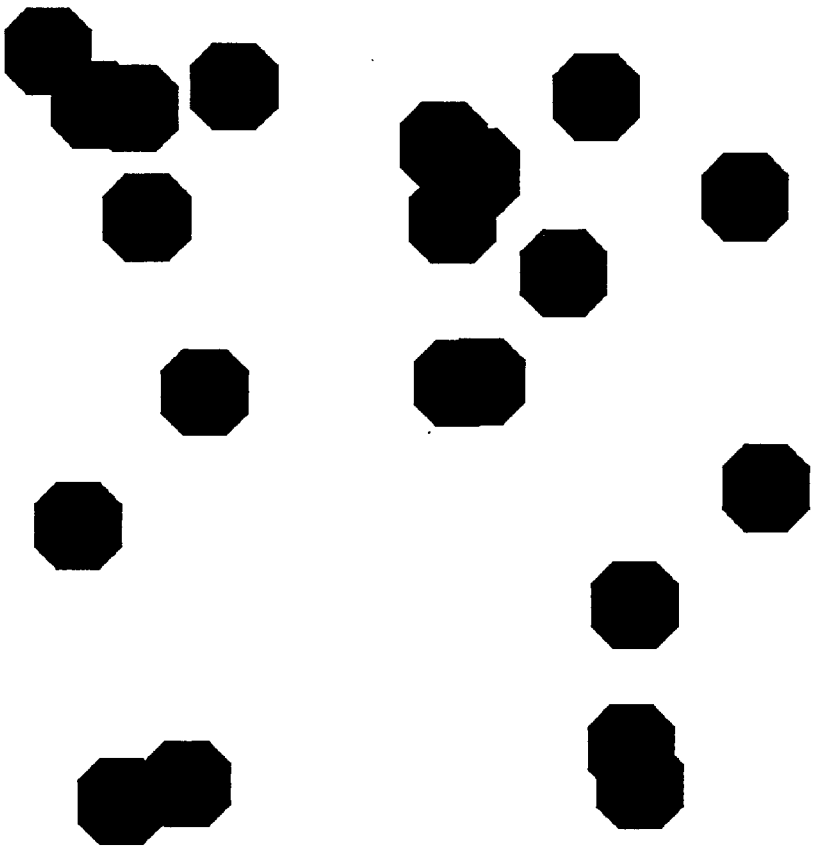


Figure 4.4: Restored image, obtained by using the optimal adaptive mask filter.

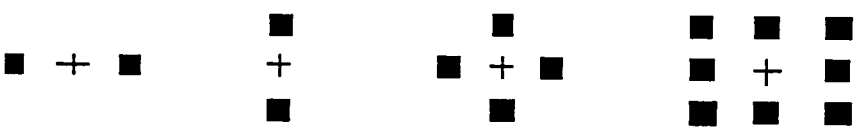


Figure 4.5: Some structuring elements that can be used in a "gap-filling" mode.

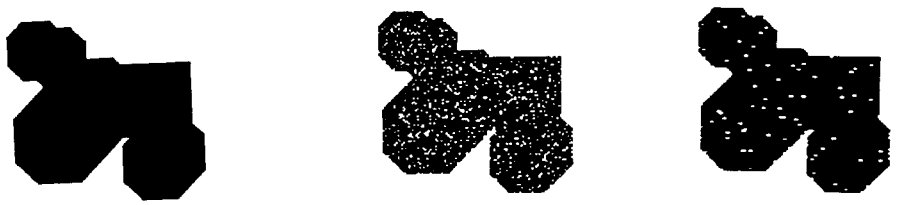


Figure 4.6: (a) Original image, (b) Intersection of the image in (a) with a Bernoulli random field, (c) Restored image.

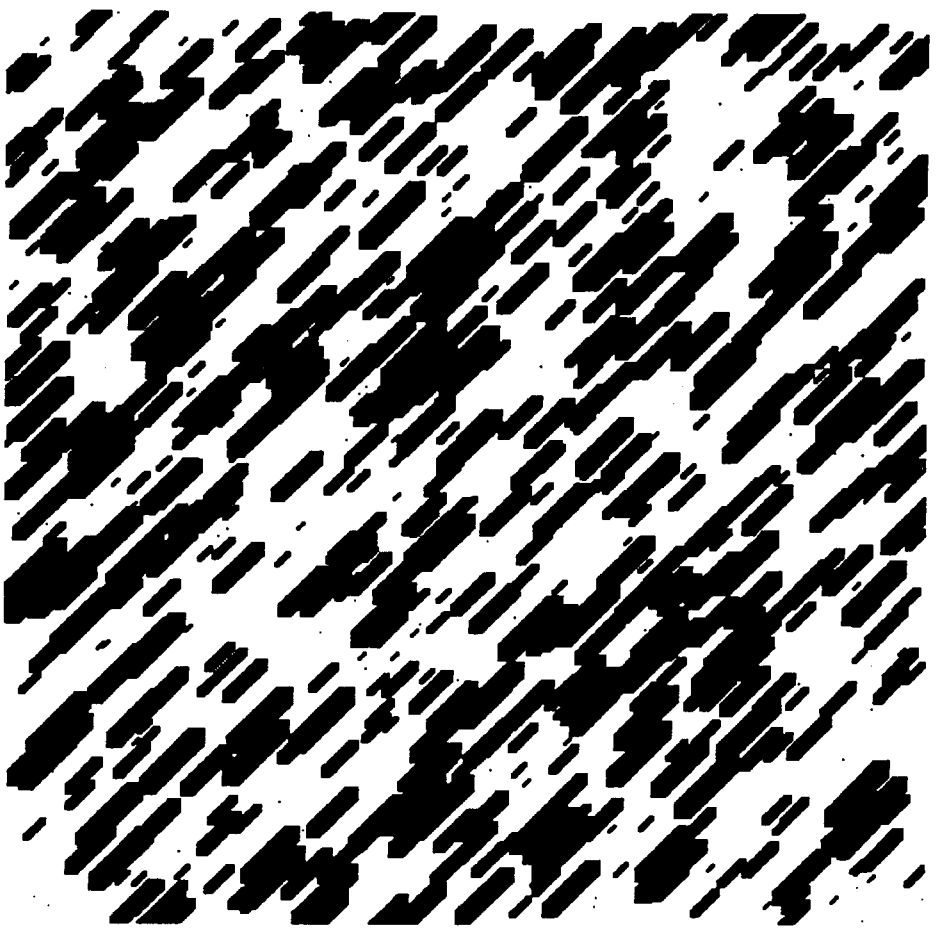


Figure 4.7: A realization of the signal DRS, X.

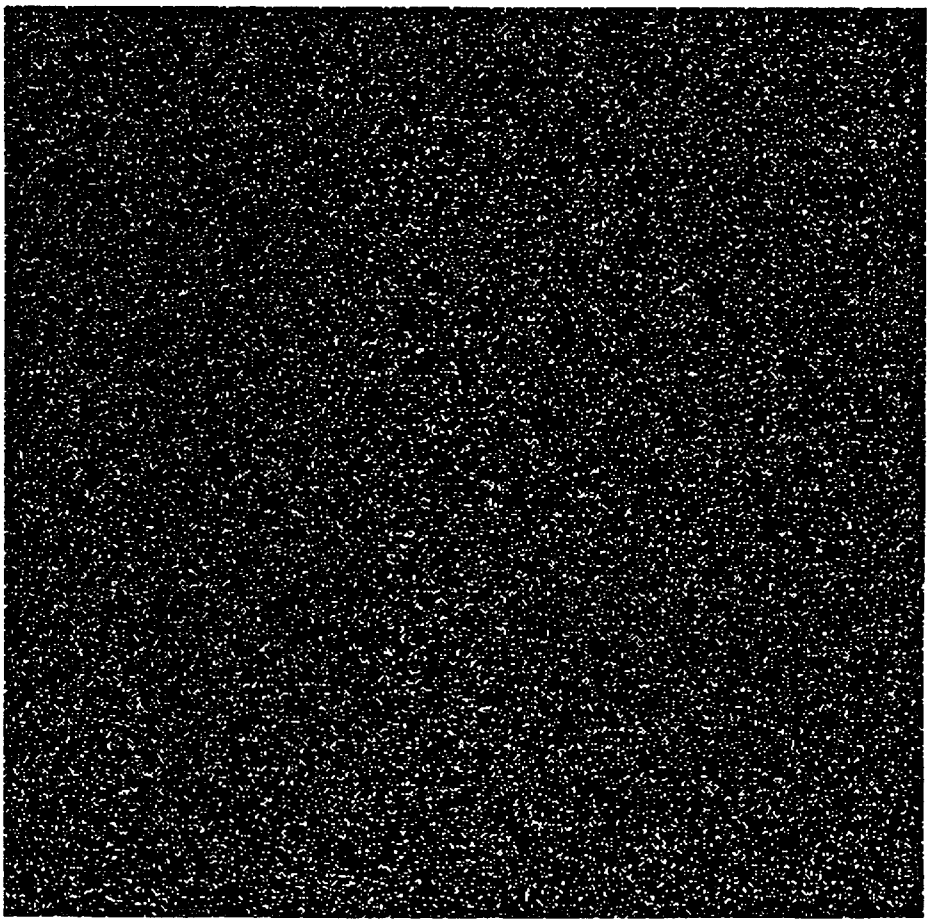


Figure 4.8: A realization of the noise DRS, N.

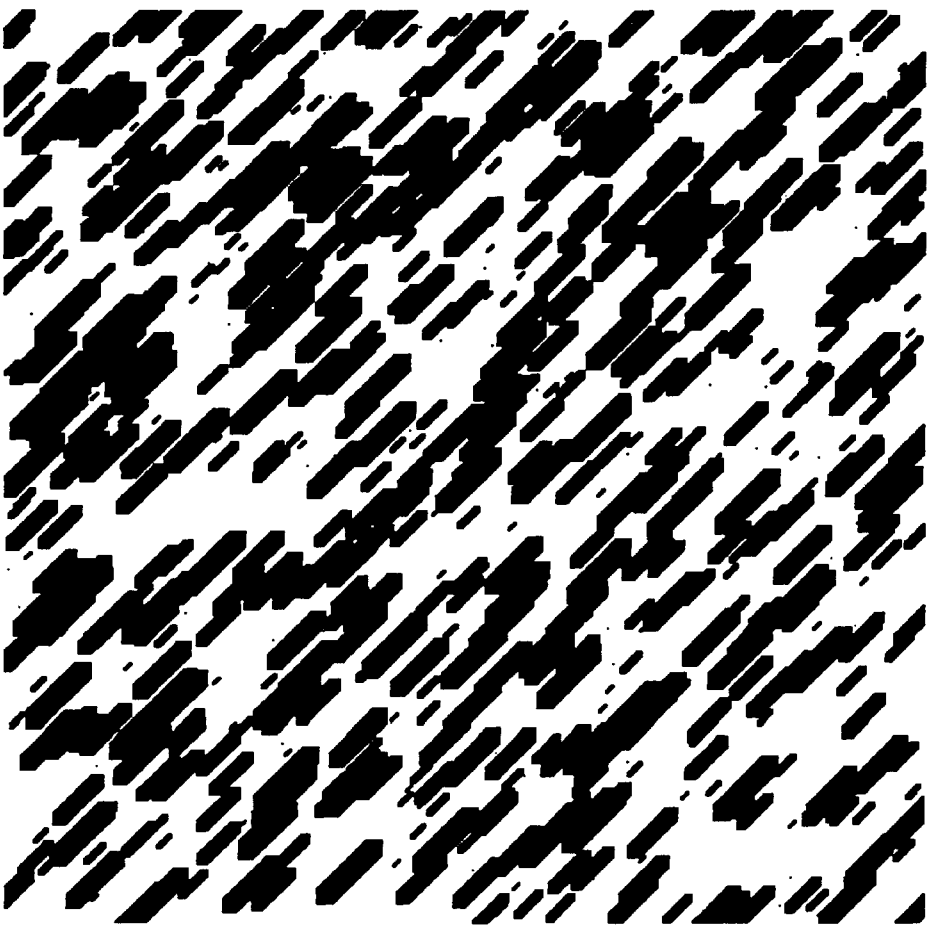


Figure 4.9: Another (independent) realization of the signal DRS, X.

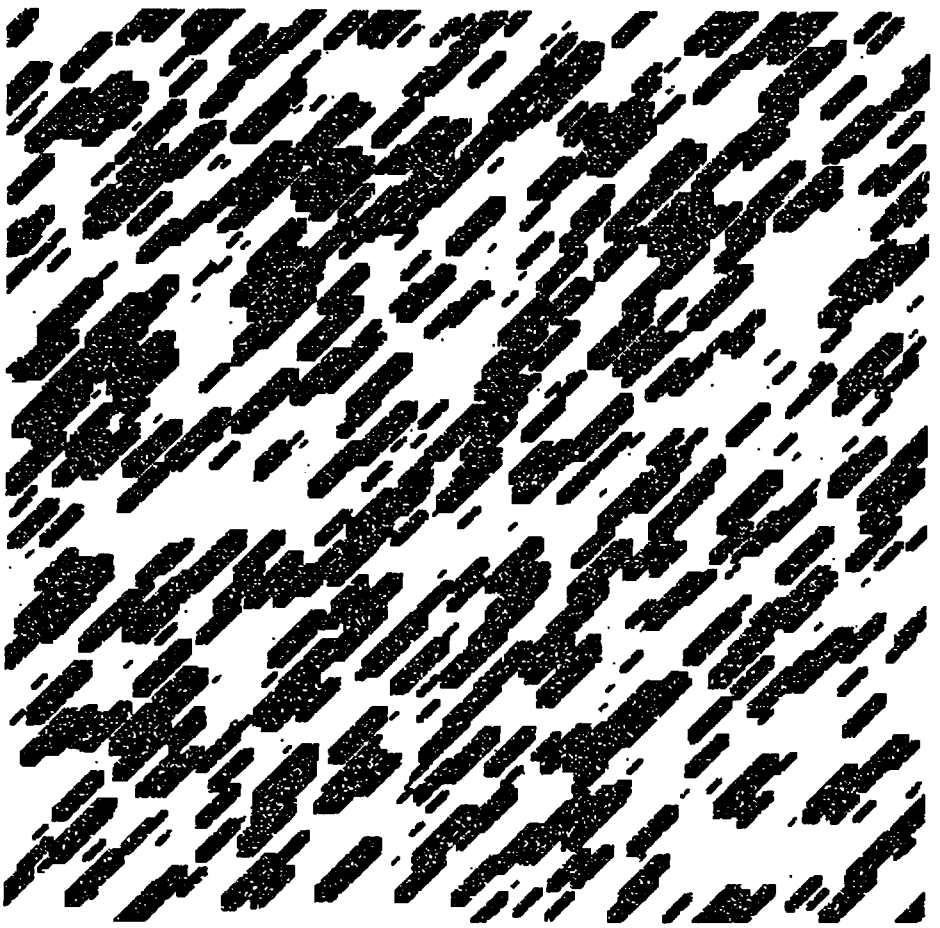


Figure 4.10: The result of intersecting the DRS realization of figure 4.9 with another (independent) realization of the noise DRS, N.

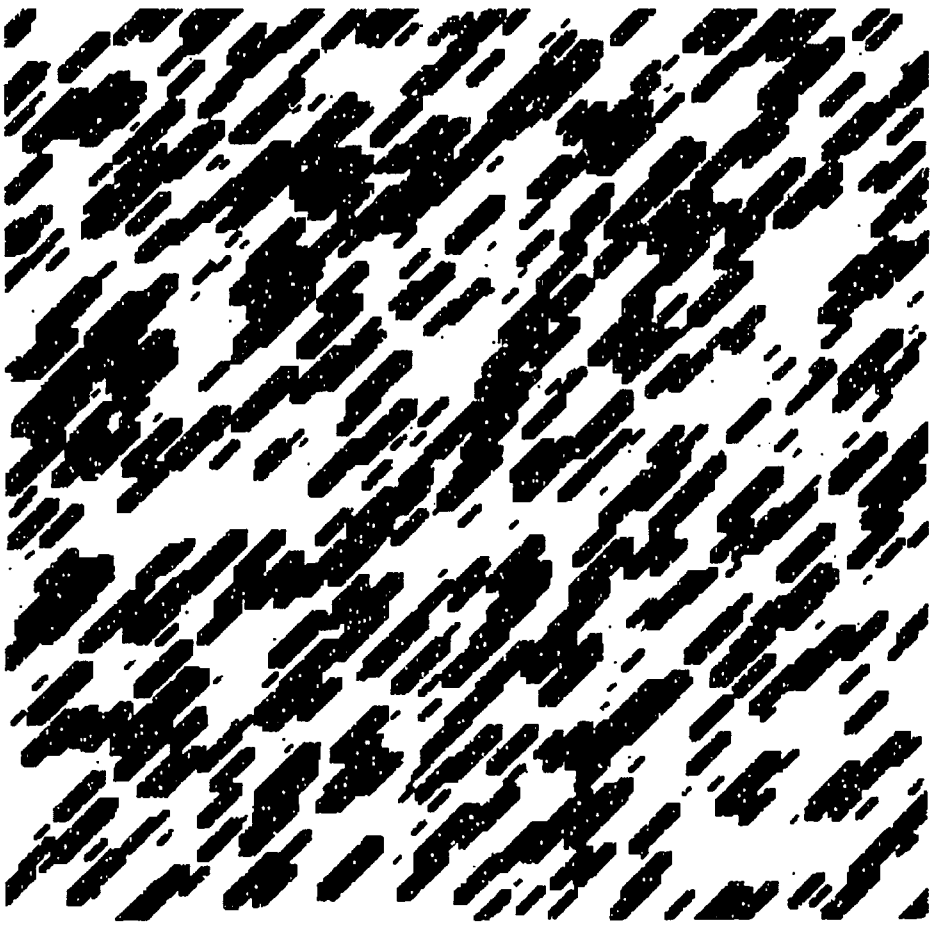


Figure 4.11: Restored image, obtained by filtering the DRS realization of figure 4.10 using structuring element W_2 (the best one).

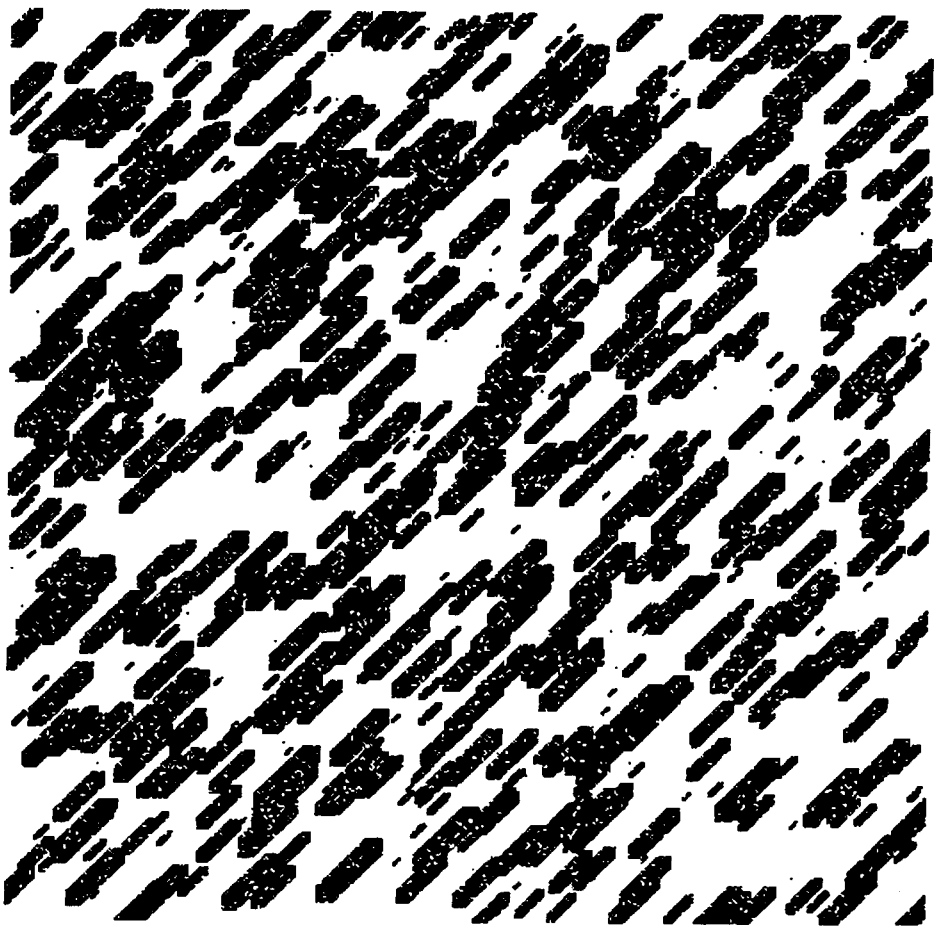


Figure 4.12: Restored image, obtained by filtering the DRS realization of figure 4.10 using structuring element W_4 (the worst one).

Figure 4.13: Binary Lena picture.

Figure 4.14: Lena picture, degraded by a combination of burst and memoryless transmission errors.

Figure 4.15: Restored Lena picture.

CHAPTER
FIVE

PROBING BINARY SHAPES IN CLUTTER

5.1 Introduction

In loose terms, adaptive probabilistic probing is a flexible model and data-driven sequential procedure which enables one to carry out incremental image data classification and hypothesis verification tasks. The method can be based on the theory of probabilistic classification trees [53, 22, 7, 59, 72]. The basic idea is to break the classification problem into a sequence of appropriately chosen, "highly informative" image feature evaluations. It is assumed that features are collected by "probing" the image using some simple primitive shape ("the probe"), which conveys information about the geometrical and topological structure of the image. During each step, the method proceeds by evaluating an appropriately chosen feature, which is selected based on all the features which were previously evaluated, as well as their respective values.

Adaptive sequential classification procedures become attractive when it is too difficult to find, design, or implement the optimal Bayesian classification

rule [6]. The basic idea is to shift the burden of complex decision-making from run-time to the design stage, with some subsequent loss of performance [22, 2]. In addition, classification tree - based procedures can provide considerable insight into the decision-making process, because they decompose the classification problem into sequences of simpler problems [6]. However, the optimal design of such procedures is very difficult.

This part of our work constitutes a *preliminary* attempt to apply the theory of Probabilistic Classification Trees to a specific problem: that of classifying discrete and binary "random" shapes hidden in clutter. No contribution is made to the theory of Probabilistic Classification Trees. Here, we are interested in *recognition algorithms*, and our aim is to bring together several key ideas, and highlight the benefits, as well as the drawbacks, of an adaptive probabilistic probing approach. The outlook is as follows. We first formulate the problem of classification of discrete and binary "random" shapes in clutter. Then we derive the MAP rule, show how it can be implemented, and discuss its computational complexity. The next step is to introduce Kendall's "trapping system" idea, show how a trapping system can be used to test hypotheses, and discuss various related design problems. A brief introduction to Probabilistic Classification Trees follows. Then all the above ideas are fused together to develop adaptive probabilistic probing.

5.2 Classification of discrete and binary shapes hidden in clutter

Assume we are given $\mathcal{C}$, a finite collection of discrete, binary, and bounded shapes (subsets of some bounded set $B \subset \mathbf{Z}^2$). Let C be a $\mathcal{C}$-valued DRS. Typically, $|\mathcal{C}|$ is small, and, therefore, it is convenient to specify C via its probability mass function

$$Pr(C = c) = \pi_0(c), \quad \forall c \in \mathcal{C}$$

Let N be another DRS on B, which is used to model the combined effect of sensor/channel noise, and "random" obscuration because of clutter. As such, N is naturally specified via its generating functional

$$Q_N(K) = Pr(N \cap K = \emptyset), \quad \forall K \in \Sigma(B)$$

It is assumed that C and N are mutually independent, and that the observable DRS, Y is

$$Y = C \cup N$$

The problem is to classify the observation, i.e. decide upon the operative C given Y. A standard approach is to pick $c \in \mathcal{C}$ to minimize the total probability of error. This is equivalent to the Maximum A Posteriori (MAP) rule. It calls for the maximization of

$$Pr(C = c \mid Y = K)$$

by means of a suitable choice of c, where K is the observed realization of Y, i.e.

$$\hat{c} = arg\ max_{c \in \mathcal{C}} Pr(C = c \mid Y = K)$$

Equivalently, using Bayes' rule

$$\hat{c} = arg\ max_{c \in \mathcal{C}} \pi_0(c) Pr(Y = K \mid C = c)$$

Hence, it is crucial to compute $Pr(Y = K \mid C = c)$. Observe that

$$Q_{Y|C=c}(K) \triangleq Pr(Y \cap K = \emptyset \mid C = c) =$$

$$Pr((C \cup N) \cap K = \emptyset \mid C = c) = \ (by\ independence\ of\ C, N)\ =$$

$$Pr((c \cup N) \cap K = \emptyset) = \begin{cases} 0 & , if\ c \cap K \neq \emptyset \\ \\ Q_N(K) & , otherwise \end{cases}$$

Employing the fundamental theorem 2.1, we obtain

$$Pr(Y = K \mid C = c) = \sum_{K' \subseteq K} (-1)^{|K'|} Q_{Y|C=c}(K^c \cup K')$$

for all $K \in \Sigma(B)$. Given an observation, $K \in \Sigma(B)$, we need to compute the above probability for each $c \in \mathcal{C}$. Each one of these $|\mathcal{C}|$ computations takes $O(2^{|K|})$ operations, for an overall complexity of $O(|\mathcal{C}| \times 2^{|K|})$ operations, which is clearly not acceptable for most practical applications. Furthermore, these computations have to be done *online*, because of the inherent enumeration and indexing problems associated with set-valued random variables. Observe though that there is almost no design cost - all computations are differed until the actual runtime. This is typical of the classical Bayesian approach. Later we will see how we can trade design (computational) complexity and some loss of performance for runtime efficiency.

5.3 Kendall's trapping system and ramifications

We now digress a bit, to describe a fundamental idea, originally due to Kendall [34], which has been instrumental in the early development of a theory of continuous domain random sets. It will also provide considerable insight into our problem.

Assume we are given a collection of "traps" (subsets of $B \subset \mathbf{Z}^2$) $\mathcal{T} = \{T_i\}$, with the property $\bigcup_{T_i \in \mathcal{T}} T_i = B$, and a set of "shapes", $\mathcal{C} = \{c_j\}$, as before. Assume some (unknown) operative $c \in \mathcal{C}$, and suppose we are given hit/miss information over $\mathcal{T}$, i.e. for each $T \in \mathcal{T}$ we are told whether the operative c hits (has a nonempty intersection with) T or misses it (has an empty intersection with T). Let us pose the following question: in what sense (if any) does

$$\hat{c} \triangleq B \setminus \left(\bigcup_{T_i \in \mathcal{T}\ ,\ T_i \cap c = \emptyset} T_i \right)$$

convey information about c? Under which conditions $\hat{c} = c$? In general, $\hat{c}$ is a "smooth version" of c. In particular, let H be a small convex primitive which contains the origin, and assume that $\mathcal{T}$ comprises of all translations of H which fit within B. Observe that the requirement $\bigcup_{T_i \in \mathcal{T}} T_i = B$ implies that $B \circ H = B$, i.e. H must be such that B is H-open. Let us further assume that $\mathcal{C}$ is such that

$$B \setminus ((B \setminus c) \circ H) = c \bullet H, \quad \forall c \in \mathcal{C}$$

This condition is needed to secure that no problems arise due to edge effects. It essentially requires that all $c \in \mathcal{C}$ lay sufficiently far (with respect to H) from

the border of B. Under these assumptions it can be seen that $\widehat{c} = c \bullet H$, i.e. the Morphological closing of c by H. In addition:

Theorem 5.1 *(Kendall [34]) If, for each $c \in \mathcal{C}$, $B \setminus c$ can be written as a union of traps $T_i \in \mathcal{T}$, $i \in I(c)$, where $I(c)$ is an index set which depends on c*

$$B \setminus c = \bigcup_{T_i \in \mathcal{T},\ i \in I(c)} T_i$$

and

$$\bigcup_{T_i \in \mathcal{T}} T_i = B$$

then $\widehat{c} = c$, and the largest collection $\mathcal{C}$ for which the first condition above is true is called the set of $\mathcal{T}$-closed sets.

In particular, if $\mathcal{T}$ is the set of all allowable translations of H, then the corresponding collection of $\mathcal{T}$-closed sets is exactly the collection of all *Morphologically H-closed* subsets of B which satisfy the condition $B \setminus ((B \setminus c) \circ H) = c \bullet H$. Finally, if H is taken to be the origin (i.e. a point trap) then the set of $\mathcal{T}$-closed sets is $\Sigma(B)$, and $\widehat{c} = c$, $\forall c \in \Sigma(B)$. Therefore, there always exists some trapping system $\mathcal{T}$ such that the range of *any* DRS Y is a subset of the set of $\mathcal{T}$-closed sets.

These simple, yet important, observations have the following implication: modulo some possible (but controllable) loss of detail, we can look at hit/miss information over a suitable trapping system instead of the image per se, and still be able to recover the operative c.

Assume now that we are given $\mathcal{T}$ and $\mathcal{C} \subseteq \Sigma(B)$, the associated set of $\mathcal{T}$-closed sets. Then the *hitting map*

$$h_{\mathcal{T}} : \mathcal{C} \mapsto \{0,1\}^N, \quad N = |\mathcal{T}|$$

defined by

$$h_{\mathcal{T}}(K) = [f_{T_1}(K), \cdots, f_{T_N}(K)]^T, \ \forall K \in \mathcal{C}$$

with

$$f_{T_i}(K) = \begin{cases} 1 \ , \ K \cap T_i \neq \emptyset \\ \\ 0 \ , \ otherwise \end{cases}$$

is *invertible*, the inverse

$$h_{\mathcal{T}}^{-1} : \{0,1\}^N \mapsto \mathcal{C}$$

given by

$$h_{\mathcal{T}}^{-1}\left([f_{T_1}, \cdots, f_{T_N}]^T\right) = B \setminus \left(\bigcup_{i=1,\cdots,N,\ f_{T_i}=0} T_i \right)$$

Therefore, returning to the classification problem of the previous section, we can always choose a suitable trapping system, $\mathcal{T}$, such that the range of the observable DRS $Y = C \cup N$ is a subset of the set of $\mathcal{T}$-closed sets, in which case $Pr(Y = K \mid C = c)$ can be computed recursively, based only on the data $h_{\mathcal{T}}(K)$ and $Q_{Y|C=c}$.

$$Pr(Y = K \mid C = c) = Pr(f_{T_1}(Y) = f_{T_1}(K), \cdots, f_{T_N}(Y) = f_{T_N}(K) \mid C = c)$$

Let us recall the inclusion-exclusion principle (cf. the proof of proposition 4.6). For an arbitrary DRS X, on B, and $n+1$ elements of $\Sigma(B)$, $K_0, K_1, \cdots, K_n$,

define

$$\Gamma_{X,n}(K_0; K_1, \cdots, K_n) \triangleq Pr(X \cap K_0 = \emptyset \ , \ X \cap K_1 \neq \emptyset, \cdots, X \cap K_n \neq \emptyset)$$

By definition, $\Gamma_{X,0}(K) = Q_X(K)$. Using Bayes' rule, one can easily show that this functional satisfies the following recursion (*the inclusion - exclusion principle*).

$$\Gamma_{X,n}(K_0; K_1, \cdots, K_n) = \Gamma_{X,n-1}(K_0; K_1, \cdots, K_{n-1})$$

$$-\Gamma_{X,n-1}(K_0 \cup K_n; K_1, \cdots, K_{n-1})$$

Let $T_{i_1}, \cdots, T_{i_n}$ be the collection of traps which are hit by K, i.e. those with $f_T(K) = 1$. Let $T_{j_1}, \cdots, T_{j_{N-n}}$ be the remaining ones. Let $T = T_{j_1} \cup \cdots \cup T_{j_{N-n}}$. Then

$$Pr(Y = K \mid C = c) = \Gamma_{Y|C=c,n}(T; T_{i_1}, \cdots, T_{i_n}) =$$

$$\Gamma_{Y|C=c,n-1}(T; T_{i_1}, \cdots, T_{i_{n-1}}) - \Gamma_{Y|C=c,n-1}(T \cup T_{i_n}; T_{i_1}, \cdots, T_{i_{n-1}})$$

By iterating, we obtain

$$Pr(Y = K \mid C = c) = \sum_{m=0}^{n} (-1)^m \sum_{T' \in S_m^n(T_{i_1}, \cdots, T_{i_n})} Q_{Y|C=c}(T \cup T')$$

where $S_m^n(T_{i_1}, \cdots, T_{i_n})$ is the collection of all possible unions of exactly m out of the n sets $T_{i_1}, \cdots, T_{i_n}$.

Let us now consider the "short horizon" case. Suppose we want to build a classification procedure based on $M << N$ traps. In this setting, we would like to pick M out of the N possible traps in some sort of optimal fashion. One way

to do this is to pick the traps to minimize the average uncertainty about the "class r.v." C, given hit/miss information over the chosen set of traps, i.e.

$$\left\{T_{i_1^*}, \cdots, T_{i_M^*}\right\} = arg\ min_{\{T_{i_1}, \cdots, T_{i_M}\} \subseteq \mathcal{T}} H\left(C \mid f_{T_{i_1}}(Y), \cdots, f_{T_{i_M}}(Y)\right) =$$

$$arg\ max_{\{T_{i_1}, \cdots, T_{i_M}\} \subseteq \mathcal{T}} \left[H(C) - H\left(C \mid f_{T_{i_1}}(Y), \cdots, f_{T_{i_M}}(Y)\right)\right] =$$

$$arg\ max_{\{T_{i_1}, \cdots, T_{i_M}\} \subseteq \mathcal{T}} I\left(C\ ;\ \left(f_{T_{i_1}}(Y), \cdots, f_{T_{i_M}}(Y)\right)^T\right)$$

which says that choosing the traps to minimize the average uncertainty is equivalent to choosing them to maximize the average mutual information over the channel of figure 5.1. At this point, we will not argue about the merits of using an average mutual information criterion to rank features (refer to Kanal [30], or Lewis [40] for a thorough discussion). To compute $H\left(C \mid f_{T_{i_1}}(Y), \cdots, f_{T_{i_M}}(Y)\right)$, one essentially needs to compute the posteriors

$$\pi_M(c) \triangleq Pr\left(C = c \mid f_{T_{i_1}}(Y) = b_{i_1}, \cdots, f_{T_{i_M}}(Y) = b_{i_M}\right)$$

for each $c \in \mathcal{C}$ and each binary M-tuple $(b_{i_1}, \cdots, b_{i_M})^T \in \{0,1\}^M$. Using Bayes' rule, it suffices to compute

$$Pr\left(f_{T_{i_1}}(Y) = b_{i_1}, \cdots, f_{T_{i_M}}(Y) = b_{i_M} \mid C = c\right)$$

which can again be computed, as for the case of using the full trap set $\mathcal{T}$, using the inclusion-exclusion principle.

$$Pr\left(f_{T_{i_1}}(Y) = b_{i_1}, \cdots, f_{T_{i_M}}(Y) = b_{i_M} \mid C = c\right) =$$

$$\sum_{m=0}^{\sigma_b} (-1)^m \sum_{T' \in S_m^{\sigma_b}(T_{j_1}, \cdots, T_{j_{\sigma_b}})} Q_{Y|C=c}(T \cup T')$$

where,

$$\sigma_b \triangleq \sum_{k=1}^{M} b_{i_k} \quad (algebraic\ sum)$$

$$T = \bigcup_{k=1,\ b_{i_k}=0}^{M} T_{i_k}$$

$\left\{T_{j_1}, \cdots, T_{j_{\sigma_b}}\right\}$ is the collection of those traps which are hit, i.e., those with associated $b_{i_k} = 1$, and $S_m^{\sigma_b}(T_{j_1}, \cdots, T_{j_{\sigma_b}})$ is the collection of all possible unions of exactly m out of the σ_b traps $T_{j_1}, \cdots, T_{j_{\sigma_b}}$. Each such computation requires 2^{σ_b} operations. For fixed M and for a given choice of M traps, one needs $O(|\mathcal{C}| \times 2^M \times 2^M) = O(|\mathcal{C}| \times 2^{2M})$ operations to compute $H(C \mid f_{T_{i_1}}(Y), \cdots, f_{T_{i_M}}(Y))$. Hence, if $|\mathcal{T}| = N$, then in order to optimize the choice of M traps under an average conditional uncertainty (or, equivalently, average mutual information) criterion, we need

$$O\left(\binom{N}{M} \times |\mathcal{C}| \times 2^{2M}\right)$$

operations. To get an idea of the numbers involved, if we only use point traps on a 64 × 64 grid (N = 4,096), probe at only 10 points (M = 10), and have 5 classes ($|\mathcal{C}|$ = 5), then we need $O(10^{36})$ operations. Therefore, the design is computationally prohibitive. In addition, we need to tabulate and store $|\mathcal{C}| \times 2^M$ posteriors $\pi_M(c) = Pr(C = c \mid f_{T_{i_1}}(Y) = b_{i_1}, \cdots, f_{T_{i_M}}(Y) = b_{i_M})$ into a 2-dimensional look-up table, in order to implement the runtime decision rule.

This discussion settles any questions concerning why joint optimization of the M trap sets is never pursued in practice. In the following we continue with one-step optimal adaptive classification procedures, and Probabilistic Classification

Trees in particular.

5.4 Probabilistic Classification Trees

The basic idea behind Probabilistic Classification Trees goes as follows. Suppose we have a "class" r.v. C, taking values in some finite alphabet, $\mathcal{C}$. Suppose that C is not directly observable. Instead we have a collection of "features" $\{f_i\}_{i\in I}$, which convey imperfect information about the class r.v. C. The idea is to build a tree-structured classifier, which at any given node of the decision tree only looks at *one* feature and then branches to one of its descendent nodes, according to the outcome of the feature evaluation[1]. The particular feature which is evaluated at a given internal node depends on the features evaluated along the path to the given node, as well as the associated feature values. When a leaf is reached, a decision is made concerning the operative c. Thus, the runtime implementation consists of traversing a path all the way down the tree, until a leaf is reached. En route, a typically small number of features are evaluated. Hence, the runtime rule is particularly simple, fast, and with prudent storage and logic requirements. However the optimal design of such a tree is a very difficult combinatorial problem [6]. Effectively, the complexity of the overall problem has been shifted from runtime to the design stage, *with some subsequent loss of performance* (since a tree-structured classifier is a *constrained* classification

[1]The basic references for this section are [22, 7, 59, 72, 30], the recent survey paper [53], and the book [6].

procedure, and, therefore, it is in general suboptimal).

The major idea behind the information-theoretic design of probabilistic classification trees is to choose the feature to be evaluated at each internal node by means of maximizing the average information gained by such an evaluation [22, 7, 59, 72]. This is conveniently visualized using a suitably chosen channel, which is associated to each internal tree node. During the tree design process, one maintains the posterior pmf of the class r.v. conditioned on the feature evaluations along the path to a given node, for all tree nodes. A node is declared a leaf if the entropy of this posterior pmf drops below a threshold (other criteria exist, e.g., see [22]). If a node is declared to be a leaf, then the label assigned to it is the mode of its associated posterior class pmf, i.e., the c with the highest conditional probability. If a node is not a leaf, then it is called "active". At each stage of the tree growing process, one active node is chosen to be split. Usually this node is the one with the highest conditional entropy. Although this is not optimal (since one should really pick the node which *after being split* results in the maximum attainable gain in average information), it is convenient.

A frequently made assumption is that the available features are *conditionally independent* (i.e. conditioned on the event that the class r.v. takes on a given value, the features are independent). Although in many cases this is not entirely justified, it makes for much simpler design. In some applications it is possible to "whiten" the features, by means of an orthogonal transformation applied prior to classification.

Instead of prescribing a maximum allowable level of conditional node entropy ("impurity"), we can specify that no more than M features can be evaluated during runtime. This requirement is more natural when the classification system must meet "hard" real - time constraints. In this case, we constrain the depth of the tree, and the result is an adaptive suboptimal algorithm for stepwise minimization of $H(C \mid f_{T_{i_1}}(Y), \cdots, f_{T_{i_M}}(Y))$ (contrast with the non - adaptive optimal design described in the previous section. Note that, due to adaptivity, the adaptive stepwise optimal algorithm may result in smaller probability of missclassification than the optimal non - adaptive algorithm).

5.5 Adaptive Sequential Probing

Let us now combine the trapping system and probabilistic classification tree ideas, to reformulate the problem of classifying discrete and binary "random" shapes in a cluttered environment. For convenience, we reproduce all relevant definitions:

- $\mathcal{C}$ = finite collection of discrete and binary shapes (subsets of a bounded $B \subset \mathbf{Z}^2$).
- C = a $\mathcal{C}$-valued DRS, specified via its pmf $\pi_0(c),\ c \in \mathcal{C}$.
- N = a DRS on B, specified via its generating functional, $Q_N(K),\ K \in \Sigma(B)$.
- $Y = C \cup N$ (not directly observable).
- $\mathcal{T}$ = finite collection of "traps", or "probes" (can be arbitrary subsets of B), with the property $\cup_{T \in \mathcal{T}} T = B$.

- $f_T(Y) = 1$, if $T \cap Y \neq \emptyset$, and 0 otherwise ("hit-or-miss" binary query, or "feature").
- $\{f_T(Y),\ T \in \mathcal{T}\}$ = collection of available features.
- The conditional probabilities:

$$\pi_k(c \mid f_{T_1}(Y), \cdots, f_{T_k}(Y)) = Pr(C = c \mid f_{T_1}(Y), \cdots, f_{T_k}(Y))$$

Let us consider a "generic" tree node at level k (root is at level 0). The path to the node under consideration involves binary queries $f_{T_1}, \cdots, f_{T_k}$, and associated answers $b_1, \cdots, b_k$. The corresponding class posterior pmf is

$$\pi_k(c \mid f_{T_1}(Y) = b_1, \cdots, f_{T_k}(Y) = b_k)$$

Suppose that the entropy of this posterior pmf is above the preset threshold, and so the node is "active". Suppose that it is also chosen to be split. According to the adopted splitting rule, we should choose the next trap, T_{k+1}, in such a way so as to minimize the average uncertainty after the split, i.e., we seek to minimize:

$$H(C \mid f_{T_{k+1}}(Y), f_{T_1}(Y) = b_1, \cdots, f_{T_k}(Y) = b_k)$$

by means of a suitable choice of $T_{k+1} \in \mathcal{T}$. Equivalently, we need to pick $T_{k+1} \in \mathcal{T}$ to minimize:

$$Pr(f_{T_{k+1}}(Y) = 0 \mid f_{T_1}(Y) = b_1, \cdots f_{T_k}(Y) = b_k) \times$$

$$H(C \mid f_{T_{k+1}}(Y) = 0, f_{T_1}(Y) = b_1, \cdots, f_{T_k}(Y) = b_k) +$$

$$Pr(f_{T_{k+1}}(Y) = 1 \mid f_{T_1}(Y) = b_1, \cdots f_{T_k}(Y) = b_k) \times$$

$$H(C \mid f_{T_{k+1}}(Y) = 1, f_{T_1}(Y) = b_1, \cdots, f_{T_k}(Y) = b_k)$$

Equivalently, we must choose $T_{k+1} \in \mathcal{T}$ to maximize the average mutual information over the channel depicted in figure 5.2. For this channel, and for $b_{k+1} = 0, 1$, we have

$$Pr(f_{T_{k+1}}(Y) = b_{k+1} \mid f_{T_1}(Y) = b_1, \cdots, f_{T_k}(Y) = b_k, \tilde{C} = c_j) =$$

$$\frac{Pr(f_{T_1}(Y) = b_1, \cdots, f_{T_k}(Y) = b_k, f_{T_{k+1}}(Y) = b_{k+1} \mid \tilde{C} = c_j)}{Pr(f_{T_1}(Y) = b_1, \cdots, f_{T_k}(Y) = b_k \mid \tilde{C} = c_j)}$$

For each $c \in \mathcal{C}$ and for $b_{k+1} = 0, 1$ (the $b_1, \cdots, b_k$ are fixed for a given node), the nominator and the denominator can be computed via the inclusion - exclusion principle. Then the new posteriors

$$\pi_{k+1}(c \mid f_{T_1}(Y) = b_1, \cdots, f_{T_{k+1}}(Y) = b_{k+1})$$

can be computed directly from the channel, and stored for future use (i.e. splitting the descendants of the node under consideration). The average mutual information gained by using T_{k+1} can be computed as

$$I_{T_{k+1}}(\tilde{C}; Q) = H(\tilde{C}) - H(\tilde{C} \mid Q)$$

$H(\tilde{C})$ is simply the entropy of the conditional pmf $\pi_k(\cdot)$. $H(\tilde{C} \mid Q)$ can be computed directly from the channel. In general, unless the r.v. $f_{T_{k+1}}(Y)$ is conditionally independent of $f_{T_1}(Y), \cdots, f_{T_k}(Y)$, for each choice of $T_{k+1} \in \mathcal{T}$, the computation of $I_{T_{k+1}}(\tilde{C}; Q)$ takes $O(|\mathcal{C}| \times 2^{k+1})$ operations, because of the need to compute the forward channel transition probabilities by means of the

inclusion - exclusion principle. Thus, in general, one needs $O(|\mathcal{T}| \times |\mathcal{C}| \times 2^{k+1})$ operations to optimally split a node at level k. In applications, the depth of classification trees is typically in the order of several hundred levels; this is the reason why the potentially dangerous assumption of conditional independence of the "features" (queries) is consistently being made over a wide spectrum of applications. Observe that if $f_{T_{k+1}}(Y)$ *is chosen* to be conditionally independent of the queries along the path to the node under consideration, then

$$Pr(f_{T_{k+1}}(Y) = 1 \mid f_{T_1}(Y) = b_1, \cdots, f_{T_k}(Y) = b_k, \tilde{C} = c_j) =$$

$$Pr(f_{T_{k+1}}(Y) = 1 \mid \tilde{C} = c_j) = \begin{cases} 1 & , \; if \; T_{k+1} \cap c_j \neq \emptyset \\ 1 - Q_N(T_{k+1}) & , \; otherwise \end{cases}$$

where Q_N is the generating functional of the "noise" DRS, N. Then the channel associated to probe selection at a generic tree node at level k reduces to the one depicted in figure 5.3. In figure 5.3, the alphabet $\mathcal{C}$ has been artificially split into two groups: those $c's$ which miss T_{k+1}, i.e. $c \cap T_{k+1} = \emptyset$ $(c_1, \cdots, c_m)$, and those which hit it, i.e. $c \cap T_{k+1} \neq \emptyset$ $(c_{m+1}, \cdots, c_{|\mathcal{C}|})$. This particular channel is easy to set up (since $Q_N(K)$, $K \in \Sigma(B)$ is assumed to be given). The problem of course is that it is only valid when $f_{T_{k+1}}(Y)$ is conditionally independent of $f_{T_1}(Y), \cdots, f_{T_k}(Y)$. The question then is for how long can we keep on growing the tree using "good" (i.e. highly informative), pathwise conditionally independent features. Theoretical results [22] indicate that in order to reach the Bayes optimal missclassification rate one needs to ask, on the average, approximately

$O(H(C))$ well chosen (highly informative) independent 0-1 questions, where $H(C)$ is the entropy of $\pi_0(\cdot)$, i.e. the entropy of the prior class pmf. Therefore, if $H(C)$ is relatively small, this approach can work out well. However, if $H(C)$ is high, and one uses the simplified channel model, in conjunction with strongly correlated queries, one will end up not only choosing the wrong (non-optimal) query at each stage, but also, more importantly, feeding the wrong posterior probabilities to the descendant nodes. In other words, *errors will propagate.*

In view of these observations, it becomes clear that we should attempt to grow the tree using pathwise conditionally independent queries, for as long as this is possible. If at some node down the tree this is no longer possible, we should look for queries which are only mildly dependent *with the ones which were previously used along the path to the node under consideration.* Then, one strategy is to use the simplified channel to compute the merit of each candidate query (feature), and pick the "best" query according to this simplified merit measure. This is not necessarily the query which maximizes the average mutual information over the full - complexity channel; however, we expect it to be a reasonably good choice. After a query has been chosen, the true posteriors for this particular choice should be computed *using the full - complexity channel,* in order to avoid error accumulation. We illustrate these concepts by means of a concrete example in the section that follows.

5.6 Probing in Boolean clutter

In this section we make two key assumptions: that N is a DRBRS of constant intensity, $\lambda_s(z) = p = 1 - q$, $\forall z \in B$, and primary grains of size $R \leq \bar{R}$, and that $\mathcal{T}$ is given by

$$\mathcal{T} = \{S_i \oplus \{z\} \mid S_i \in \mathcal{S},\ S_i \oplus \{z\} \subseteq B\}$$

For example, $\mathcal{S}$ might be

$$\mathcal{S} = \{\{\bar{0}\}, P, 2P, 3P, \cdots, (l-1)P\}$$

where P is a small convex structuring element ("the unit probe"), or, more generally,

$$\mathcal{S} = \{\{\bar{0}\}, P_1, \cdots, (l_1 - 1)P_1, \cdots, P_m, \cdots (l_m - 1)P_m\}$$

For simplicity, we will restrict our attention to one structuring element; the generalization to multiple structuring elements is straightforward. Note that we need not assume that $\mathcal{T}$ is rich enough to make the range of the DRS $Y = C \cup N$ a subset of the set of $\mathcal{T}$-closed sets. However, if the range of Y is a subset of the set of $\mathcal{T}$-closed sets, and we use *all* available "features" (queries), $\{f_T(Y),\ T \in \mathcal{T}\}$, then we can reconstruct Y. Thus, we stand a better chance of approaching the Bayes missclassification rate. Finally, we assume that the shapes $c \in \mathcal{C}$ are "roughly centered" at the origin, as a result of some preprocessing (more on this later).

Assuming that the queries are pathwise conditionally independent, the average mutual information over the (simplified) channel of figure 5.3, for a generic node at level k, is given by

$$I_{T_{k+1}}(\tilde{C}; Q) = I_{T_{k+1}}(\tilde{C}; f_{T_{k+1}}(Y)) = H(f_{T_{k+1}}(Y)) - H(f_{T_{k+1}}(Y) \mid \tilde{C})$$

with

$$H(f_{T_{k+1}}(Y)) = H_b\left([Q_N(T_{k+1})] \sum_{c \in \mathcal{C}} \pi_k(c) 1(T_{k+1} \cap c = \emptyset)\right)$$

and

$$H(f_{T_{k+1}}(Y) \mid \tilde{C}) = \left(\sum_{c \in \mathcal{C}} \pi_k(c) 1(T_{k+1} \cap c = \emptyset)\right) H_b(1 - Q_N(T_{k+1}))$$

where $H_b(\cdot)$ is the binary entropy function

$$H_b(t) = -t\ log(t) - (1-t)\ log(1-t), \quad t \in [0,1]$$

Thus, we need to pick $T_{k+1} \in \mathcal{T}$, to maximize this expression. Each $T \in \mathcal{T}$ can be written as $rP \oplus \{z\}$, for some $r \in \{0, 1, \cdots, (l-1)\}$, and $z \in B$, such that $rP \oplus \{z\} \subseteq B$. Observe that

$$rP \oplus \{z\} \subseteq B \Longleftrightarrow z \in B \ominus rP^s$$

Therefore, we need to pick r^*, z^* as follows:

$$(r^*, z^*) = arg\ max_{r \in \{0,1,\cdots,(l-1)\},\ z \in B \ominus rP^s}$$

$$\left\{H_b\left([Q_N(rP \oplus \{z\})] \sum_{c \in \mathcal{C}} \pi_k(c) 1((rP \oplus \{z\}) \cap c = \emptyset)\right) - \left(\sum_{c \in \mathcal{C}} \pi_k(c) 1((rP \oplus \{z\}) \cap c = \emptyset)\right) H_b(1 - Q_N(rP \oplus \{z\}))\right\}$$

Let

$$\eta \triangleq 1 - Q_N(rP \oplus \{z\})$$

and

$$\zeta \triangleq \sum_{c \in \mathcal{C}} \pi_k(c) 1((rP \oplus \{z\}) \cap c = \emptyset)$$

and observe that $I_{T_{k+1}}(\tilde{C}; Q)$ can be written as a function of η, ζ as

$$I_{T_{k+1}}(\tilde{C}; Q) = H_b((1-\eta)\zeta) - H_b(\eta)\zeta$$

We have the following elementary results:

• Fix $\eta \in (0,1)$. As a function of ζ, $H_b((1-\eta)\zeta) - H_b(\eta)\zeta$ is *concave* in $(0,1)$, and it attains its maximum at

$$\zeta = \frac{1}{(1-\eta)\left(1 + 2^{\frac{H_b(\eta)}{1-\eta}}\right)}$$

• Fix $\zeta \in (0,1)$. As a function of η, $H_b((1-\eta)\zeta) - H_b(\eta)\zeta$ is *convex*, and *strictly decreasing* in $(0,1)$.

In order to gain information, we have to probe at locations where $\zeta \neq 0, 1$, since for $\zeta = 0, 1$ the mutual information is zero. In particular, let

$$L \triangleq \bigcap_{c \in \mathcal{C}} c\,, \qquad U \triangleq \bigcup_{c \in \mathcal{C}} c$$

Then, for a fixed r, it suffices to optimize z over

$$(U \oplus rP^s) \backslash (L \oplus rP^s)$$

For any z outside this set, $rP \oplus \{z\}$ will either hit *all* of the c's in $\mathcal{C}$, or *none* of the c's in $\mathcal{C}$. In the former case $\zeta = 0$, whereas in the latter case $\zeta = 1$.

In order to guarantee that queries are pathwise conditionally independent, we must further restrict the domain of z. Let $T_1, \cdots, T_k$ be the collection of traps used along the path to the node under consideration. Then, in order to assure that $f_{T_{k+1}}(Y) = f_{rP \oplus \{z\}}(Y)$ is conditionally independent of $f_{T_1}(Y), \cdots, f_{T_k}(Y)$, we must exclude all z such that $rP \oplus \{z\}$ can be hit by a grain of the DRBRS N which can also hit one or more of the traps $T_1, \cdots, T_k$, i.e. we must exclude all z in

$$(T_1 \cup \cdots \cup T_k) \oplus \bar{R}H^s \oplus \bar{R}H \oplus rP^s$$

where H is the unit size primary grain of the DRBRS N. Thus, for a fixed r, z should be optimized over

$$D(r) \triangleq ((U \oplus rP^s) \backslash (L \oplus rP^s)) \setminus \left((T_1 \cup \cdots \cup T_k) \oplus \bar{R}H^s \oplus \bar{R}H \oplus rP^s\right)$$

In practice, $|D(r)|$ will be significantly smaller than $|B|$, and it will decrease fast with k.

It has been assumed that the shapes $c \in \mathcal{C}$ are "roughly centered" at the origin. Let us be more specific. We assume that

$$\left((U \oplus \bar{R}P^s) \backslash (L \oplus \bar{R}P^s)\right) \oplus \bar{R}P \oplus \bar{R}P^s \subseteq B$$

This condition amounts to requiring that edge effects can be safely ignored. Along with the assumption of constant intensity of the DRBRS N, it implies that, for $z \in D(r)$, $Q_N(rP \oplus \{z\}) = Q_N(rP)$, i.e. it does not depend on z, the location of the probe. Note that the values $Q_N(rP)$, $r = 0, 1, \cdots, l-1$ can be

easily and accurately estimated from a training sample of the DRBRS N, by taking spatial averages. Thus, for each r, z can be easily optimized over $D(r)$.

As we have seen earlier on, the generating functional of any DRS is constrained to be decreasing. It follows that η is an increasing function of r. Hence, in terms of maximizing the mutual information, it generally helps to keep r small. Furthermore, it makes sense to keep the size of the traps as small as possible, because this slows down the rate of decrease of $|D(r)|$ as a function of k, and, in effect, allows us to go deeper down the tree using pathwise conditionally independent queries.

With all these in mind, we propose the following query selection algorithm:

ALGORITHM (query selection at generic node at level k):

• **INPUT:** $T = T_1 \cup \cdots \cup T_k$, the union of the traps associated to queries along the path to the given node. $\pi_k(\cdot)$, the node posterior pmf.

• Let r = -1;

• do { Let r = r + 1; Compute

$$D(r) \triangleq ((U \oplus rP^s) \backslash (L \oplus rP^s)) \setminus \left(T \oplus \bar{R}H^s \oplus \bar{R}H \oplus rP^s\right)$$

Let

$$z^*(r) = arg\ max_{z \in D(r)} I(r, z)$$

where

$$I(r,z) \triangleq \left\{ H_b \left([Q_N(rP)] \sum_{c \in \mathcal{C}} \pi_k(c) 1((rP \oplus \{z\}) \cap c = \emptyset) \right) - \right.$$

$$\left(\sum_{c\in\mathcal{C}} \pi_k(c)1((rP\oplus\{z\})\cap c=\emptyset)\right) H_b(1-Q_N(rP))\Bigg\}$$

}

while $((r \leq (l-1))$ *and* $(I(r,z^*(r))$ *is not sufficiently close to* $1))$;

• If $I(r,z^*(r))$ is sufficiently close to 1, then $f_{T_{k+1}}(Y) = f_{rP\oplus\{z^*(r)\}}(Y)$ is a good query; otherwise no acceptable conditionally independent query can be found. □

Note that the average information that can be gained by $0-1$ queries is bounded above by 1. This "uniform" upper bound allows us to stop as soon as we find a "good" query (i.e. one that results in an average gain in information which is "sufficiently" close to 1). If no acceptable conditionally independent query can be found, then it is necessary to switch to the full - complexity channel, and, for each r, optimize z over $(U\oplus rP^s)\backslash(L\oplus rP^s)$ instead of $D(r)$. The complexity of this optimization is exponential in k. This is the biggest single drawback of the method. It remains to be seen whether there exist efficient approximate optimization algorithms for query selection when the queries are not pathwise conditionally independent. This will be the subject of future research.

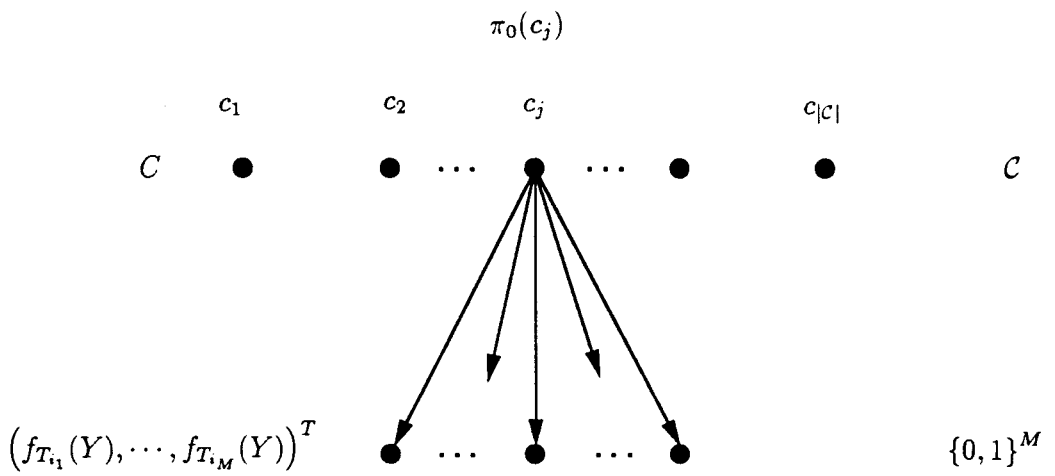


Figure 5.1: Channel associated to the "short horizon" problem.

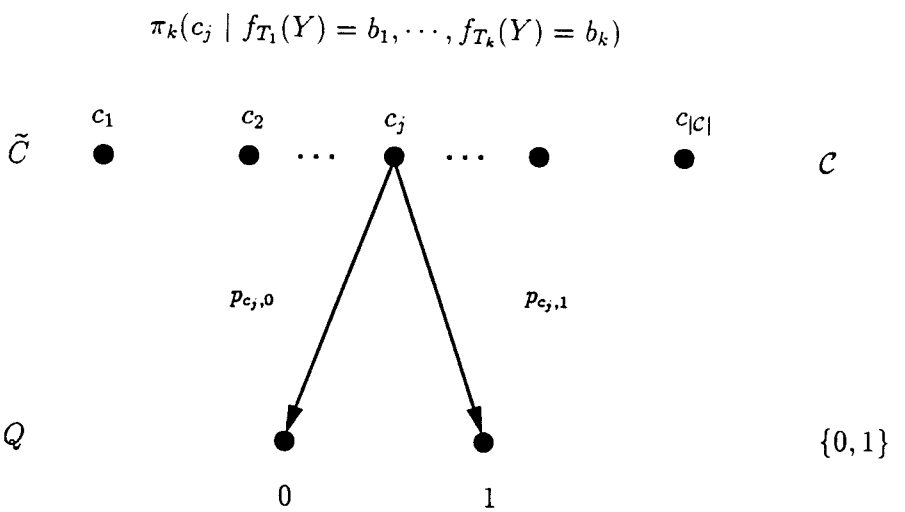


Figure 5.2: Channel associated to probe selection at a "generic" tree node at level k.

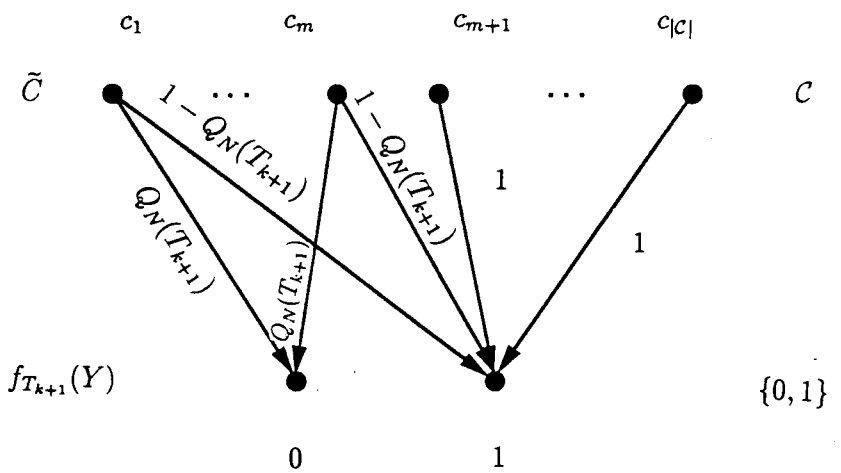


Figure 5.3: Simplified channel for pathwise conditionally independent queries.

CHAPTER
SIX

CONCLUSIONS AND FURTHER RESEARCH

In this dissertation, we have taken the approach of modeling discrete and binary digital images as realizations of a uniformly bounded discrete random set, a mathematical object which can be axiomatically defined directly on a finite sample space. We have argued for the merits of such an approach, most notably the ability to *recover* the associated probability measure by means of a Moebius-type transformation and knowledge of the generating functional. Based on this result, and some tools of Morphological shape analysis, we have developed a discrete analog of the Boolean random set, obtained its complete probabilistic specification, and provided various tools for its statistical inference. Although, in reality, binary digital image data are sampled versions of an underlying physical process, which lives in a continuum, the data *per se* can only assume a finite number of realizations. This is the case in many applications, in which there exist physical barriers that limit the available resolution. Although a uniformly bounded discrete random set approach may ignore the "fine letter" of the underlying physical structure, it provides a useful, and, most importantly, *tractable* idealization, which, as demonstrated, can lead to practical inference procedures.

Various important questions remain unresolved. With one exception (the so-called *hard-core* case of the last proposition of chapter 3), little is known about the consistency and/or efficiency of the proposed estimators. This is not a matter of coincidence; it reflects fundamental analytical difficulties, which, at this point, seem insurmountable.

The DRS modeling of highly correlated (e.g. man-made) "random" spatial structures is an important next step. Some progress towards this direction has been reported recently in [26]. It is based on the idea of using a 2-dimensional discrete random walk to generate "random skeletons" which can be tuned to resemble the skeletons of "real-world" shapes. Preliminary simulations indicate that this technique may hold promise. Still, at this point, DRS modeling of man-made structures is a far cry. One notable exception is the case of DRS structures which assume only a "small" number of realizations. These DRS structures can be specified and manipulated via their pmf, in much the same way as discrete random variables.

In the second part of this work, we have considered optimal filtering of binary digital images. Mask filtering is a natural approach to the problem of binary digital image restoration, under a union/intersection degradation model. We have discussed both optimal fixed-mask filtering, and optimal adaptive mask filtering. Although adaptive mask filtering is superior, it essentially requires knowledge of the capacity functionals of the signal and noise. This is the case when both the signal, X, and the noise, N, can be modeled as DBRS's. On the

other hand, fixed-mask filtering only requires knowledge of first-order statistics (pixel hitting probabilities), which can be easily and accurately estimated from training data. Therefore, it provides a simple and robust alternative, when the signal and noise processes are not known in detail.

The combined syntactical/statistical analysis of Morphological filters has been another focal point of our research. We have demonstrated that certain popular Morphological filtering schemes are indeed optimal under some fairly plausible assumptions. We have also described a general optimal Morphological binary image filtering approach, which is more appropriate when the signal and noise DRS's exhibit a statistical behavior which is spatially invariant. We have seen that by choosing the right expansion of the optimal filter, namely as a union of erosions (intersection of dilations), under an intersection (union) noise model, we can obtain universal optimal filtering results, which do not rely on strong assumptions concerning the nature of the signal and noise, and the mode of their spatial interaction. In particular, they are valid when the signal and noise patterns are spatially overlapping. This situation contrasts with the optimality results of Haralick, Dougherty, and Katz [27], which are based on the assumption that the signal and noise patterns are "non-interfering", and the results of Schonfeld and Goutsias [55], which rely on strong separability of the noise patterns. In contrast with the aforementioned model-based approaches, we have chosen to avoid restricting the class of input signals under consideration. Obviously, a model-based approach is superior when the underlying assumptions

are justified in practice. However, if this is not the case, then our approach may prove safer.

From the viewpoint of applications, it would be useful to develop software for the automatic generation of expressions for the probability of pixel error. We have seen that "manual" generation is rather straightforward, but cumbersome. As the size of the basis increases, the required manipulations become more and more tedious and boring. It should not be too difficult to develop a program for this purpose, possibly using a macro symbolic assembler.

The optimal design of increasing, shift-invariant filters, under a combined union/intersection noise model, is still open. This seems to be a difficult problem. The reason is that neither erosion, nor dilation are distributive over *both* union *and* intersection. Thus, the trick of using a convenient (under the given degradation model) expansion of the optimal filter does not work. As pointed out in chapter 4, another open problem is how to compute the "projection" (in the sense of the $d(\cdot,\cdot)$ distance metric) of an arbitrary subset of B, onto a collection of "smooth" subsets of B.

Finally, we have made some progress towards the development of adaptive probing procedures for the classification of "random" known shapes in clutter. Clearly, there remains much to be conquered along this route. At this point, it is not clear how to make optimal use of pathwise conditionally dependent geometric features, while simultaneously avoiding combinatorial explosion at the design stage. This is an important problem. We feel that extensive experimentation

is necessary in order to shed some light in this direction, and suggest "good" algorithms for probe selection.

We have reserved an important remark for the very end. A question that frequently arises is what is the most important contribution of this work. We believe that we have succeeded in demonstrating that a uniformly bounded discrete random set approach to statistical binary digital image modeling, filtering, and analysis, is meaningful and fruitful.

CHAPTER
SEVEN

APPENDIX

7.1 Maximum Likelihood parameter estimation

This appendix serves a double purpose. First, it is a self-contained summary of basic ML parameter estimation, as it pertains to our work. Second, it states and proves two lemmas, which we use in chapter 3. All this material is well known, and it can be found (albeit in scattered form) in many standard textbooks (e.g. [48]).

Let Y be a random variable taking values in $\mathcal{A}$, according to a pmf $f_\theta(y)$, $y \in \mathcal{A}$ (we assume that $|\mathcal{A}| < \infty$), where $\theta \in \Theta$ is some unknown (fixed) parameter. Let $\mathcal{F}(\mathcal{A})$ denote the power set of $\mathcal{A}$, let P_θ denote the probability measure induced on $\mathcal{F}(\mathcal{A})$ by f_θ, and let $\mathcal{F}(\Theta)$ denote a σ-algebra on Θ.

Definition 7.1.20 *An* **estimator**, $\widehat{\theta}(\cdot)$, *of* θ, *on the basis of* Y, *is any mapping of the probability space* $(\mathcal{A}, \mathcal{F}(\mathcal{A}), P_\theta)$ *into the measurable space* $(\Theta, \mathcal{F}(\Theta))$.

Note that measurability is automatically satisfied here, since $\mathcal{F}(\mathcal{A})$ is the power set of $\mathcal{A}$. An *estimator*, $\widehat{\theta}(\cdot)$, of θ, on the basis of Y, provides an *estimate*, $\widehat{\theta}(Y)$,

which is a random variable. In what follows E_θ denotes expectation with respect to Y, under the unknown fixed value of the parameter θ.

Definition 7.1.21 *An estimator, $\widehat{\theta}(\cdot)$, of θ, on the basis of Y, is* **unbiased** *if*

$$E_\theta \widehat{\theta}(Y) = \theta, \ \forall \theta \in \Theta$$

Theorem 7.1 *[48, Cramer-Rao lower bound] Define the Fisher information matrix*

$$F(\theta) \triangleq E_\theta \left[(\nabla_\theta ln f_\theta(Y)) (\nabla_\theta ln f_\theta(Y))^T \right]$$

If $F(\theta)$ is not singular, then the covariance

$$\Sigma(\widehat{\theta}(\cdot)/\theta) \triangleq E_\theta \left[\left(\widehat{\theta}(Y) - E_\theta \widehat{\theta}(Y) \right) \left(\widehat{\theta}(Y) - E_\theta \widehat{\theta}(Y) \right)^T \right]$$

of any unbiased estimator, $\widehat{\theta}(\cdot)$, of θ, on the basis of Y, will satisfy the following inequality

$$\Sigma(\widehat{\theta}(\cdot)/\theta) \geq F^{-1}(\theta), \ \forall \theta \in \Theta$$

Definition 7.1.22 *An unbiased estimator, $\widehat{\theta}(\cdot)$, of θ, on the basis of Y, is said to be* **efficient** *if it achieves the Cramer-Rao lower bound, i.e. if*

$$\Sigma(\widehat{\theta}(\cdot)/\theta) = F^{-1}(\theta), \ \forall \theta \in \Theta$$

in which case, $\widehat{\theta}(\cdot)$ is also called a **Minimum Variance Unbiased Estimator (MVUE)** *of θ, on the basis of Y.*

Definition 7.1.23 *Let* $\{Y_i\}_{i=1}^N$ *be a sequence of* N *independent r.v.'s, each taking values in* $\mathcal{A}$*, according to a pmf* $f_\theta(y)$*,* $y \in \mathcal{A}$*. The sequence of estimators,* $\left\{\widehat{\theta}^{(N)}(\cdot)\right\}_{N=1}^{\infty}$*, of* θ *on the basis of* $\left\{\{Y_i\}_{i=1}^N\right\}_{N=1}^{\infty}$*, respectively, is said to be* **strongly consistent** *if*

$$\widehat{\theta}^{(N)}(Y^{(N)}) \overset{a.s. \ as \ N\to\infty}{\longrightarrow} \theta$$

Here, a.s. means almost surely, *i.e. convergence for almost all sample paths to a unique limit, except for a set of sample paths of measure zero.*

Definition 7.1.24 *An estimator,* $\widehat{\theta}(\cdot)$*, of* θ*, on the basis of* Y*, is said to be a* **Maximum Likelihood Estimator (MLE)** *of* θ*, on the basis of* Y*, if it satisfies the ML-equation*

$$\nabla_\theta ln f_\theta(Y)|_{\theta=\widehat{\theta}(Y)} = \bar{0}$$

Lemma 7.1 *Let* $\{Y_i\}_{i=1}^N$ *be a sequence of* N *iid binary random variables, each with marginal* $Pr\{Y_i = 1\} = p$*,* $i = 1, \ldots, N$*. The sequence of empirical estimators,* $\left\{\widehat{p}^{(N)}(\cdot)\right\}_{N=1}^{\infty}$*, of* p *on the basis of* $\left\{\{Y_i\}_{i=1}^N\right\}_{N=1}^{\infty}$*, respectively, defined by*

$$\widehat{p}^{(N)}(Y^{(N)}) \triangleq \frac{1}{N}\sum_{i=1}^{N} Y_i, \ \forall N \in \mathcal{Z}_+^*$$

is a strongly consistent sequence of MVUE's. Furthermore, it is a sequence of MLE's of p *on the basis of* $\left\{\{Y_i\}_{i=1}^N\right\}_{N=1}^{\infty}$*.*

Proof:

The proof of the second part is trivial, and will be omitted. Since $\{Y_i\}_{i=1}^N$ is

an iid sequence of r.v.'s of finite mean and variance, by the strong law of large numbers

$$\frac{1}{N}\sum_{i=1}^{N} Y_i \overset{a.s.\ as\ N\to\infty}{\longrightarrow} E_p Y_1 = p$$

Hence, $\left\{\widehat{p}^{(N)}(\cdot)\right\}_{N=1}^{\infty}$ is a *strongly consistent* sequence of estimators of p on the basis of $\left\{\{Y_i\}_{i=1}^{N}\right\}_{N=1}^{\infty}$ respectively.

Next, observe that

$$E_p\widehat{p}^{(N)}(Y^{(N)}) = \frac{1}{N}\sum_{i=1}^{N} E_p Y_i = \frac{1}{N}\sum_{i=1}^{N} p = \frac{1}{N}pN = p,\ \forall N \in \mathcal{Z}_+^*$$

and, therefore, $\left\{\widehat{p}^{(N)}(\cdot)\right\}_{N=1}^{\infty}$ is a sequence of *unbiased* estimators of p on the basis of $\left\{\{Y_i\}_{i=1}^{N}\right\}_{N=1}^{\infty}$ respectively. Let us check whether it is efficient. Let $f_p(Y^{(N)})$ denote the joint pmf of the first N observations. The Cramer-Rao lower bound on the variance of unbiased estimators of p on the basis of $\{Y_i\}_{i=1}^{N}$, is expressed by the following inequality

$$E_p\left[\left(\widehat{p}^{(N)}(Y^{(N)}) - p\right)^2\right] \geq \frac{1}{F^{(N)}(p)},\ \forall p \in [0,1]$$

where the N-step Fisher information is defined as

$$F^{(N)}(p) \triangleq E_p\left[\left(\frac{\partial}{\partial p} ln f_p(Y^{(N)})\right)^2\right]$$

Under the iid assumption, it can be shown that

$$F^{(N)}(p) = NF^{(1)}(p)$$

Here, the one-step Fisher information is defined as

$$F^{(1)}(p) \triangleq E_p\left[\left(\frac{\partial}{\partial p} ln f_p(Y_1)\right)^2\right]$$

where $f_p(Y_1)$ is the pmf of Y_1 under p, given by

$$f_p(Y_1) = p^{Y_1}(1-p)^{1-Y_1}, \ \forall p \in [0,1]$$

The one-step Fisher information can be easily found to be

$$F^{(1)}(p) = \frac{1}{p(1-p)}$$

Thus

$$F^{(N)}(p) = \frac{N}{p(1-p)}$$

and the Cramer-Rao lower bound follows

$$\frac{1}{F^{(N)}(p)} = \frac{p(1-p)}{N}$$

Finally, it is straightforward to calculate the variance of the estimator (which is also the variance of the error here, since the estimator is unbiased) and verify that it is given by

$$E_p\left[\left(\widehat{p}^{(N)}(Y^{(N)}) - p\right)^2\right] = \frac{p(1-p)}{N}$$

Therefore, for all $N \in \mathcal{Z}_+^*$, the Cramer-Rao lower bound is met with equality. Hence $\left\{\widehat{p}^{(N)}(\cdot)\right\}_{N=1}^{\infty}$ is an *efficient* sequence of estimators of p on the basis of $\left\{\{Y_i\}_{i=1}^N\right\}_{N=1}^{\infty}$ respectively. Since it's also unbiased, it is a sequence of MVUE's of p on the basis of $\left\{\{Y_i\}_{i=1}^N\right\}_{N=1}^{\infty}$. □

The following lemma represents a generalization of the above result, for the case of finite alphabet iid observations, each taking values in $\{1,\ldots,M\}$, according to some unknown marginal pmf. Note that in attempting to uncover

the unknown probabilities, $\{p_j\}_{j=1}^M$, we are faced with a constrained parameter estimation problem, since $\sum_{j=1}^M p_j = 1$. Therefore, we only need to estimate $p_1, \ldots, p_{M-1}$, since $p_M = 1 - \sum_{j=1}^{M-1} p_j$.

Lemma 7.2 *Let $\{Y_i\}_{i=1}^N$ be a sequence of N iid random variables, each with marginal pmf $Pr\{Y_i = j\} = p_j$, $j = 1, \ldots, M$, $i = 1, \ldots, N$. Define*

$$\bar{p} \triangleq [p_1, \cdots, p_{M-1}]^T$$

and observe that the marginal pmf of the Y_i's can be written in terms of the elements of $\bar{p}$ as follows

$$f_{\bar{p}}(k) \triangleq Pr\{Y_i = k\} = \left(1 - \sum_{j=1}^{M-1} p_j\right)^{1_M(k)} \prod_{j=1}^{M-1} p_j^{1_j(k)}, \; k = 1, \ldots, M$$

The sequence of empirical frequency of occurrence estimators (or normalized histogram estimators) $\left\{\widehat{\bar{p}}^{(N)}(\cdot)\right\}_{N=1}^{\infty}$, of $\bar{p}$ on the basis of $\left\{\{Y_i\}_{i=1}^N\right\}_{N=1}^{\infty}$, respectively, defined by

$$\widehat{\bar{p}}^{(N)}(\cdot) \triangleq \left[\widehat{p}_1^{(N)}(\cdot), \cdots, \widehat{p}_{M-1}^{(N)}(\cdot)\right]^T$$

where, $\forall N \in \mathcal{Z}_+^*$,

$$\widehat{p}_j^{(N)}(Y^{(N)}) \triangleq \frac{1}{N} \sum_{i=1}^{N} 1_j(Y_i), \; j = 1, \ldots, M-1$$

is a strongly consistent sequence of MVUE's. Furthermore, it is a sequence of MLE's of $\bar{p}$ on the basis of $\left\{\{Y_i\}_{i=1}^N\right\}_{N=1}^{\infty}$.

Proof:

Since $\{Y_i\}_{i=1}^N$ is an iid sequence, it follows that, for each $j = 1, \ldots, M-1$,

$\{1_j(Y_i)\}_{i=1}^N$ is an iid sequence of binary valued r.v.'s of (obviously) finite mean and variance. Therefore, by the strong law of large numbers, for each $j = 1, \ldots, M-1$

$$\hat{p}_j^{(N)}(Y^{(N)}) \triangleq \frac{1}{N}\sum_{i=1}^{N} 1_j(Y_i) \overset{a.s.\ as\ N\to\infty}{\longrightarrow} E_{\bar{p}} 1_j(Y_1) = Pr_{\bar{p}}\{Y_1 = j\} = p_j$$

Hence, $\left\{\hat{\bar{p}}^{(N)}(\cdot)\right\}_{N=1}^{\infty}$ is a *strongly consistent* sequence of estimators of $\bar{p}$ on the basis of $\left\{\{Y_i\}_{i=1}^N\right\}_{N=1}^{\infty}$, respectively.

Next, observe that, for each $j = 1, \ldots, M-1$, and for all $N \in \mathcal{Z}_+^*$

$$E_{\bar{p}}\hat{p}_j^{(N)}(Y^{(N)}) = \frac{1}{N}\sum_{i=1}^{N} E_{\bar{p}} 1_j(Y_i) = \frac{1}{N}\sum_{i=1}^{N} p_j = p_j$$

Hence, $\left\{\hat{\bar{p}}^{(N)}(\cdot)\right\}_{N=1}^{\infty}$ is an *unbiased* sequence of estimators of $\bar{p}$ on the basis of $\left\{\{Y_i\}_{i=1}^N\right\}_{N=1}^{\infty}$, respectively.

Let us check for efficiency. Let $f_{\bar{p}}(Y^{(N)})$ denote the joint pmf of the first N observations under $\bar{p}$, given by

$$f_{\bar{p}}(Y^{(N)}) = \left(1 - \sum_{j=1}^{M-1} p_j\right)^{\sum_{i=1}^{N} 1_M(Y_i)} \prod_{j=1}^{M-1} p_j^{\sum_{i=1}^{N} 1_j(Y_i)}$$

The Cramer-Rao lower bound on the covariance matrix of unbiased estimators of $\bar{p}$ on the basis of $\{Y_i\}_{i=1}^N$, is expressed by the following inequality

$$\Sigma\left(\hat{\bar{p}}^{(N)}(\cdot)/\bar{p}\right) \geq \left(F^{(N)}(\bar{p})\right)^{-1}$$

where the $(M-1) \times (M-1)$ covariance matrix $\Sigma\left(\hat{\bar{p}}^{(N)}(\cdot)/\bar{p}\right)$, is defined as

$$\Sigma\left(\hat{\bar{p}}^{(N)}(\cdot)/\bar{p}\right) \triangleq E_{\bar{p}}\left[\left(\hat{\bar{p}}^{(N)}(Y^{(N)}) - \bar{p}\right)\left(\hat{\bar{p}}^{(N)}(Y^{(N)}) - \bar{p}\right)^T\right]$$

Note that this is also the error covariance matrix, since the estimator is unbiased.

The $(M-1)\times(M-1)$ N-step Fisher information matrix $F^{(N)}(\bar{p})$ is defined as

$$F^{(N)}(\bar{p}) \triangleq E_{\bar{p}}\left[\left(\nabla_{\bar{p}} ln f_{\bar{p}}(Y^{(N)})\right)\left(\nabla_{\bar{p}} ln f_{\bar{p}}(Y^{(N)})\right)^T\right]$$

It can be verified that the covariance matrix is given by

$$\left[\Sigma\left(\hat{\bar{p}}^{(N)}(\cdot)/\bar{p}\right)\right]_{k,l} = \begin{cases} -\frac{1}{N}p_k p_l, & 1 \le k,l \le M-1,\ l \ne k \\ \\ \frac{1}{N}p_k - \frac{1}{N}p_k^2, & 1 \le k = l \le M-1 \end{cases}$$

Again, under the iid assumption,

$$F^{(N)}(\bar{p}) = NF^{(1)}(\bar{p})$$

therefore, the Cramer-Rao Lower bound is

$$\Sigma\left(\hat{\bar{p}}^{(N)}(\cdot)/\bar{p}\right) \ge \frac{1}{N}\left(F^{(1)}(\bar{p})\right)^{-1}$$

where

$$F^{(1)}(\bar{p}) \triangleq E_{\bar{p}}\left[\left(\nabla_{\bar{p}} ln f_{\bar{p}}(Y_1)\right)\left(\nabla_{\bar{p}} ln f_{\bar{p}}(Y_1)\right)^T\right]$$

with

$$f_{\bar{p}}(Y_1) = \left(1 - \sum_{j=1}^{M-1} p_j\right)^{1_M(Y_1)} \prod_{j=1}^{M-1} p_j^{1_j(Y_1)}$$

and it can be verified that

$$\left[F^{(1)}(\bar{p})\right]_{k,l} = \begin{cases} \frac{1}{p_M}, & 1 \le k,l \le M-1,\ l \ne k \\ \\ \frac{1}{p_M} + \frac{1}{p_k}, & 1 \le k = l \le M-1 \end{cases}$$

Observe that $F^{(1)}(\bar{p})$ can be expanded as

$$F^{(1)}(\bar{p}) = (A + BCD)$$

with A being a $(M-1) \times (M-1)$ diagonal matrix, with k-th diagonal element equal to $1/p_k$, $k = 1, \ldots, M-1$, B being a $(M-1) \times 1$ column vector of units, C being the scalar $1/p_M$ and D being a $1 \times (M-1)$ row vector of units. Therefore, we can use the modified matrices formula of Woodbury to compute its inverse

$$(A + BCD)^{-1} = A^{-1} - A^{-1}B\left(DA^{-1}B + C^{-1}\right)^{-1} DA^{-1}$$

from which we obtain

$$\left[\left(F^{(1)}(\bar{p})\right)^{-1}\right]_{k,l} = \begin{cases} -p_k p_l, & 1 \leq k, l \leq M-1,\ l \neq k \\ \\ p_k - p_k^2, & 1 \leq k = l \leq M-1 \end{cases}$$

Therefore, the Cramer-Rao lower bound on the covariance matrix is

$$\left[\frac{1}{N}\left(F^{(1)}(\bar{p})\right)^{-1}\right]_{k,l} = \begin{cases} -\frac{1}{N}p_k p_l, & 1 \leq k, l \leq M-1,\ l \neq k \\ \\ \frac{1}{N}p_k - \frac{1}{N}p_k^2, & 1 \leq k = l \leq M-1 \end{cases}$$

which is exactly the covariance matrix of the proposed estimator, for all $N \in \mathcal{Z}_+^*$. Hence, $\left\{\hat{\bar{p}}^{(N)}(\cdot)\right\}_{N=1}^{\infty}$ is an *efficient* sequence of estimators of $\bar{p}$ on the basis of $\left\{\{Y_i\}_{i=1}^N\right\}_{N=1}^{\infty}$, respectively. Since it is also unbiased, it is a sequence of MVUE's of $\bar{p}$ on the basis of $\left\{\{Y_i\}_{i=1}^N\right\}_{N=1}^{\infty}$.

For the second part of the lemma, the ML-equation

$$\nabla_{\bar{p}} ln f_{\bar{p}}(Y^{(N)}) = \bar{0}$$

gives

$$\widehat{p}_{k,ML}^{(N)}(Y^{(N)}) = \frac{\left(1 - \sum_{j=1,j\neq k}^{M-1} \widehat{p}_{j,ML}^{(N)}(Y^{(N)})\right) \sum_{i=1}^{N} 1_k(Y_i)}{\sum_{i=1}^{N} 1_M(Y_i) + \sum_{i=1}^{N} 1_k(Y_i)}$$

for all $k = 1, \ldots, M-1$. It can be verified by direct substitution that

$$\widehat{p}_k^{(N)}(Y^{(N)}) \triangleq \frac{1}{N} \sum_{i=1}^{N} 1_k(Y_i), \; k = 1, \ldots, M-1$$

for all $N \in \mathcal{Z}_+^*$, is a valid solution to the ML equation above, and, therefore, $\left\{\widehat{\bar{p}}^{(N)}(\cdot)\right\}_{N=1}^{\infty}$ is a sequence of maximum likelihood estimators of $\bar{p}$ on the basis of $\left\{\{Y_i\}_{i=1}^{N}\right\}_{N=1}^{\infty}$. □

BIBLIOGRAPHY

[1] M. Aigner. *Combinatorial Theory.* Springer-Verlag, New York, 1979.

[2] P. Argentiero, R. Chin, and P. Beaudet. An Automated Approach to the Design of Decision Tree Classifiers. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 4(1):51–57, January 1982.

[3] Z. Artstein and R.A. Vitale. A Strong Law of Large Numbers for Random Compact Sets. *The Annals of Probability*, 3(5):879–882, 1975.

[4] G. Ayala, J. Ferrandiz, and F. Montes. Boolean models: ML estimation from circular clumps. *Biomedical Journal*, 32:73–78, 1990.

[5] M. Baudin. Note on the determination of cluster centers from a realization of a multidimensional Poisson cluster process. *Journal of Applied Probability*, 20:136–143, 1983.

[6] L. Breiman, J.H. Friedman, R.A. Olshen, and C.J. Stone. *Classification and Regression Trees.* Wadsworth, Belmont, CA, 1984.

[7] R.G. Casey and G. Nagy. Decision Tree Design Using a Probabilistic Model. *IEEE Trans. Information Theory*, 30(1):93–99, January 1984.

[8] F. Chen and P.A. Kelly. Algorithms for generating and segmenting Morphologically smooth binary images. In *Proc. 26th Annual Conference on Information Sciences and Systems*, Princeton University, Princeton, N.J., March 1992.

[9] G. Choquet. Theory of capacities. *Ann. Institute Fourier*, 5:131–295, 1953.

[10] N. Cressie. A strong limit theorem for random sets. *Supplement to Advances in Applied Probability*, 10:36–46, 1978.

[11] N. Cressie and G.M. Laslett. Random set theory and problems of modeling. *SIAM Review*, 29:557–574, 1987.

[12] Pamela J. Davy. Aspects of random set theory. *Supplement to Advances in Applied Probability*, 10:28–35, 1978.

[13] P.J. Diggle. Binary mosaics and the spatial pattern of heather. *Biometrics*, 37:531–539, 1981.

[14] E. Dougherty. Optimal Mean Square N-Observation Digital Morphological Filters - I. Optimal Binary Filters. *Computer Vision, Graphics, and Image Processing: Image Understanding*, 55(1):36–54, January 1992.

[15] E. Dougherty. Optimal Mean Square N-Observation Digital Morphological Filters - II. Optimal Gray - Scale Filters. *Computer Vision, Graphics, and Image Processing: Image Understanding*, 55(1):55–72, January 1992.

[16] E. Dougherty and C. Giardina. *Morphological Methods in Image and Signal Processing.* Prentice-Hall, Englewood Cliff, 1988.

[17] E.R. Dougherty, R.M. Haralick, Y. Chen, B. Li, C. Agerskov, U. Jacobi, and P. H. Sloth. Morphological pattern-spectra-based tau-opening optimization. In *Proc. SPIE Vol. 1606, Boston, Massachusetts.* Society for Optical Engineering, November 1991.

[18] E.R. Dougherty and R.P. Loce. Constrained optimal digital morphological filters. In *Proc. of the 25th Annual Conference on Information Sciences and Systems, The Johns Hopkins University, Baltimore, Maryland*, March 1991.

[19] E.R. Dougherty, A. Mathew, and V. Swarnakar. A conditional-expectation-based implementation of the optimal mean-square binary morphological filter. In *Proc. SPIE Vol. 1451, San Jose, California.* Society for Optical Engineering, February 1991.

[20] V. Dupac. Parameter estimation in the Poisson field of discs. *Biometrika*, 67:187–190, 1980.

[21] J. Serra Ed. *Image Analysis and Mathematical Morphology, vol. 2, Theoretical Advances.* Academic, San Diego, 1988.

[22] R.M. Goodman and P. Smyth. Decision Tree Design from a Communication Theory Standpoint. *IEEE Trans. Information Theory*, 34(5):979–994, September 1988.

[23] J. Goutsias. Modeling random shapes: An introduction to random set theory. *To appear in: Mathematical Morphology. Theory and Hardware, R. M. Haralick, Ed., Oxford University Press*, 1993.

[24] J. Goutsias and D. Schonfeld. Morphological representation of discrete and binary images. *IEEE Transactions on Signal Processing*, 39(6):1369–1379, June 1991.

[25] J. Goutsias and C. Wen. Modeling Discrete Random Shapes: A Random Set Theory Approach. Technical Report JHU/ECE 90-13, The Johns Hopkins University, 1990.

[26] J. Goutsias and C. Wen. Discrete random set models for shape synthesis and analysis. In *Proc. SPIE Vol. 1606 Visual Communications and Image Processing '91: Image Processing*, pages 174–185. Society for Optical Engineering, 1991.

[27] R.M. Haralick, E.R. Dougherty, and P.L. Katz. Model-based morphology. In *Proc. SPIE Vol. 1472, Orlando, Florida.* Society for Optical Engineering, April 1991.

[28] R.M. Haralick, X. Zhuang, C. Lin, and J.S.J. Lee. The Digital Morphological Sampling Theorem. *IEEE Trans. Acoustics, Speech and Signal Processing*, 37(12):2067–2089, 1989.

[29] E.B. Jensen and H. J. G. Gundersen. The stereological estimation of moments of particle volume. *Journal of Applied Probability*, 22:82–98, 1985.

[30] L.N. Kanal. Problem-solving methods and search strategies for pattern recognition. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 1:193–201, April 1979.

[31] A. F. Karr. *Point Processes and their Statistical Inference.* Marcel Dekker, New York and Basel, 1990.

[32] A.M. Kellerer. On the number of clumps resulting from the overlap of randomly placed figures in a plane. *Journal of Applied Probability*, 20:126–135, 1983.

[33] A.M. Kellerer. Counting figures in planar random configurations. *Journal of Applied Probability*, 22:68–81, 1985.

[34] D.G. Kendall. Foundations of a theory of random sets. In E.F. Harding and D.G. Kendall, editors, *Stochastic Geometry*, pages 322–376. John Wiley, London, England, 1974.

[35] D.G. Kendall. Shape manifolds, Procrustean metrics, and complex projective spaces. *The Bulletin of the London Mathematical Society*, 16:81–121, 1984.

[36] D.G. Kendall. Exact distributions for shapes of random triangles in convex sets. *Advances in Applied Probability*, 17:308–329, 1985.

[37] D.G. Kendall. A survey of the Statistical Theory of Shape. *Statistical Science*, 4(2):87–120, 1989.

[38] D.G. Kendall and Hui-Lin Le. The structure and explicit determination of convex-polygonally generated shape densities. *Advances in Applied Probability*, 19:896–916, 1987.

[39] T.Y. Kong and A. Rosenfeld. Digital topology: Introduction and survey. *Computer Vision, Graphics and Image Processing*, 48:357–393, 1989.

[40] P.M. Lewis. The characteristic selection problem in recognition systems. *IRE Trans. Information Theory*, IT-8:171–178, 1962.

[41] R.P. Loce and E.R. Dougherty. Using Structuring Element Libraries to Design Suboptimal Morphological Filters. In P.D. Gader and E.R. Dougherty, editors, *Proc. of Conference on Image Algebra and Morphological Image*

Processing II, SPIE vol. 1568, San Diego, CA, pages 233–246. Society for Optical Engineering, July 1991.

[42] P. Maragos. A Representation Theory for Morphological Image and Signal Processing. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 11(6):586–599, June 1989.

[43] P. Maragos and R.W. Schafer. Morphological skeleton representation and coding of binary images. *IEEE Trans. on Acoustics, Speech, and Signal Processing*, 34:1228–1244, 1986.

[44] G. Matheron. *Elements pour une theorie des Milieux Poreux.* Masson, 1967.

[45] G. Matheron. *Random Sets and Integral Geometry.* Wiley, New York, 1975.

[46] I. Pitas and N.D. Sidiropoulos. Pattern recognition of binary image objects using Morphological shape decomposition. *Computer Vision and Image Processing*, pages 279–305. L. Shapiro and A. Rosenfeld, Eds., Academic Press, 1992 (Collection of refereed articles from the journal: Computer Vision, Graphics, and Image Processing).

[47] G. Polya and G. Szego. *Problems and Theorems in Analysis, Vol. II.* Springer-Verlag, New York, 1976.

[48] H. V. Poor. *An introduction to Signal Detection and Estimation.* Springer-Verlag, New York, 1988.

[49] B.D. Ripley. Locally finite random sets: foundations for point process theory. *Ann. Probab.*, 4:983–994, 1976.

[50] B.D. Ripley. On stationarity and superposition of point processes. *Ann. Probab.*, 4:999–1005, 1976.

[51] B.D. Ripley. *Spatial Statistics.* John Wiley, New York City, New York, 1981.

[52] B.D. Ripley. *Statistical Inference for Spatial Processes.* Cambridge University Press, Cambridge, England, 1988.

[53] S.R. Safavian and D. Landgrebe. A Survey of Decision Tree Classifier Methodology. *IEEE Trans. Systems, Man, and Cybernetics*, 21(3):660–674, May/June 1991.

[54] Michel Schmitt. Estimation of the density in a stationary Boolean model. *Journal of Applied Probability*, 28:702–708, September 1991.

[55] D. Schonfeld and J. Goutsias. Optimal morphological pattern restoration from noisy binary images. *IEEE Transactions on Pattern Anal. Mach. Intell.*, 13(1):14–29, Jan. 1991.

[56] J. Serra. The Boolean Model and Random Sets. *Computer Graphics and Image Processing*, 12:99–126, 1980.

[57] J. Serra. *Image Analysis and Mathematical Morphology.* Academic, New York, 1982.

[58] J. Serra and L. Vincent. An Overview of Morphological Filtering. *Circuits, Systems and Signal Processing*, 11(1):47–108, January 1992.

[59] I.K. Sethi and G.P.R. Sarvarayudu. Hierarchical Classifier Design Using Mutual Information. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 4(4):441–445, July 1982.

[60] N.D. Sidiropoulos, J.S. Baras, and C.A. Berenstein. Discrete Random Sets: an Inverse Problem, plus tools for the Statistical Inference of the Discrete Boolean model. In P.D. Gader, E.R. Dougherty, and J. Serra, editors, *Proc. of Conference on Image Algebra and Morphological Image Processing III, SPIE vol. 1769, San Diego, CA.* Society for Optical Engineering, July 1992.

[61] N.D. Sidiropoulos, J.S. Baras, and C.A. Berenstein. Optimal Mask Filtering of Discrete Random Sets under a Union / Intersection Noise Model. In *Proc. 26th Annual Conference on Information Sciences and Systems*, Princeton University, Princeton, N.J., March 1992.

[62] N.D. Sidiropoulos, J.S. Baras, and C.A. Berenstein. Optimal Morphological Filters for Discrete Random Sets under a Union or Intersection Noise Model. In *Proc. of Conference on Visual Communications and Image Processing,*

SPIE vol. 1818, Boston, Mass. Society for Optical Engineering, November 1992.

[63] D. L. Snyder. *Random Point Processes.* Wiley, New York, 1975.

[64] J. Song and E.J. Delp. A Study of the Generalized Morphological Filter. *Circuits, Systems and Signal Processing*, 11(1):229–252, January 1992.

[65] D. Stoyan, W.S. Kendall, and J. Mecke. *Stochastic Geometry and its Applications.* Wiley, Berlin, 1987.

[66] R.A. Vitale. Random Convex Hulls: Floating Bodies and Expectations. *Journal of Approximation Theory.* To Appear.

[67] R.A. Vitale. Some developments in the theory of random sets. *Bulletin of the International Statistical Institute*, 50:863–871, 1983.

[68] R.A. Vitale. Symmetric Statistics and Random Shape. In *First World Congress of the Bernoulli Society, Tashkent, USSR*, September 1986. Invited talk.

[69] R.A. Vitale. Expected Convex Hulls, Order Statistics, and Banach Space Probabilities. *Acta Applicandae Mathematicae*, 9:97–102, 1987.

[70] R.A. Vitale. An alternative formulation of mean value for random geometric figures. *Journal of Microscopy*, 151(3):197–204, September 1988.

[71] R.A. Vitale. The Translative Expectation of a Random Set. *Journal of Mathematical Analysis and Applications*, 160(2):556–562, September 1991.

[72] Q.R. Wang and C.Y. Suen. Analysis and Design of a Decision Tree Based on Entropy Reduction and Its Application to Large Character Set Recognition. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 6(4):406–417, July 1984.

[73] J. Woods. Two-Dimensional discrete Markovian fields. *IEEE Transactions on Information Theory*, 18:232–240, 1972.