PROBLEMS OF COMPUTATIONAL AND INFORMATION COMPLEXITY IN MACHINE VISION AND LEARNING

by

SANJEEV RAMESH KULKARNI

B.S.,B.S.,M.S., Clarkson University (1983,1984,1985)
M.S., Stanford University (1985)

SUBMITTED TO THE DEPARTMENT OF
ELECTRICAL ENGINEERING AND COMPUTER SCIENCE
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY
IN ELECTRICAL ENGINEERING AND COMPUTER SCIENCE

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

May 1991

Signature of Author_____
Department of Electrical Engineering and Computer Science
May 20, 1991

Certified by_____
Sanjoy K. Mitter
Professor of Electrical Engineering
Thesis Supervisor

Accepted by_____
Arthur C. Smith
Chairman, Department Committee on Graduate Students

Problems of Computational and Information Complexity in Machine Vision and Learning

by

Sanjeev Ramesh Kulkarni

Submitted to the Department of Electrical Engineering and Computer Science
on May 20, 1991, in partial fulfillment of the requirements for
the degree of Doctor of Philosophy

Abstract

In this thesis, we consider a number of problems in the areas of machine vision and learning. Our results take steps towards understanding the computational and information complexity of problems in areas such as machine vision and signal processing. In the first part of the thesis, we study problems concerning computational requirements/limitations in machine vision. We first consider relationships between variational methods and discrete Markov random field formulations for the problem of image restoration and segmentation. Several discrete formulations are presented which correctly approximate the continuous segmentation problem. The results for the segmentation problem lead us to consider a question concerning the computation of the length of a digitized contour. It is shown that for a particular model of parallel computation, length cannot be computed locally with a rectangular digitization, but can be computed locally using a random tesselation and an appropriate deterministic one. Finally, we study the complexity of model based recognition and show that certain formulations of model based recognition are NP-complete.

In the second part of the thesis, we study a number of extensions to models in machine learning with a view towards obtaining information complexity results applicable to areas such as machine vision and signal processing. We first consider extensions to the Probably Approximately Correct (PAC) learning model, including learning over a class of distributions, active learning, and learning with generalized samples. We study a particular application of learning with generalized samples to a problem of reconstructing a curve by counting intersections with straight lines. Our results refine a classical result from stochastic geometry. Finally, we consider a problem concerning the classification of an unknown probability measure from empirical data. Using large deviations techniques, we simplify and extend previous results on classifying the mean of a random variable. We also study the much more general case of classifying the measure itself, and consider applications to density estimation and the problem of order determination of a Markov chain.

Thesis Supervisor: Sanjoy K. Mitter
Title: Professor of Electrical Engineering

Acknowledgments

First, I would like to thank my thesis advisor, Sanjoy Mitter, for his guidance, support, and enthusiasm not only for my thesis, but for all aspects of my professional development as well. My thesis committee, John Tsitsiklis, Bob Gallager, and Ron Rivest, provided many valuable comments suggestions.

Much of this thesis was a result of joint work with various other people. Chapter 2 was joint work with Sanjoy Mitter, Chapter 3 with Sanjoy, John Tsitsiklis, and Tom Richardson, Chapter 4 with Haim Schweitzer, Chapter 6 with Sanjoy and John, Chapter 7 with Sanjoy, John, and Ofer Zeitouni, Chapter 8 with Ofer and Jack Koplowitz, and Chapter 9 with Ofer. I would like to thank them for their permission to include the joint work in this thesis. I have thoroughly enjoyed and greatly benefited from working with all of them.

I have also benefited from interaction with many other students, faculty, and staff at M.I.T. Avi Lele has been a great friend, an ideal person to work with, and has contributed much to this thesis. Tom Richardson was a great office mate, a source of inspiration, and has also been a great friend. I have deeply appreciated and learned much through working with Alan Willsky. Technical discussions early on with John Wyatt, Luigi Ambrosio, Jayant Shah, and Charles Rockland, and more recently with Bob Gallager and David Tse, have been very valuable. Jeff Shapiro has provided much guidance and advice as my academic advisor. Peter Doerschuk and, more recently, Clem Karl have been helpful with a number of computer related matters. Kathleen O'Sullivan, Betty Lou McClanahan, Sheila Hegarty, and the rest of the staff at LIDS have always kept things running smoothly. All those involved in intramural sports, reading room discussions, and social activities helped maintain a nice balance at LIDS.

There are several people at Lincoln Laboratory who deserve special mention. In particular, I would like to thank Bill Keicher, Herb Kleiman, Chuck Niessen, Leo Sullivan, and Brian Edwards for their support for my thesis work, and Fred Knight, Alan Stein, Ellen Breau, and Wayne Schollenberger for various technical assistance.

Before coming to M.I.T., Jack Koplowitz, Eytan Barouch, and Mark Ablowitz at Clarkson and Tom Cover at Stanford were particularly influential in shaping my academic outlook. I am deeply indebted to Jack and Eytan for taking special interest in my educational development far beyond any call of duty.

Throughout the years, my parents and siblings have always been there for me as a source of motivation, support, and happiness. Many friends over the years and my recent extended family in the past few years have been there for me as well. Finally, this thesis would not have been possible without the love, support, and inspiration provided by my wife, Molly, and our children, Mykel and Kristina.

My work on this thesis was supported by the Office of Naval Research under Air Force Contract F19628-90-C-0002 (Lincoln Laboratory), and in part by the U.S. Army Research Office under Contract DAAL03-86-K-0171 (Center for Intelligent Control Systems). I greatly appreciated this support.

Contents

Chapter 1

Introduction

In this thesis, we consider a variety of problems in the areas of machine vision and machine learning. The underlying goal motivating our choice of problems is to take steps towards understanding the computational and information complexity of new classes of problems – particularly those arising in vision. There has been a tremendous amount of work done in the area of machine vision (for example, see the annual reviews by Rosenfeld [103, 104]. However, comparatively little work has been done on understanding the computational or information requirements/limitations of vision tasks.

The first part of the thesis (Chapters 2 through 4) deals with questions related to computational limitations. Chapters 2 through 3 concern variational and Markov random field (MRF) approaches to problems in early vision which have been a subject of great interest recently. We focus on a particular problem of image restoration and segmentation, but our methodology and results are suggestive of the types of questions that should be considered for other early vision problems using variational or MRF formulations. In Chapter 4, we consider the complexity of model based recognition, an approach which has been widely considered for later stages of vision. Again, our results are suggestive of the types of questions that should be considered for other approaches to the problem of object recognition.

The second part of the thesis (Chapters 5 through 9) deals with various extensions to models in machine learning. Standard approaches in estimation theory, signal processing, information theory, etc., are not applicable to most vision tasks due to the complex and highly nonlinear nature of the observations and objects to be reconstructed/estimated. Thus, there is little theoretical machinery available to address questions concerning the amount of data required to adequately perform a

particular vision task. There has been a great deal of work done in the area of machine learning on what is known as the Probably Approximately Correct (PAC) learning model. In Chapters 5 through 7, we study some extensions of the PAC model with a view towards extending the domain of applicability of this model. In Chapters 8 and 9, we study a different type of inference problem (or "learning model") from empirical data. We feel that this approach may be further extended and related to other learning approaches. Our results on both learning models are of interest in themselves, but it is hoped that they also serve as step towards developing machinery to answer questions of information complexity for a much broader range of problems, such as those arising in vision.

In the remainder of this chapter, we first briefly discuss the general areas considered in the thesis. Section 1.1 is related to the material in Chapters 2 through 3, Section 1.2 to the material in Chapter 4, Section 1.3 to the material in Chapters 5 through 7, and Section 1.4 to the material in Chapters 8 and 9. In Section 1.5, we describe the organization of the thesis and (very briefly) the main contributions in each of the chapters.

1.1 Variational and Markov Random Field Methods in Machine Vision

Recently, many problems in early vision have been formulated using variational approaches (e.g., see [82]). These variational methods are appealing for a number of reasons. The terms in the cost function are intuitively plausible and correspond in a natural way to constraints generally expected to be present in the environment (for example certain invariants and/or smoothness). Also, these methods provide a unifying approach to the wide variety of early vision tasks, and in fact suggest ways in which various early vision modules might be fused. Finally, it is possible to analyze certain problems (such as for segmentation) to obtain properties of the solutions provided by these variational approaches.

In practice, these variational methods are generally implemented using finite-difference-like approximations with discrete versions of the cost function defined on digitized domains. Interestingly, the resulting discrete problems are closely related to Markov random field (MRF) models, which are conceptually and computationally appealing. In fact, for many problems the MRF formulation is the starting point from

which a variational principle on a continuous domain can be deduced if necessary.

A continuous formulation is useful for a number of reasons. For example, it may be easier to impose or identify certain constraints such as invariance under arbitrary rotations and translations. Also, analytical techniques can be more readily applied to derive properties of the continuous formulation. However, since analytic solutions are not generally available, the problem must eventually be digitized to obtain numerical solutions. The discrete problem has the advantages of being more directly amenable to computer implementations, particularly with parallel algorithms or hardware. Also, as mentioned above, the probabilistic interpretation in terms of MRF's is conceptually appealing.

A natural question is whether these discrete formulations are in fact approximations of the continuous formulations in the sense that solutions to the discrete problems are close to solutions of the continuous problem as the lattice spacing tends to zero. This question is important if one wants to guarantee that the advantages of the continuous formulations are retained, at least approximately, by solving the discrete problem. Hence, for these reasons, our criteria for considering a discrete formulation to be an "approximation" to a continuous problem is not whether the cost functions are approximations of one another in the usual sense, but, rather, whether the solutions provided by the two problems (i.e., the minimizers of the cost functions) are approximations of one another.

We consider these questions of suitable discrete approximations for a particular formulation for image restoration and segmentation. Various discrete formulations in terms of MRF's were studied by Geman and Geman [46], Marroquin [81] and others. A variational approach to the problem was proposed by Mumford and Shah in [88] (see also Blake and Zisserman [22, 23]). It appears that the standard discrete formulations used do not properly approximate the continuous problem. We study some discrete formulations for which we prove desirable convergence properties in the continuum limit. Motivated by our results for the segmentation problem, we then study questions on whether the length of a curve can be computed locally from discrete approximations.

1.2 Model Based Recognition

Many tasks of perceptual information processing that are easy and natural for humans appear to be much harder for machines. For example, although locating an object

such as a pen on a table appears to us an easy task, it requires the ability to identify all possible shapes of pens as such, and is difficult to implement in a machine. These difficulties can be avoided in many computer vision applications that take place in a controlled environment. In these cases it is assumed that the objects of interest can be modeled and catalogued in a library. The problem of model based recognition can be informally described in the following way: given a library of modeled objects and a set of sensed data, identify and locate the objects from the library that are present in the data.

Reviews of the extensive literature on model based recognition in computer vision can be found in [19, 21, 26]; more recent studies include [48, 49, 74, 123]. The standard computational approach is to represent the modeled objects and the data in terms of discrete features so that the recognition can be solved as a search problem. These results indicate that by applying rigidity constraints in various ways, model based recognition can be efficiently applied to recognize a small number of object even from partial views and in the presence of non-malicious noise. The relevant complexity parameter in such cases is the number of features that comprise each object.

The generic model based recognition problem that we consider is noise free and assumes no occlusion. We analyze the case in which objects are represented by a small number of features. The relevant complexity parameter in this case is the number of objects. Instead of analyzing the performance of specific algorithms, our approach is to apply techniques from complexity theory to identify cases in which model based recognition appears to be inherently difficult. Specifically, we show that the problem is NP-complete, and thus, its complexity (modulo standard complexity assumptions, i.e., P $\neq$ NP) is super-polynomial in the size of the library.

Proving that a problem is NP-complete is a common technique in complexity analysis for identifying the problem as intrinsically difficult. In a (well defined) sense, an NP-complete problem is the most difficult problem in the class NP, which includes many difficult problems such as the traveling salesman. However, an NP-complete problem is not completely unapproachable; a standard method for coping with such problems is to identify easily solved sub-problems. In the case of model based recognition this might correspond to exploiting additional structure of the modeled objects and the way they are viewed. The negative results that we provide can be used to determine constraints that may simplify the problem of model based recognition. For more information on the theory of NP-complete problems see [44]. For applications of NP-completeness results to vision tasks see [64, 117].

1.3 Formal Models for Machine Learning

In defining a formal model for learning, one needs to provide precise specifications for the concepts being learned, the information gathering mechanisms, and the performance criteria. A number of different models have been studied by varying one or more of the features above. For example, Gold [47] considered a model known as "identification in the limit" (referring to the performance criterion), which has been studied extensively in the context of language learning. Another interesting class of leaning models which has been considered is known as mistake bound (or on-line) learning [78, 79].

We will focus on a learning model popularized by the work of Valiant [118] which he referred to as "distribution-free learning" (referring to an aspect of the performance criteria). This model is also known as the "probably approximately correct" (or PAC) learning model (referring to a different aspect of the performance criteria). A more general model (ignoring computational complexity requirements) was studied independently by Vapnik [121], and fundamental results related to this framework have been obtained in the probability and statistics literature [119, 120, 36, 94]. A very general formulation for PAC learning was presented by Haussler [55, 56]. There has been a tremendous amount of work done on analyzing and extending the original model. Most of the variations retain the PAC criterion, and so we use the term "PAC learning" to refer to the original model and its variants.

In the PAC model, the learner attempts to approximate a function unknown to him, but chosen from a known class of functions. The data available to the learner consists of random samples of the unknown function. After seeing some bounded number of samples, the learner is required to produce a hypothesis which with high probability is close to the true function (hence the name "probably approximately correct"). Precise definitions for this learning model are provided in Chapter 5.

One goal of studying such a formal framework is to be able to characterize in a precise sense the tractability of learning problems. That is, to address questions concerning the amount of data required to learn a particular concept. We feel that this general paradigm may be extended to help provide results on the information complexity for a wide range of problems in areas such as machine vision, system identification, and signal/image processing. We consider some extensions of the PAC learning framework with a view towards extending the domain of applicability to such areas.

1.4 Classifying Probability Measures from Empirical Data

In [28], Cover presented some interesting results concerning the type of information that can be extracted about the mean of an unknown random variable from a sequence of i.i.d. samples. Koplowitz [66] extended and refined some of Cover's results. The problem considered in [28, 66] can be stated as follows. Given a sequence of i.i.d. samples of an unknown random variable, we wish to decide whether the mean of the random variable is in a particular set or its complement. It is required that we eventually stop making incorrect decisions as long as the true mean is not in some set of measure zero.

This problem is interesting since the type of information to be extracted differs from the usual objective in estimation/statistics, and cannot be directly obtained from the usual convergence results on empirical means (e.g., Chebycheff bounds or laws of large numbers). The success criteria is also highly reminiscent of an "identification in the limit" criteria used in machine learning [47].

We extend some results of [28, 66], and study a much more general framework. We feel that the framework discussed may have connections with other learning paradigms, and it may be possible to further extend the framework to apply to a much broader class of problems.

1.5 Contributions and Organization of Thesis

The remainder of this thesis is organized as follows. In Chapter 2, we first describe the variational method for image segmentation, and discuss some useful results from geometric measure theory. We then derive a number of properties of Minkowski content. These are used to show that a new discrete formulation that we present appropriately approximates the continuous formulation. We then present two other discrete approximations which also correctly approximate the continuous segmentation problem. The proofs of convergence for these methods are much simpler, and they may also lead to more efficient implementations.

A problem suggested by these discrete approximations is discussed in Chapter 3. This problem concerns the local versus global nature of computing the length of a curve from discrete approximations. We define a particular model for parallel computation, and following definitions of Minsky and Papert [84] (see also [1]), we

consider notions of local and non-local computation for this model. We show that for the usual rectangular digitizations, length cannot be computed locally, but that using appropriate random or deterministic digitizations, the length of line segments can be computed locally.

In Chapter 4, we consider the computational complexity for the problem of model matching in object recognition. We show that certain formulations of model matching are NP-complete, so that without further restrictions this approach to object recognition is computationally difficult.

Starting in Chapter 5, we turn our attention to some problems in machine learning. After introducing the PAC learning model, we derive some new results on the relationships between the metric entropy of a concept class with respect to various distributions and its VC dimension. We then prove some partial results regarding learnability for a class of distributions which give some indication of whether prior knowledge of the distribution helps in terms of learnability. Our results suggest that a substantial amount of prior knowledge regarding the distribution is required before this prior information impacts learnability.

In Chapter 6, we study the question of how much oracles can help the learnability of a concept class. Specifically, we consider the effect of allowing access to an oracle capable of answering arbitrary binary valued queries. We show that, surprisingly, for both fixed distribution and distribution-free cases, the set of learnable concept classes is not enlarged by allowing active learning, although the sample complexity can be reduced.

In Chapter 7, we consider an extension of the PAC learning model which allows the use of more general types of samples. This extension substantially increases the range of problems to which the PAC learning framework can be applied. We consider a specific application to a problem of reconstructing a curve by counting intersections with straight lines. Our results refine a classical result from stochastic geometry.

In Chapters 8 and 9, we consider a problem involving classifying probability measures from empirical data. Our results greatly extend previous results on problems of this type. The model we consider is a kind of generalization to an identification in the limit criterion, and may be applicable to more standard learning formulations.

At the end of Chapters 2 through 9 we discuss a number of open problems and directions to pursue which are directly related to the content of the particular chapter. In Chapter 10, we discuss a number of potential research directions of a more general nature. Some of the directions we discuss are relatively independent of the work

presented in the thesis, but concern ideas which arose during the course of this work. Other directions concern extensions of or connections between various parts of the thesis.

Chapter 2

Convergent Discrete Approximations to a Variational Method for Image Segmentation

2.1 Variational and Markov Random Field Methods for Image Segmentation

A variational approach to the problem of restoring and segmenting an image degraded by noise was recently proposed by Mumford and Shah in [88] (see also Blake and Zisserman [22, 23]). The method involves minimizing a cost functional over a space of boundaries with suitably smooth functions within the boundaries. Specifically, if g represents the observed image defined on $\Omega \subset \mathbf{R}^2$, then a reconstructed image f and its associated edges Γ are found by minimizing

$$E(f,\Gamma) = c_1 \iint_{\Omega} (f-g)^2 \, dx \, dy + c_2 \iint_{\Omega \setminus \Gamma} \| \nabla f \|^2 \, dx \, dy + c_3 L(\Gamma) \qquad (2.1)$$

where c_1, c_2, c_3 are constants and $L(\Gamma)$ denotes the length of Γ. An interesting special case of this problem is obtained if f is restricted to be constant within connected components of $\Omega \setminus \Gamma$. In this case, the optimal value of f on a connected component of $\Omega \setminus \Gamma$ is simply the mean of g over the connected component. Hence, the solution

depends only on Γ and is obtained by minimizing

$$E(\Gamma) = c_1 \sum_{i=1}^{k} \iint_{\Omega_i} (g - \overline{g}_i)^2 \, dx \, dy + c_3 L(\Gamma) \tag{2.2}$$

where $\Omega_1, \ldots, \Omega_k$ are the connected components of $\Omega \backslash \Gamma$, and $\overline{g}_i$ is the mean of g over Ω_i.

Discrete versions of these problems have also been proposed [23, 88]. In these discrete problems, the original image g is defined on a subset of the lattice $\frac{1}{n}\mathbf{Z}^2$ with lattice spacing $\frac{1}{n}$. The reconstructed image f is defined on the same lattice, while the boundary Γ consists of a subset of line segments joining neighboring points of the dual lattice. For the discrete problem, f and Γ are found by minimizing

$$E(f, \Gamma) = c_1 \sum_{i \in \Omega} \frac{1}{n^2} (f_i - g_i)^2 + c_2 \sum_{\substack{i,i' \in \Omega \\ \text{adjacent} \\ \overline{ii'} \cap \Gamma = \emptyset}} (f_i - f_{i'})^2 + c_3 L(\Gamma) \tag{2.3}$$

Similar discrete problems arise in the context of using Markov random fields for problems in vision as proposed by Geman and Geman [46] and studied by Marroquin [81] and others.

The continuous formulation has some distinct advantages over the discrete formulation. For example, the continuous problem is invariant under arbitrary rotations and translations. Also, results from the calculus of variations can be applied in the continuous case. In fact, such methods have yielded interesting results concerning the properties of the minimizing f and Γ [89, 108, 122]. However, since analytic solutions are not available, the problem must eventually be digitized to obtain numerical solutions. The discrete problem has the advantages of being more directly amenable to computer implementations, particularly with parallel algorithms or hardware.

A desirable property of any discrete version of a continuous problem would be for solutions of the discrete problem to converge to solutions of the continuous problem in the continuum limit. In the examples above, one would like convergence of the discrete solutions as the lattice spacing tends to zero. It seems that this is not the case for the problems as defined above. Specifically, in Section 2.4 we present an example for which there is evidence indicating that solutions to the discrete problem fail to converge to a solution of the continuous problem. In this chapter, we consider formulations involving modifications to both the cost functional and the discretization procedure for which we can ensure convergence in the continuum limit. We consider

three discrete approximations and prove some desirable convergence results for these methods. For the cost functional, we propose the use of different penalty terms for the boundaries instead of Hausdorff measure which has been previously used [4, 5, 96]. For the discretization procedure, we consider only digitizing the boundary. The observed and reconstructed images are still defined on continuous domains.

In Section 2.2, we introduce some preliminary definitions and results from geometric measure theory, and in Section 2.3 some additional properties of Minkowski content are derived. Section 2.4 contains results on the application of these ideas to the variational formulation of the segmentation problem.

2.2 Metrics and Measures on the Space of Boundaries

In this section, we introduce a variety of notions useful in dealing with the 'boundaries' or 'edges' of an image. The 'image' is usually a real valued function defined on a bounded open set $\Omega \subset \mathbf{R}^2$, although some of the results consider the more general case of $\Omega \subset \mathbf{R}^n$. A *boundary* generally refers to a closed subset of $\overline{\Omega}$. However, sometimes the boundary may be restricted to have certain additional properties such as having a finite number of connected components. A topology on the space of boundaries is required for the notion of convergence, and a measure of the 'cost' of a boundary is required for the variational problem.

For $A \subset \mathbf{R}^n$, the *δ-neighborhood* of A will be denoted by $A^{(\delta)}$ and is defined as

$$A^{(\delta)} = \{x \in \mathbf{R}^n : \inf_{y \in A} |x - y| < \delta\}$$

The notion of distance between boundaries which we will use is the Hausdorff metric $d_H(\cdot, \cdot)$ defined as

$$d_H(A_1, A_2) = \inf\{\rho : A_1 \subset A_2^{(\rho)} \text{ and } A_2 \subset A_1^{(\rho)}\}$$

It is elementary to show that $d_H(\cdot, \cdot)$ is in fact a metric on the space of all non-empty compact subsets of $\mathbf{R}^n$. An important property of this metric is that it induces a topology which makes the space of boundaries compact.

Theorem 2.1 *Let $\mathcal{C}$ be an infinite collection of non-empty closed subsets of a bounded closed set $\overline{\Omega}$. Then there exists a sequence $\{\Gamma_n\}$ of distinct sets of $\mathcal{C}$ and a non-empty*

closed set $\Gamma \subset \overline{\Omega}$ *such that* $\Gamma_n \to \Gamma$ *in the Hausdorff metric.*

Proof: See [41], Theorem 3.16.

□

For the 'cost' of a boundary, the usual notion of length cannot be applied to highly irregular boundaries. Hence a measure on the space of boundaries which generalizes the usual notion of length is desired. A variety of such measures for subsets of $\mathbf{R}^n$ have been investigated. (e.g., see [42]). Perhaps the most widely used and studied are Hausdorff measures [41, 42, 102].

For a non-empty subset A of $\mathbf{R}^n$, the *diameter* of A is defined by $|A| = \sup\{|x-y| : x, y \in A\}$. Let

$$\omega_s = \frac{\Gamma(\frac{1}{2})^s}{\Gamma(\frac{s}{2}+1)}$$

where $\Gamma(\cdot)$ is the usual Gamma function. For integer values of s, ω_s is the volume of the unit ball in $\mathbf{R}^s$. For $s > 0$ and $\delta > 0$ define

$$\mathcal{H}^s_\delta(A) = 2^{-s}\omega_s \ \inf\{\sum_{i=1}^{\infty} |U_i|^s \ : A \subset \bigcup_{i=1}^{\infty} U_i, \ |U_i| \leq \delta\}$$

The *Hausdorff s-dimensional measure* of A is then given by

$$\mathcal{H}^s(A) = \lim_{\delta \to 0} \mathcal{H}^s_\delta(A) = \sup_{\delta > 0} \mathcal{H}^s_\delta(A)$$

Note that the factor $2^{-s}\omega_s$ in the definition of $H^s_\delta(\cdot)$ is included for proper normalization. With this definition, for integer values of s, Hausdorff measure gives the desired value on sets where the usual notions of length, area, and volume apply.

Many properties of Hausdorff measure can be found in [41, 42, 102]. The following definitions are required to state several useful properties. A *curve* $\Gamma \subset \mathbf{R}^n$ is the image of a continuous injection $\psi : [0,1] \to \mathbf{R}^n$. The *length* of a curve Γ is defined as

$$L(\Gamma) = \sup\{\sum_{i=1}^{m} |\psi(t_i) - \psi(t_{i-1})| \ : \ 0 = t_0 < t_1 < \cdots < t_m = 1\}$$

and Γ is said to be *rectifiable* if $L(\Gamma) < \infty$. Finally, a compact connected set is called a *continuum.*

Theorem 2.2 *If* $\Gamma \subset \mathbf{R}^n$ *is a curve, then* $\mathcal{H}^1(\Gamma) = L(\Gamma)$.

Proof: See [41] Lemma 3.2.

□

Theorem 2.3 *If Γ is a continuum with $\mathcal{H}^1(\Gamma) < \infty$, then Γ consists of a countable union of rectifiable curves together with a set of $\mathcal{H}^1$-measure zero.*

Proof: See [41], Theorem 3.14.

□

Theorem 2.4 *If $\{\Gamma_n\}$ is a sequence of continua in $\mathbf{R}^n$ that converges (in Hausdorff metric) to a compact set Γ, then Γ is a continuum and $\mathcal{H}^1(\Gamma) \leq \liminf_{n\to\infty} \mathcal{H}^1(\Gamma_n)$.*

Proof: See [41], Theorem 3.18.

□

Theorem 2.4 asserts that $\mathcal{H}^1$-measure is lower-semicontinuous on the set of connected boundaries with respect to the Hausdorff metric. Richardson [96] extended this result to a cost term for boundaries which depends on the number of connected components. Specifically, define $\nu(\Gamma) = \mathcal{H}^1(\Gamma) + F(\#(\Gamma))$ where $\#(\Gamma)$ denotes the number of connected components of Γ, and F is any non-decreasing function such that $\lim_{n\to\infty} F(n) = \infty$. The following result was shown in [96].

Theorem 2.5 *$\#(\cdot)$ and $\nu(\cdot)$ are lower-semicontinuous on the space of boundaries with respect to the Hausdorff metric.*

Proof: See [96], Lemma 1 and Theorem 2.

□

This result was used in [96] to prove an existence theorem for the variational problems of interest. The essential properties required are the compactness of the space of boundaries and the lower-semicontinuity of the cost functional. However, with the discrete approximation suggested in [88], $\mathcal{H}^1$-measure can be strictly lower-semicontinuous on the space of boundaries with the topology induced by the Hausdorff metric. That is, one can find a set of boundaries Γ_n in the discrete approximations converging to a boundary Γ but with $\mathcal{H}^1(\Gamma) < \liminf_{n\to\infty} \mathcal{H}^1(\Gamma_n)$. It is for this reason that discrete solutions may fail to converge in the continuum limit to a solution of

the continuous problem (see Section 2.4). It may be possible to resolve this problem by modifying the cost functional and/or the discretization process. Here we consider the use of alternate notions for the cost of boundaries and modified discretization procedures (discussed in Section 2.4).

To measure the cost of the boundaries, we suggest the use of Minkowski content [42]. Let $\mu(\cdot)$ denote Lebesgue measure in $\mathbf{R}^n$. For any $A \subset \mathbf{R}^n$, $0 \leq s \leq n$, and $\delta > 0$, define

$$\mathcal{M}_\delta^s(A) = \frac{\mu(A^{(\delta)})}{\delta^{n-s}\omega_{n-s}}$$

As in the definition of Hausdorff measure, the term ω_{n-s} is included for proper normalization. Recall that $A^{(\delta)}$ is the δ-neighborhood of A — i.e. those points within distance δ of A. Equivalently, $A^{(\delta)}$ is the Minkowski set sum of A and the open ball of radius δ; or in the terminology of mathematical morphology [107] it is the dilation of A with the open ball of radius δ. In general, $\lim_{\delta \to 0} \mathcal{M}_\delta^s(A)$ may not exist (for an example see [42], Section 3.2.40). However, *lower* and *upper Minkowski contents* can be defined by

$$\mathcal{M}_*^s(A) = \liminf_{\delta \to 0^+} \mathcal{M}_\delta^s(A)$$

and

$$\mathcal{M}^{*s}(A) = \limsup_{\delta \to 0^+} \mathcal{M}_\delta^s(A)$$

respectively. If these two values agree (i.e if $\lim_{\delta \to 0} \mathcal{M}_\delta^s(A)$ exists) then the common value is simply called the *s-dimensional Minkowski content* and is denoted by $\mathcal{M}^s(A)$.

2.3 Properties of Minkowski Content

In this section, we develop several properties of Minkowski content some of which will be used in Section 2.4. The results can roughly be categorized as properties of δ-neighborhoods, continuity and regularity properties of Minkowski content, and relationships between Minkowski content and Hausdorff measure.

First, we state two elementary properties. Two sets A_1, A_2 are said to be *positively separated* if

$$d(A_1, A_2) \equiv \inf\{|a_1 - a_2| : a_1 \in A_1, a_2 \in A_2\} > 0$$

The sets $A_1, A_2, \ldots, A_m$ are called positively separated if $\min_{i \neq j} d(A_i, A_j) > 0$. The first property is that $\mathcal{M}^s$ is additive on positively separated sets, i.e. if $A_1, A_2, \ldots, A_m$ are positively separated then $\mathcal{M}^s(\cup_{i=1}^m A_i) = \sum_{i=1}^m \mathcal{M}^s(A_i)$. This follows from the fact

that for sufficiently small δ, the δ-neighborhoods of the A_i are disjoint. The second property is that for any set A, $A^{(\delta)} = \overline{A}^{(\delta)}$ and so $\mathcal{M}_\delta^s(A) = \mathcal{M}_\delta^s(\overline{A})$ for every $\delta > 0$, where $\overline{A}$ denotes the closure of A. Clearly $A^{(\delta)} \subset \overline{A}^{(\delta)}$. On the other hand, if $x \in \overline{A}^{(\delta)}$ then $|x - y| = \eta < \delta$ for some $y \in \overline{A}$. But $|y - a| < \delta - \eta$ for some $a \in A$, so that $|x - a| \leq |x - y| + |y - a| < \delta$. Hence, $x \in A^{(\delta)}$ and so the result follows.

The following two lemmas give properties of δ-neighborhoods which will be useful in showing continuity properties of Minkowski content. $B_r(x)$ and $\overline{B}_r(x)$ denote the open and closed balls, respectively, of radius r centered at x, and for a set A, ∂A denotes the topological boundary of A, i.e., the closure of A minus the interior of A.

Lemma 2.1 *$\mu(\partial\,\Gamma^{(\delta)}) = 0$ for every $\Gamma \subset \mathbf{R}^2$.*

Proof: Let $\Gamma \subset \mathbf{R}^2$ and let $E = \partial\,\Gamma^{(\delta)}$. The Lebesgue density of E at x, $D_\mu(E, x)$, is defined as

$$D_\mu(E, x) = \lim_{r \to 0} \frac{\mu(E \cap B_r(x))}{\mu(B_r(x))}$$

when the limit exists. We will show that the Lebesgue density of E is less than 1 for all $x \in E$. Hence, $\mu(E) = 0$ will follow from the Lebesgue Density Theorem.

Let $x \in E = \partial\,\Gamma^{(\delta)}$. Then for each $r > 0$, there exists $c(r) \in \Gamma$ with $|x - c(r)| < \delta + r^2$. If $w \in B_\delta(c(r))$ then $w \notin E$, so that

$$\mu(E \cap B_r(x)) \leq \mu(B_r(x)) - \mu(B_r(x) \cap B_\delta(c(r)))$$

The circle of radius δ centered at $c(r)$ intersects the circle of radius r centered at x in two points which determine a chord C. Let S denote the segment of $B_r(x)$ determined by C, θ the central angle at x subtended by C, and a the distance from x to C. Then

$$\mu(B_r(x) \cap B_\delta(c(r))) \geq \mu(S) = \frac{1}{2} r^2(\theta - \sin\theta)$$

and

$$\lim_{r \to 0} \theta = \lim_{r \to 0} 2\cos^{-1}(\frac{d}{r}) = \lim_{r \to 0} 2\cos^{-1}(\frac{r^2 + 2\delta r^2 + r^4}{2r(\delta + r^2)}) = \pi$$

Therefore,

$$D_\mu(E, x) = \lim_{r \to 0} \frac{\mu(E \cap B_r(x))}{\mu(B_r(x))} \leq \lim_{r \to 0} \frac{\mu(B_r(x)) - \mu(S)}{\mu(B_r(x))} = \lim_{r \to 0}(1 - \frac{1}{2\pi}(\theta - \sin\theta)) = \frac{1}{2}$$

□

Lemma 2.2 *If $\Gamma_n \to \Gamma$ in Hausdorff metric, then $\Gamma_n^{(\delta)} \to \Gamma^{(\delta)}$.*

Proof: Let $\epsilon > 0$. Since $\Gamma_n \to \Gamma$, $\exists\ N < \infty$ such that $d_H(\Gamma_n, \Gamma) < \epsilon\ \forall n > N$. If $x \in \Gamma^{(\delta)}$ then $x = a + \rho$ with $a \in \Gamma$ and $\| \rho \| < \delta$. For all $n > N$, there exists $a_n \in \Gamma_n$ with $\| a - a_n \| < \epsilon$. Then $x_n \equiv a_n + \rho \in \Gamma_n^{(\delta)}$ and $\| x - x_n \| = \| a - a_n \| < \epsilon$. Hence, $\Gamma^{(\delta)} \subset (\Gamma_n^{(\delta)})^{(\epsilon)}$. Similarly, $\Gamma_n^{(\delta)} \subset (\Gamma^{(\delta)})^{(\epsilon)}$. Thus, $d_H(\Gamma_n^{(\delta)}, \Gamma^{(\delta)}) < \epsilon\ \forall n > N$.

□

Two continuity properties of $\mathcal{M}_\delta^s$ may now be deduced. These follow directly from the corresponding continuity properties of Lebesgue measure on δ-neighborhoods.

Theorem 2.6 *If $\Gamma_n \to \Gamma$ in Hausdorff metric then $\mu(\Gamma_n^{(\delta)}) \to \mu(\Gamma^{(\delta)})$ and so $\mathcal{M}_\delta^s(\Gamma_n) \to \mathcal{M}_\delta^s(\Gamma)$. I.e., $\mathcal{M}_\delta^s(\Gamma)$ is continuous in Γ with respect to Hausdorff metric.*

Proof: Since $\Gamma_n \to \Gamma$, by Lemma 2.2 we have $\Gamma_n^{(\delta)} \to \Gamma^{(\delta)}$. Let $\epsilon > 0$. Then there exists $N < \infty$ such that $\Gamma_n^{(\delta)} \subset (\Gamma^{(\delta)})^{(\epsilon)}\ \forall n \geq N$. Therefore, $\sup_{n \geq N} \mu(\Gamma_n^{(\delta)}) \leq \mu(\Gamma^{(\delta+\epsilon)})$. As $\epsilon \downarrow 0$, $\Gamma^{(\delta+\epsilon)} \downarrow \overline{\Gamma^{(\delta)}}$ so that $\limsup_{n\to\infty} \mu(\Gamma_n^{(\delta)}) \leq \mu(\overline{\Gamma^{(\delta)}})$. Then by Lemma 2.1 it follows that $\limsup_{n\to\infty} \mu(\Gamma_n^{(\delta)}) \leq \mu(\Gamma^{(\delta)})$.

Let K be a compact subset of $\Gamma^{(\delta)}$. Since $\{B_\delta(x) : x \in \Gamma\}$ is an open cover of K, there exists a finite subcover $B_\delta(x_1), \ldots, B_\delta(x_m)$. Let $\epsilon > 0$. Since $\Gamma_n \to \Gamma$, there exists $N < \infty$ such that $\forall n \geq N$ we can find $y_{n,1}, \ldots, y_{n,m} \in \Gamma_n$ with $|y_{n,i} - x_i| < \epsilon$ for $i = 1, \ldots, m$. Then $\mu(B_\delta(x_i) \setminus B_\delta(y_{n,i})) < f(\epsilon)$ where $f(\epsilon) = \mu(B_1 \setminus B_2) \leq 2\delta\epsilon$ where B_1 and B_2 are balls of radius δ whose centers are ϵ apart. Therefore,

$$\mu(\Gamma_n^{(\delta)}) \geq \mu(\bigcup_{i=1}^{m} B_\delta(y_{n,i})) > \mu(K) - mf(\epsilon) \quad \forall n \geq N$$

and so $\inf_{n \geq N} \mu(\Gamma_n^{(\delta)}) > \mu(K) - mf(\epsilon)$. Since $\epsilon > 0$ is arbitrary and $f(\epsilon) \to 0$ as $\epsilon \to 0$ we have $\liminf_{n\to\infty} \mu(\Gamma_n^{(\delta)}) \geq \mu(K)$. Finally, since this is true for every compact $K \subset \Gamma^{(\delta)}$, we have $\liminf_{n\to\infty} \mu(\Gamma_n^{(\delta)}) \geq \sup_{K \subset \Gamma^{(\delta)}} \mu(K) = \mu(\Gamma^{(\delta)})$.

Thus,

$$\liminf_{n\to\infty} \mu(\Gamma_n^{(\delta)}) = \limsup_{n\to\infty} \mu(\Gamma_n^{(\delta)}) = \lim_{n\to\infty} \mu(\Gamma_n^{(\delta)}) = \mu(\Gamma^{(\delta)})$$

□

Proposition 2.1 *$\mathcal{M}_\delta^s(\Gamma)$ is continuous in δ for all $\delta > 0$.*

Proof: As $\eta \uparrow \delta$, we have $\Gamma^{(\eta)} \uparrow \Gamma^{(\delta)}$ so that $\mu(\Gamma^{(\eta)}) \uparrow \mu(\Gamma^{(\delta)})$. As $\eta \downarrow \delta$, we have $\Gamma^{(\eta)} \downarrow \overline{\Gamma^{(\delta)}}$. Then by Lemma 2.1, $\mu(\Gamma^{(\eta)}) \downarrow \mu(\overline{\Gamma^{(\delta)}}) = \mu(\Gamma^{(\delta)})$. Thus, $\lim_{\eta\to\delta} \mu(\Gamma^{(\eta)}) = \mu(\Gamma^{(\delta)})$.

□

All the results given so far in this section were proved for $\Gamma \subset \mathbf{R}^2$. However, these results and proofs can easily be extended to $\mathbf{R}^n$.

We now state a result given in Federer [42] relating Minkowski content to Hausdorff measure. A subset Γ of $\mathbf{R}^n$ is called *m-rectifiable* if there exists a Lipschitzian function mapping a bounded subset of $\mathbf{R}^m$ onto Γ.

Theorem 2.7 *If Γ is a closed m-rectifiable subset of $\mathbf{R}^n$ then $\mathcal{M}^m(\Gamma) = \mathcal{H}^m(\Gamma)$.*

Proof: See [42] Theorem 3.2.39.

□

We will present a proof of Theorem 2.7 in the restricted case of 1-dimensional measure in $\mathbf{R}^2$ (i.e., $m = 1, n = 2$), which is stated as Theorem 2.8. The basic idea of our proof is contained in the proof of Proposition 2.4. This idea will be used again in the proof of Theorem 2.9 on the Γ-convergence of Minkowski content, which is true only for 1-dimensional measures.

The following two results give upper and lower bounds on $M^1_\delta(\Gamma)$ for rectifiable and connected sets respectively. These two results could be appropriately extended to s-dimensional measure in $\mathbf{R}^n$.

Proposition 2.2 *If $\Gamma \subset \mathbf{R}^2$ is rectifiable then $\mu(\Gamma^{(\delta)}) \leq 2\delta\mathcal{H}^1(\Gamma) + \pi\delta^2$ and so $\mathcal{M}^1_\delta(\Gamma) \leq \mathcal{H}^1(\Gamma) + \frac{1}{2}\pi\delta$.*

Proof: Since Γ is rectifiable, $\Gamma = \{\gamma(t) : 0 \leq t \leq 1\}$ where $\gamma : [0,1] \to \mathbf{R}^2$ is rectifiable and $\mathcal{H}^1(\Gamma) = \sup\{\sum_{i=1}^m |\gamma(t_i) - \gamma(t_{i-1})| : 0 = t_0 < t_1 < \cdots < t_m = 1\}$. For $j = 1, 2, \ldots$ let $\{t_{ij}\}$ be a sequence of dissections such that $\max_i\{|t_{ij} - t_{i-1,j}|\} \to 0$ and $\mathcal{H}^1(\Gamma) = \lim_{j\to\infty} \sum_{i=1}^{m(j)} |\gamma(t_{ij}) - \gamma(t_{i-1,j})|$ Let $C_j = \cup_{i=1}^{m(j)} S_i$ where S_i is the straight line joining $\gamma(t_{i-1,j})$ and $\gamma(t_{ij})$. Then $\mu(S_{ij}^{(\delta)}) = |\gamma(t_{ij}) - \gamma(t_{i-1,j})| + \pi\delta^2$, and

$$\begin{aligned}
\mu(\cup_{i=1}^k S_{ij}^{(\delta)}) &= \mu(\cup_{i=1}^{k-1} S_{ij}^{(\delta)}) + \mu(S_{kj}^{(\delta)}) - \mu(S_{kj}^{\delta} \cap \bigcup_{i=1}^{k-1} S_{ij}^{(\delta)}) \\
&\leq \mu(\cup_{i=1}^{k-1} S_{ij}^{(\delta)}) + \mu(S_{kj}^{(\delta)}) - \pi\delta^2 \\
&= \mu(\cup_{i=1}^{k-1} S_{ij}^{(\delta)}) + |\gamma(t_{kj}) - \gamma(t_{k-1,j})|
\end{aligned}$$

By induction on i, we get

$$\mu(C_j^\delta) \leq \sum_{i=1}^{m(j)} |\gamma(t_{ij}) - \gamma(t_{i-1,j})| + \pi\delta^2$$

Since $C_j \to \Gamma$ in Hausdorff metric, by Theorem 2.6

$$\mu(\Gamma^{(\delta)}) = \lim_{j\to\infty} \mu(C_j^{(\delta)}) \leq \mathcal{H}^1(\Gamma) + \pi\delta^2$$

□

Proposition 2.3 *If $\Gamma \subset \mathbf{R}^2$ is connected, then $\mathcal{M}_\delta^1(\Gamma) \geq |\Gamma|$.*

Proof: Let $x, y \in \Gamma$, and let $\epsilon > 0$. Since Γ is connected, we can find $x = x_0, x_1, \ldots, x_k = y$ in Γ with $|x_i - x_{i-1}| < \epsilon$ for $1 \leq i \leq k$. Let $P(w)$ denote the point obtained by the orthogonal projection of w onto the straight line T through x and y, and let $p(w)$ be the coordinate of $P(w)$ considering T as the real line with origin at x and positive direction towards y. I.e.,

$$p(w) = \frac{\langle w - x, y - x\rangle}{|y - x|}$$

where $\langle \cdot, \cdot \rangle$ denotes the usual inner product. Note that $|p(x_i) - p(x_{i-1})| = |P(x_i) - P(x_{i-1})| \leq |x_i - x_{i-1}| < \epsilon$. By deleting intermediate points and reordering the indices as necessary, we can assume that $0 = p(x_0) < p(x_1) < \cdots < p(x_k) = |x - y|$ and $p(x_i) - p(x_{i-1}) < \epsilon$.

For $u, v \in \mathbf{R}^2$ with $p(u) < p(v)$, let $R(u, v) = \{w \in \mathbf{R}^2 : p(u) < p(w) < p(v)\}$. Then

$$\begin{aligned} \Gamma^{(\delta)} &\supset \cup_{i=0}^k B_\delta(x_i) \supset \cup_{i=1}^k B_\delta(x_i) \cap R(x_i, x_{i-1}) \\ &= \cup_{i=1}^k B_\delta(P(x_i)) \cap R(x_i, x_{i-1}) \end{aligned}$$

Since the $R(x_i, x_{i-1})$ for $i = 1, 2, \ldots, k$ are disjoint,

$$\begin{aligned} \mathcal{M}_\delta^1(\Gamma) &\geq \frac{\mu(\cup_{i=1}^k B_\delta(P(x_i)) \cap R(x_i, x_{i-1}))}{2\delta} \\ &= \frac{1}{2\delta} \sum_{i=1}^k \mu(B_\delta(P(x_i)) \cap R(x_i, x_{i-1})) \end{aligned}$$

$$\begin{aligned} &\geq \frac{1}{2\delta}\sum_{i=1}^{k} 2\sqrt{\delta^2-\epsilon^2}\,(p(x_i)-p(x_{i-1})) \\ &= |x-y|\sqrt{1-\frac{\epsilon^2}{\delta^2}} \end{aligned}$$

Since $\epsilon > 0$ is arbitrary we have $\mathcal{M}^1_\delta(\Gamma) \geq |x-y|$. Finally, the result follows since $x, y \in \Gamma$ are arbitrary.

□

Using the bounds of Propositions 2.2 and 2.3, the following proposition can be shown.

Proposition 2.4 *If $\Gamma \subset \mathbf{R}^2$ is connected and consists of a countable union of rectifiable curves then $\mathcal{M}^1(\Gamma) = \mathcal{H}^1(\Gamma)$.*

Proof: First, we prove the result when Γ is a rectifiable curve which does not intersect itself. Let $\Gamma = \{\gamma(t) : 0 \leq t \leq 1\}$ where $\gamma : [0,1] \to \mathbf{R}^2$ is rectifiable and $\gamma(s) \neq \gamma(t)$ if $s \neq t$. Let $0 = t_0 < t_1 < \cdots < t_m = 1$, and for $i = 1, 2, \ldots, m$ let $\Gamma_i = \{\gamma(t) : t_{i-1} < t < t_i\}$. If $K_i \subset \Gamma_i$ $i = 1, 2, \ldots, m$ are continua then they are positively separated. Therefore, for sufficiently small δ the $K_i^{(\delta)}$ are disjoint. From Proposition 2.3 we have

$$\mathcal{M}^1_\delta(\Gamma) \geq \sum_{i=1}^{m} \mathcal{M}^1_\delta(K_i) \geq \sum_{i=1}^{m} |K_i|$$

for all sufficiently small δ. Hence,

$$\liminf_{\delta \to 0} \mathcal{M}^1_\delta(\Gamma) \geq \sum_{i=1}^{m} |\gamma(t_i) - \gamma(t_{i-1})|$$

and since the dissection $\{t_i\}$ is arbitrary

$$\liminf_{\delta \to 0} \mathcal{M}^1_\delta(\Gamma) \geq \mathcal{H}^1(\Gamma)$$

On the other hand, from Proposition 2.2, $\mathcal{M}^1_\delta(\Gamma) \leq \mathcal{H}^1(\Gamma) + \frac{1}{2}\pi\delta$ so that

$$\limsup_{\delta \to 0} \mathcal{M}^1_\delta(\Gamma) \leq \mathcal{H}^1(\Gamma)$$

Thus,

$$\mathcal{M}^1(\Gamma) = \lim_{\delta \to 0} \mathcal{M}^1_\delta(\Gamma) = \mathcal{H}^1(\Gamma)$$

Now, suppose $\Gamma = \cup_{i=1}^{\infty} C_i$ is connected where the C_i are rectifiable curves. By decomposing the C_i as necessary, we can assume that they are not self-intersecting and that C_i intersects C_j in at most a finite number of points for $i \neq j$. Then $\mathcal{H}^1(\Gamma) = \sum_{i=1}^{\infty} \mathcal{H}^1(C_i)$. Let $E_k = \cup_{i=1}^{k} C_i$. Then by a dissection argument similar to that used above we get

$$\liminf_{\delta \to 0} \mathcal{M}^1_\delta(E_k) \geq \sum_{i=1}^{k} \mathcal{H}^1(C_i)$$

and so

$$\liminf_{\delta \to 0} \mathcal{M}^1_\delta(\Gamma) \geq \sup_k \liminf_{\delta \to 0} \mathcal{M}^1_\delta(E_k) \geq \mathcal{H}^1(\Gamma)$$

Also, from Proposition 2.2 and the fact that Γ is connected we have

$$\begin{aligned} \mu(E_k^{(\delta)}) &= \mu(E_{k-1}^{(\delta)}) + \mu(C_k^{(\delta)}) - \mu(E_{k-1}^{(\delta)} \cap C_k^{(\delta)}) \\ &\leq \mu(E_{k-1}^{(\delta)}) + \mu(C_k^{(\delta)}) - \pi\delta^2 \leq \mu(E_{k-1}^{(\delta)}) + 2\delta\mathcal{H}^1(C_k) \end{aligned}$$

By induction we get

$$\mu(E_k^{(\delta)}) \leq 2\delta \sum_{i=1}^{k} \mathcal{H}^1(C_k) + \pi\delta^2$$

for every integer k. Since $E_k^{(\delta)}$ is an increasing sequence of sets with $\Gamma^{(\delta)} = \cup_{i=k}^{\infty} E_k$ we have

$$\mu(\Gamma^{(\delta)}) = \lim_{k \to \infty} \mu(E_k^{(\delta)}) \leq 2\delta\mathcal{H}^1(\Gamma) + \pi\delta^2$$

Thus

$$\limsup_{\delta \to 0} \mathcal{M}^1_\delta(\Gamma) \leq \mathcal{H}^1(\Gamma)$$

and so the result follows.

□

The next inequality gives bounds for s-dimensional Minkowski content in $\mathbf{R}^2$ which are valid for *every* subset of $\mathbf{R}^2$. This could also be appropriately extended to $\mathbf{R}^n$. Here, we use the notation

$$\mathcal{H}^s_{\delta,2\delta}(\Gamma) = 2^{-s}\omega_s \ \inf\{\sum_{i=1}^{\infty} |U_i|^s \ : \Gamma \subset \bigcup_{i=1}^{\infty} U_i, \ \delta \leq |U_i| \leq 2\delta\}$$

Proposition 2.5 *For every* $\Gamma \subset \mathbf{R}^2$ *and* $0 \leq s \leq 2$,

$$\frac{2^{s-1}}{\omega_s \omega_{2-s}} \mathcal{H}^s_\delta(\Gamma) \leq \mathcal{M}^s_\delta(\Gamma) \leq \frac{16}{\omega_s \omega_{2-s}} \mathcal{H}^s_{\delta,2\delta}(\Gamma)$$

and so

$$\frac{2^{s-1}}{\omega_s \omega_{2-s}} \mathcal{H}^s(\Gamma) \leq \mathcal{M}^s_*(\Gamma) \leq \liminf_{\delta \to 0} \frac{16}{\omega_s \omega_{2-s}} \mathcal{H}^s_{\delta,2\delta}(\Gamma)$$

where $\mathcal{H}^s_{\delta,2\delta}(\Gamma) = 2^{-s}\omega_s \ \inf\{\sum_{i=1}^\infty |U_i|^s \ : \Gamma \subset \bigcup_{i=1}^\infty U_i, \ \delta \leq |U_i| \leq 2\delta\}$

Proof: Consider the closed lattice squares formed by the points $\frac{1}{\sqrt{2}\delta}\mathbf{Z}^2$. Form a cover $\{U_i\}$ of Γ by taking all lattice squares whose intersection with Γ is non-empty. Then $\{U_i\}$ is a δ-cover of Γ and $\bigcup_i U_i \subset \overline{\Gamma^{(\delta)}}$. Hence,

$$\begin{aligned} \frac{2^s}{\omega_s}\mathcal{H}^s_\delta(\Gamma) &\leq \sum_i |U_i|^s = \frac{2}{\delta^{2-s}} \sum_i (\frac{\delta}{\sqrt{2}})^2 = \frac{2}{\delta^{2-s}} \mu(\bigcup_i U_i) \\ &\leq \frac{2}{\delta^{2-s}} \mu(\overline{\Gamma^{(\delta)}}) = \frac{2}{\delta^{2-s}} \mu(\Gamma^{(\delta)}) = 2\omega_{2-s}\mathcal{M}^s_\delta(\Gamma) \end{aligned}$$

To show the second part of the first inequality, let $\{U_i\}$ be any cover of Γ with $\delta \leq |U_i| \leq 2\delta$. Without loss of generality, we assume that $U_i \cap \Gamma$ is non-empty for each i. Select $x_i \in \Gamma \cap U_i$. Then $\cup_i \overline{B}_{|U_i|}(x_i) \supset \cup_i U_i \supset \Gamma$ so that $\cup_i \overline{B}_{2|U_i|}(x_i) \supset \Gamma^{(\delta)}$ since $|U_i| \geq \delta$. Therefore,

$$\mathcal{M}^s_\delta(\Gamma) \leq \frac{\mu(\cup_i \overline{B}_{2|U_i|}(x_i)}{\delta^{2-s}\omega_{2-s}} \leq \frac{\sum_i 4\pi |U_i|^2}{\delta^{2-s}\omega_{2-s}} \leq \frac{4\pi}{\omega_{2-s}} \sum_i \frac{|U_i|^2}{(\frac{|U_i|}{2})^{2-s}} = \frac{2^{4-s}}{\omega_{2-s}} \sum_i |U_i|^s$$

and so

$$\mathcal{M}^s_\delta(\Gamma) \leq \frac{2^{4-s}}{\omega_{2-s}} \inf\{\sum_{i=1}^\infty |U_i|^s \ : \Gamma \subset \bigcup_{i=1}^\infty U_i, \ \delta \leq |U_i| \leq \delta\} = \frac{16}{\omega_s \omega_{2-s}} \mathcal{H}^s_{\delta,2\delta}(\Gamma)$$

□

Note that the definition of $\mathcal{H}^s_{\delta,2\delta}$ is similar to Hausdorff measure, except that the diameter of the covering sets is bounded below as well as above. Hence, its value may be quite different from Hausdorff measure. As an aside, one consequence of the above proposition is the known result that the *Minkowski dimension* of a set is greater than or equal to its *Hausdorff dimension* [35, 80].

We can now prove the following special case of Theorem 2.7.

Theorem 2.8 *If $\Gamma \subset \mathbf{R}^2$ is a compact set with a finite number of connected components then $\mathcal{M}^1(\Gamma) = \mathcal{H}^1(\Gamma)$.*

Proof: Since the connected components of Γ are compact, disjoint, and finite in number, they are positively separated. By additivity of both $\mathcal{M}^1$ and $\mathcal{H}^1$, we need only consider the case in which Γ has one connected component. Hence, we assume that Γ is a continuum. If $\mathcal{H}^1(\Gamma) = \infty$ then $\mathcal{M}^1(\Gamma) = \infty$ from Proposition 2.5. Therefore, we can assume that $\mathcal{H}^1(\Gamma) < \infty$.

Then from Lemma 3.12 of [41], Γ is arcwise connected. Since Γ is compact, we can define a sequence of curves C_j inductively as follows (as in the proof of Lemma 3.13 of [41]). Let C_1 be a curve in Γ joining two of the most distant points of Γ. Given $C_1, C_2, \ldots, C_j$, let $x \in \Gamma$ be at a maximum distance from $\cup_{i=1}^{j} C_i$ and let d_j denote this maximum distance. If $d_j = 0$ then the procedure terminates and we let $C_i = \emptyset$ for $i \geq j+1$. Otherwise, let C_{j+1} be a curve in Γ joining x and $\cup_{i=1}^{j} C_i$ that is disjoint from $\cup_{i=1}^{j} C_i$ except for an endpoint.

Let $E_k = \cup_{j=1}^{k} C_j$. It is shown in [41] (proof of lemma 3.13) that $\mathcal{H}^1(\Gamma) = \mathcal{H}^1(\cup_{i=1}^{\infty} E_k)$. Also,

$$\sum_{j=1}^{\infty} d_j \leq \sum_{j=1}^{\infty} \mathcal{H}^1(C_j) = \mathcal{H}^1(\Gamma) < \infty$$

so that $d_j \to 0$. This implies that $E_k = \cup_{j=1}^{k} C_j \to E$ in Hausdorff metric as $k \to \infty$ and so $\overline{\cup_{k=1}^{\infty} E_k} = \Gamma$. Hence, from Proposition 2.4 and using the fact that $\mathcal{M}^1(A) = \mathcal{M}^1(\overline{A})$ for any A, we get

$$\mathcal{H}^1(\Gamma) = \mathcal{H}^1(\cup_{k=1}^{\infty} E_k) = \mathcal{M}^1(\cup_{k=1}^{\infty} E_k) = \mathcal{M}^1(\overline{\cup_{k=1}^{\infty} E_k}) = \mathcal{M}^1(\Gamma)$$

□

Note that $\mathcal{M}^1$ and $\mathcal{H}^1$ do not agree on all compact sets. An example of a compact set on which they disagree is given in [42] (section 3.2.40).

The final result shown in this section is that Minkowski content possesses a useful type variational convergence property known as Γ-convergence (or epi-convergence). This notion of convergence, introduced by De Giorgi [29, 30] and independently by Attouch [10], is useful in problems involving the convergence of functionals. The result on Γ-convergence will be used in Section 2.4 to prove some convergence properties of solutions to certain formulations of the segmentation problem. Given a topological space (X, τ), and functions $F_n, F : X \to \mathbf{R} \cup \{-\infty, +\infty\}$, the sequence $\{F_n\}$ is said

to be Γ-*convergent* (or *epi-convergent*) to F at $x \in X$ if the following two conditions hold:

(i) for every sequence $\{x_n\}$ converging to x in (X, τ), $F(x) \leq \liminf_{n\to\infty} F_n(x_n)$, and

(ii) there exists a sequence $\{x_n\}$ converging to x in (X, τ) such that $F(x) \geq \limsup_{n\to\infty} F_n(x_n)$.

We will show that for every sequence $\delta_n \to 0$, $\mathcal{M}^1_{\delta_n}$ is Γ-convergent to $\mathcal{M}^1$ on the space of compact subsets of $\mathbf{R}^2$ with a bounded number of connected components and with the topology induced by the Hausdorff metric.

First, we need the following lemma as stated in [41].

Lemma 2.3 *Let C be a collection of balls contained in a bounded subset of $\mathbf{R}^n$. Then there exists a finite or countably infinite disjoint subcollection $\{B_i\}$ such that*

$$\bigcup_{B\in C} B \subset \bigcup_i B_i'$$

where B_i' is the ball concentric with B_i and of three times the radius.

Proof: See [41], Lemma 1.9.

Now the the Γ-convergence of Minkowski content can be shown.

Theorem 2.9 *For every sequence $\delta_n \to 0^+$, $\mathcal{M}^1_{\delta_n}$ is Γ-convergent to $\mathcal{M}^1$ on the space of compact subsets of $\mathbf{R}^2$ with a bounded number of connected components and with the topology induced by the Hausdorff metric. I.e., let $\Gamma \subset \mathbf{R}^2$ be compact with $\#(\Gamma) \leq M < \infty$, and let $\delta_n > 0$ satisfy $\lim_{n\to\infty} \delta_n = 0$. Then the following two conditions hold:*

(i) *For every sequence of compact sets $\Gamma_n \subset \mathbf{R}^2$ with $\Gamma_n \to \Gamma$ in Hausdorff metric and $\#(\Gamma_n) \leq M$ $\forall n$ we have*

$$\mathcal{M}^1(\Gamma) \leq \liminf_{n\to\infty} \mathcal{M}^1_{\delta_n}(\Gamma_n)$$

(ii) *There exists a sequence of compact sets $\Gamma_n \subset \mathbf{R}^2$ with $\Gamma_n \to \Gamma$ in Hausdorff metric and $\#(\Gamma_n) \leq M$ $\forall n$ such that*

$$\mathcal{M}^1(\Gamma) \geq \limsup_{n\to\infty} \mathcal{M}^1_{\delta_n}(\Gamma_n)$$

Proof: Since $\#(\Gamma) \leq M$ we have $\Gamma = \cup_{i=1}^{k} F_i$ where $k \leq M$ and $F_1, F_2, \ldots, F_k$ are the connected components of Γ. Since the F_i are compact and disjoint, they are positively separated, i.e there exists $\eta > 0$ such that $F_i^{(\eta)} \cap F_j^{(\eta)} = \emptyset$ for $i \neq j$. Then $\mathcal{M}^1(\Gamma) = \sum_{i=1}^{k} \mathcal{M}^1(F_i)$, and for sufficiently large n, $\mathcal{M}^1_{\delta_n}(\Gamma_n) = \sum_{i=1}^{k} \mathcal{M}^1_{\delta_n}(\Gamma_n \cap F_i^{(\eta)})$. Thus, it is sufficient to prove the result under the assumption that Γ is connected.

Suppose $\mathcal{H}^1(\Gamma) = \infty$. Form a δ-covering of Γ by placing a closed ball of radius δ about each point of Γ. Then by Lemma 2.3, we can find a disjoint subcollection (necessarily finite) of balls such that concentric balls of radius 3δ cover $\Gamma^{(\delta)}$. Let $N(\delta)$ be the number of balls in this finite disjoint subcollection. Then $6\delta N(\delta) \geq \mathcal{H}^1_{6\delta}(\Gamma) \rightarrow \infty$ as $\delta \rightarrow 0$. Let $\epsilon > 0$. Since $\Gamma_n \rightarrow \Gamma$, for sufficiently large n we have $\Gamma_n \cap B_{\frac{\delta}{2}}(x_i) \neq \emptyset$. Also, since $\#(\Gamma_n) \leq K$, there is a connected component of $\Gamma_n \cap B_\delta(x_i)$ with diameter greater than or equal to $\frac{\delta}{2}$ for at least $N(\delta) - K$ values of i. Using Proposition 2.3 and the fact that the balls are positively separated, we have for sufficiently large n

$$\begin{aligned} \mathcal{M}^1_{\delta_n}(\Gamma_n) &\geq \mathcal{M}^1_{\delta_n}(\Gamma_n \cap \cup_{i=1}^{N(\delta)} B_\delta(x_i)) = \sum_{i=1}^{N(\delta)} \mathcal{M}^1_{\delta_n}(\Gamma_n \cap B_\delta(x_i)) \\ &\geq \sum_{i=1}^{N(\delta)} |\Gamma_n \cap B_\delta(x_i)| \geq \frac{\delta}{2}(N(\delta) - K) \end{aligned}$$

Since $\delta N(\delta) \rightarrow \infty$ as $\delta \rightarrow 0$, $\liminf_{n\rightarrow\infty} \mathcal{M}^1_{\delta_n}(\Gamma_n) = \infty$, and so the result follows.

Now, suppose $\mathcal{H}^1(\Gamma) < \infty$. From Theorem 2.3 we have $\Gamma = S \cup (\bigcup_{i=1}^{\infty} C_i)$ where $\mathcal{H}^1(S) = 0$, and C_i are rectifiable curves. From the construction used in the proof of this result (see [41], also part of the proof is reproduced in the proof of Proposition 2.4), $\mathcal{H}^1(\Gamma) = \sum_{i=1}^{\infty} \mathcal{H}^1(C_i)$ and if $x \in C_i \cap C_j$ then x is an endpoint of at least one of C_i or C_j.

Consider $\cup_{i=1}^{k} C_i$. By decomposing the C_i, we can assume that they are simple curves which meet each other only at endpoints. The C_i are rectifiable curves, so that $C_i : [0,1] \rightarrow \mathbf{R}^2$ and

$$\mathcal{H}^1(C_i) = \mathcal{M}^1(C_i) = \sup\{\sum_{j=1}^{m(i)} |C_i(t_{i,j-1} - C_i(t_{ij})| : 0 = t_{i0} < t_{i1} < \ldots < t_{i,m(i)} = 1\}$$

For each $i = 1, 2, \ldots, k$, let $0 = t_{i0} < t_{i1} < \cdots < t_{i,m(i)} = 1$, and consider the points $x_{ij} = C_i(t_{ij})$.

The connected components of $\cup_{i=1}^{k} C_i \setminus \{x_{ij}\}$ are given by $G_{ij} = \{C_i(t) : t_{i,j-1} <$

$t < t_{ij}\}$ for $1 \leq i \leq k$, $1 \leq j \leq m(i)$. For each i, j, let K_{ij} be a compact subset of G_{ij}. Then the K_{ij} are positively separated since they are a finite collection of disjoint compact sets. Therefore, for some $\eta > 0$, the $\overline{K_{ij}^{(\eta)}}$ are disjoint. Since $\Gamma_n \to \Gamma$ and $\#(\Gamma_n) < M$, for n sufficiently large $\Gamma_n \cap \overline{K_{ij}^{(\eta)}}$ has a connected component whose diameter approaches the diameter of K_{ij} except for at most M values of i, j. I.e., except for at most M values of i, j, there is a connected component T_{nij} of $\Gamma_n \cap \overline{K_{ij}^{(\eta)}}$ such that for every $\epsilon > 0$ there exists $N > 0$ with $|T_{nij}| > |K_{ij}| - \epsilon$ and $\delta_n < \eta$ for all $n \geq N$. Hence, by Proposition 2.3, for all $n \geq N$

$$\begin{aligned}
\mathcal{M}^1_{\delta_n}(\Gamma_n) &\geq \mathcal{M}^1_{\delta_n}(\Gamma_n \cap \bigcup_{i,j} \overline{K_{ij}^{(\eta)}}) \\
&= \sum_{i=1}^{k} \sum_{j=1}^{m(i)} \mathcal{M}^1_{\delta_n}(\Gamma_n \cap \overline{K_{ij}^{(\eta)}}) \\
&\geq \sum_{i=1}^{k} \sum_{j=1}^{m(i)} (|K_{ij}| - \epsilon) - M(\max_{i,j}\{|K_{ij}|\})
\end{aligned}$$

and so

$$\liminf_{n\to\infty} \mathcal{M}^1_{\delta_n}(\Gamma_n) \geq \sum_{i=1}^{k} \sum_{j=1}^{m(i)} |K_{ij}| - M(\max_{i,j}\{|K_{ij}|\})$$

Taking the sup over the compact sets K_{ij} gives

$$\begin{aligned}
\liminf_{n\to\infty} \mathcal{M}^1_{\delta_n}(\Gamma_n) &\geq \sup_{K_{ij}}\{\sum_{i=1}^{k} \sum_{j=1}^{m(i)} |K_{ij}| - M(\max_{i,j}\{|K_{ij}|\})\} \\
&= \sum_{i=1}^{k} \sum_{j=1}^{m(i)} |C_i(t_{i,j-1}) - C_i(t_{ij})| - M(\max_{i,j}\{|C_i(t_{i,j-1}) - C_i(t_{ij})|\})
\end{aligned}$$

Then, taking the sup over the t_{ij} gives

$$\liminf_{n\to\infty} \mathcal{M}^1_{\delta_n}(\Gamma_n) \geq \sum_{i=1}^{k} \mathcal{H}^1(C_i)$$

since $M < \infty$ and $\max_{i,j}\{|C_i(t_{i,j-1}) - C_i(t_{ij})|\} \to 0$ as $\max_{i,j}\{|t_{i,j-1} - t_{ij})|\} \to 0$. Finally, letting $k \to \infty$ gives

$$\liminf_{n\to\infty} \mathcal{M}^1_{\delta_n}(\Gamma_n) \geq \mathcal{H}^1(\Gamma) = \mathcal{M}^1(\Gamma)$$

which proves (i).

To show (ii), take $\Gamma_n = \Gamma$. From Theorem 2.8, $\mathcal{M}^1(\Gamma) = \mathcal{H}^1(\Gamma)$ so that in particular $\lim_{\delta \to 0} \mathcal{M}^1_\delta(\Gamma) = \mathcal{M}^1(\Gamma)$ exists. Hence, for every sequence $\delta_n \to 0$, condition (ii) is satisfied by taking $\Gamma_n = \Gamma$.

□

Note that Theorem 2.9 is not true in general if the bound on the number of connected components is dropped. For example, let $r_1, r_2, \ldots$ denote an enumeration of the rationals between 0 and 1. Take $\Gamma_n = \{(r_i, 0) : 1 \leq i \leq n\}$ and $\delta_n = 1/n^2$. Then $\Gamma_n \to \Gamma = \{(x, 0) : 0 \leq x \leq 1\}$, but $\mathcal{M}^1_{\delta_n}(\Gamma_n) \leq \frac{1}{2}\pi n \delta_n \to 0$ while $\mathcal{M}^1(\Gamma) = 1$. However, the restriction on the number of connected components can be dropped if we impose the additional assumption that $d_H(\Gamma_n, \Gamma)/\delta_n \to 0$ as $n \to \infty$.

2.4 Discrete Approximations to the Segmentation Problem

In this section, we apply some results of the previous sections to the variational method for image segmentation discussed in the introduction. As before, g represents an observed image defined on a bounded open set $\Omega \subset \mathbf{R}^2$, f is the reconstructed image, and Γ are the boundaries of the image. In the variational approach, f and Γ are obtained by minimizing the cost functional 2.1 or 2.2. Normally, g is assumed to be in $L^\infty(\Omega)$, Γ is a closed subset of $\overline{\Omega}$, and f is in the Sobolev space $W^{1,2}(\Omega \setminus \Gamma)$. Under certain regularity assumptions, a number of interesting results concerning the nature of the minimizing f and Γ have been obtained [23, 89, 97, 122]. Also, the existence of a minimizing pair (f, Γ) for various versions of the problem has been shown [4, 5, 96].

Here we are concerned with the behavior of solutions to discrete versions of the problem as the lattice spacing tends to zero. Specifically, we are interested in whether or not solutions to the discrete problem converge to a solution of the continuous problem. It seems that this may not necessarily be the case for the discrete problem of 2.3. For example, consider the segmentation problem 2.2 where f is required to be piecewise constant. Take $\Omega = (0,1) \times (0,1)$, $g(x,y) = 0$ for $x < y$ and $g(x,y) = 1$ otherwise, and $4\sqrt{2}c_3 < c_1 < 8c_3$. It is generally difficult to prove optimality of solutions to these types of problems, but for simple examples one can obtain evidence of optimality by comparing the costs of various natural candidate solutions. For the

example above, the optimal solution to the discrete problem with sufficiently small lattice spacing seems to be $\Gamma = \emptyset$, while the optimal solution to the continuous problem seems to be $\Gamma = \{(x, x) : 0 \leq x \leq 1\}$. For completeness, it would be nice to actually prove optimality of these solutions.

The failure of convergence in the continuum limit appears to be a result of the possible strict lower semicontinuity of the length of curves with respect to the Hausdorff metric. E.g., in this case, if $\Gamma = \{(x, x) : 0 \leq x \leq 1\}$ and Γ_n is the discrete approximation to Γ with lattice spacing $1/n$, then $\Gamma_n \to \Gamma$ but $L(\Gamma) = \sqrt{2}$ while $\lim_{n\to\infty} L(\Gamma_n) = 2$. The notion of length in the discrete case does not coincide in the continuum limit with the usual measure of length of the limiting boundary.

As previously mentioned, it may be possible to resolve this problem by modifying the cost functional and/or the discretization process. In this section, we present three methods for discretizing the variational method for image segmentation for which we prove desirable convergence properties in the continuum limit. Other approaches to digitizing the problem have been considered and implemented (e.g., in [97, 45]), although proofs of convergence of these methods are lacking (see Section 2.5 for more discussion).

2.4.1 Minkowski Content as Cost for Boundaries

Here we consider the use of Minkowski content for the cost of the boundaries and propose a modified discrete version of the problem. Specifically, given an observed image $g \in L^\infty(\Omega)$ we consider the problem of minimizing

$$E_\delta(f, \Gamma) = c_1 \iint_\Omega (f - g)^2 \, dx \, dy + c_2 \iint_{\Omega \setminus \Gamma} \| \nabla f \|^2 \, dx \, dy + c_3 \mathcal{M}_\delta(\Gamma)$$

with Γ a closed subset of $\overline{\Omega}$ and $f \in W^{1,2}(\Omega \setminus \Gamma)$.

For the discrete version of the problem with lattice spacing $\frac{1}{n}$, we simply restrict Γ to be composed of a union of closed lattice squares of $\frac{1}{n}\mathbf{Z}^2$. Specifically, for integers i, j let

$$S_{ij} = \{(x_1, x_2) \in \mathbf{R}^2 : i \leq x_1 \leq i + 1, j \leq x_2 \leq j + 1\}$$

S_{ij} is the closed unit square whose lower left corner is at (i, j). Then $\frac{1}{n} S_{ij}$ is a square with side $\frac{1}{n}$ whose lower left corner is at $(i/n, j/n)$. For the discrete problem with lattice spacing $\frac{1}{n}$, the boundaries consist of a union of the $\frac{1}{n} S_{ij}$ — i.e., a union of closed lattice squares of $\frac{1}{n}\mathbf{Z}^2$. The discrete version with spacing $\frac{1}{n}$ of a boundary

$\Gamma \subset \mathbf{R}^2$ is given by $P_n(\Gamma)$ where $P_n(\cdot)$ maps Γ to the subset of $\mathbf{R}^2$ consisting of the union of the $\frac{1}{n}S_{ij}$ that Γ intersects:

$$P_n(\Gamma) = \bigcup_{\substack{i,j \text{with} \\ \frac{1}{n}S_{ij} \cap \Gamma \neq \emptyset}} S_{ij}$$

However, we still take g and f to be defined on the continuous domains Ω and $\Omega \setminus \Gamma$ respectively. Hence, we have only incorporated a partial discretization, i.e. we have only discretized Γ. However, the primary difficulty in numerical solutions is to properly deal with the boundary. For a fixed Γ, the minimization reduces to a standard variational problem whose Euler-Lagrange equations can be solved by standard algorithms for partial differential equations.

We now give some results concerning the problem of minimizing E_δ.

Theorem 2.10 *For every $\delta > 0$, there exists a pair $(f_\delta, \Gamma_\delta)$ which minimizes E_δ.*

Proof: Since we have shown that $\mathcal{M}_\delta$ is continuous (Theorem 2.6), the existence proof of Richardson [96] can be applied.

□

Note that for any bounded Γ, $\mathcal{M}_\delta(\Gamma) < \infty$. Hence, a minimizing boundary may quite possibly have nonzero Lebesgue measure.

The next theorem establishes the desirable property of discrete to continuous convergence for E_δ with a fixed $\delta > 0$. We will use the same notion of convergence as used in [96]. For $f \in W^{1,2}(\Omega \setminus \Gamma)$, f and its weak first order derivatives $D_{x_i} f$, $i = 1, 2$, can be considered as functions in $L^2(\Omega)$ by defining them to be zero on Γ. By $(f_n, \Gamma_n) \rightarrow (f, \Gamma)$ we mean that $\Gamma_n \rightarrow \Gamma$ in Hausdorff metric and that for the extended functions $f_n \rightarrow f$, $D_{x_i} f_n \rightarrow D_{x_i} f$, $i = 1, 2$ weakly in $L^2(\Omega)$.

Theorem 2.11 *Let $(f^*_{\delta,n}, \Gamma^*_{\delta,n})$ denote a minimizing pair for $E_{\delta,n}$, i.e. for the discrete problem E_δ with lattice spacing $\frac{1}{n}$. Then there exists a subsequence (still denoted $(f^*_{\delta,n}, \Gamma^*_{\delta,n})$) and a pair $(f_\delta, \Gamma_\delta)$ such that $(f^*_{\delta,n}, \Gamma^*_{\delta,n}) \rightarrow (f_\delta, \Gamma_\delta)$ and $(f_\delta, \Gamma_\delta)$ minimizes E_δ.*

Proof: The existence of a pair $(f_\delta, \Gamma_\delta)$ with $(f^*_{\delta,n}, \Gamma^*_{\delta,n}) \rightarrow (f_\delta, \Gamma_\delta)$ follows from Lemma 3 of [96]. We only need to show that $(f_\delta, \Gamma_\delta)$ minimizes E_δ.

Let $(f_\delta^*, \Gamma_\delta^*)$ minimize E_δ. For each n, let Λ_n be obtained from Γ_δ^* by taking the smallest cover of Γ_δ^* using the closed lattice squares of the lattice with spacing $\frac{1}{n}$. Let h_n be the restriction of f_δ^* to $\Omega \setminus \Lambda_n$. From Theorem 2.6, $\lim_{n\to\infty} E_\delta(h_n, \Lambda_n) = E_\delta(f_\delta^*, \Gamma_\delta^*)$. Then, by the lower-semicontinuity of E_δ and the optimality of $(f_{\delta,n}^*, \Gamma_{\delta,n}^*)$ for the discrete problem with lattice spacing $\frac{1}{n}$, we have

$$\begin{aligned} E_\delta(f_\delta, \Gamma_\delta) &\leq \liminf_{n\to\infty} E_\delta(f_{\delta,n}, \Gamma_{\delta,n}) \leq \liminf_{n\to\infty} E_\delta(h_n, \Lambda_n) \\ &= \lim_{n\to\infty} E_\delta(h_n, \Lambda_n) = E_\delta(f_\delta^*, \Gamma_\delta^*) \end{aligned}$$

Therefore, $E_\delta(f_\delta, \Gamma_\delta) = E_\delta(f_\delta^*, \Gamma_\delta^*)$ so that $(f_\delta, \Gamma_\delta)$ minimizes E_δ.

□

A natural question at this point concerns the behavior of $(f_\delta^*, \Gamma_\delta^*)$ as $\delta \to 0$. One would like $(f_\delta^*, \Gamma_\delta^*)$ to converge to a minimizing solution of the original cost functional E. We can show a convergence result if the number of connected components of the admissible boundaries is uniformly bounded. I.e., following [96], we let the cost term for the boundaries be

$$\nu_\delta(\Gamma) = \mathcal{M}_\delta^1(\Gamma) + F(\#(\Gamma))$$

where $F(k) = 0$ for $k \leq M < \infty$ and $F(k) = \infty$ for $k > M$. Let E_δ^M denote the cost functional with the above boundary term, and let E^M denote the cost functional whose boundary term is

$$\nu(\Gamma) = \mathcal{M}^1(\Gamma) + F(\#(\Gamma))$$

By Theorem 2.8, $\mathcal{M}^1(\Gamma)$ in the equation for $\nu(\Gamma)$ could equivalently be replaced by $\mathcal{H}^1(\Gamma)$. For these variational problems, we have the following convergence result, which essentially follows from the result on the Γ-convergence of Minkowski content (Theorem 2.9).

Theorem 2.12 *Let $(f_\delta^*, \Gamma_\delta^*)$ denote a minimizing pair for E_δ^M, and let $\delta_n \to 0^+$. Then there is a subsequence (which we still denote by δ_n) such that $(f_{\delta_n}^*, \Gamma_{\delta_n}^*) \to (f, \Gamma)$ for some (f, Γ) which minimizes E^M. Furthermore, $E_{\delta_n}^M(f_{\delta_n}^*, \Gamma_{\delta_n}^*) \to E^M(f, \Gamma)$.*

Proof: The existence of a pair (f, Γ) with $(f_{\delta,n}^*, \Gamma_{\delta,n}^*) \to (f, \Gamma)$ follows from Lemma 3 of [96]. We need to show that (f, Γ) minimizes E_δ and that $E_{\delta_n}^M(f_{\delta_n}^*, \Gamma_{\delta_n}^*) \to E^M(f, \Gamma)$.

This follows from Theorem 2.9 on the epi-convergence of Minkowski content in the case of a bounded number of connected components. Specifically,

$$E^M(f,\Gamma) \leq \liminf_{n\to\infty} E^M_{\delta_n}(f_{\delta_n},\Gamma_{\delta_n}) \leq \liminf_{n\to\infty} E^M_{\delta_n}(f^*,\Gamma^*) = E^M(f^*,\Gamma^*)$$

so that (f,Γ) minimizes E^M.

Also, we have

$$E^M(f,G) = \limsup_{n\to\infty} E^M_{\delta_n}(f,\Gamma) \geq \limsup_{n\to\infty} E^M_{\delta_n}(f_{\delta_n},\Gamma_{\delta_n})$$

Thus,

$$\limsup_{n\to\infty} E^M_{\delta_n}(f,\Gamma) \leq E^M(f,\Gamma) \leq \liminf_{n\to\infty} E^M_{\delta_n}(f,\Gamma)$$

and so

$$E^M(f,\Gamma) = \lim_{n\to\infty} E^M_{\delta_n}(f,\Gamma)$$

□

Finally, we give a result concerning the convergence of solutions when the lattice spacing and δ are simultaneously allowed to go to zero. The following theorem guarantees convergence of a subsequence to a solution of the continuous problem if $\delta \to 0$ at a rate slower than the lattice spacing.

Theorem 2.13 *Let $\delta_n > 0$ with $\delta_n \to 0$ and let $(f^*_{\delta_n,n},\Gamma^*_{\delta_n,n})$ denote a minimizing pair for $E^M_{\delta_n,n}$, i.e. for the discrete problem $E^M_{\delta_n}$ with lattice spacing $\frac{1}{n}$. If $n\delta_n \to \infty$ as $n \to \infty$ then there exists a subsequence (still denoted $(f^*_{\delta_n,n},\Gamma^*_{\delta_n,n})$) and a pair (f,Γ) such that $(f^*_{\delta_n,n},\Gamma^*_{\delta_n,n}) \to (f,\Gamma)$ and (f,Γ) minimizes E^M.*

Proof: As before, the existence of a pair (f,Γ) with $(f^*_{\delta_n,n},\Gamma^*_{\delta_n,n}) \to (f,\Gamma)$ follows from Lemma 3 of [96], and so we need to show that (f,Γ) minimizes E^M.

Let (f^*,Γ^*) minimize E^M, and for each n let (h_n,Λ_n) be obtained from (f^*,Γ^*) as in the proof of Theorem 2.11. Namely, Λ_n is the smallest cover of Γ^* using lattice squares of the lattice with spacing $\frac{1}{n}$, and h_n is the restriction of f^* to $\Omega\backslash\Lambda_n$. Then using Theorem 2.9 and the optimality of $(f^*_{\delta_n,n},\Gamma^*_{\delta_n,n})$ we have

$$E^M(f,\Gamma) \leq \liminf_{n\to\infty} E^M_{\delta_n}(f^*_{\delta_n,n},\Gamma^*_{\delta_n,n}) \leq \liminf_{n\to\infty} E^M_{\delta_n}(h_n,\Lambda_n)$$

Since Λ_n is the minimal cover of Γ^* on the lattice with spacing $\frac{1}{n}$, we have $\Lambda_n \subset$

$(\Gamma^*)^{(\frac{\sqrt{2}}{n})}$ so that

$$\begin{aligned}\liminf_{n\to\infty} \mathcal{M}^1_{\delta_n}(\Lambda_n) &\leq \liminf_{n\to\infty} \frac{\mu((\Gamma^*)^{(\delta_n+\frac{\sqrt{2}}{n})})}{2\delta_n} \\ &= \liminf_{n\to\infty} \frac{\mu((\Gamma^*)^{(\delta_n+\frac{\sqrt{2}}{n})})}{2(\delta_n+\frac{\sqrt{2}}{n})} \frac{\delta_n+\frac{\sqrt{2}}{n}}{\delta_n} \\ &= \lim_{n\to\infty} \frac{\mu((\Gamma^*)^{(\delta_n+\frac{\sqrt{2}}{n})})}{2(\delta_n+\frac{\sqrt{2}}{n})}(1+\frac{\sqrt{2}}{n\delta_n}) = \mathcal{M}^1(\Gamma^*)\end{aligned}$$

It follows that

$$\liminf_{n\to\infty} E^M_{\delta_n}(h_n, \Lambda_n) \leq E^M(f^*, \Gamma^*)$$

Therefore, $E^M(f,\Gamma) \leq E^M(f^*,\Gamma^*)$ so that (f,Γ) minimizes E^M.

□

2.4.2 Alternative Cost for Discrete Boundaries

In this section we propose another discrete version of 2.1 which possesses the desirable convergence property in the continuum limit. As in the discretized version using Minkowski content in the previous section, the discrete boundaries consist of unions of discrete closed lattice squares. However, we assign a different cost (measure of length) to the discrete boundaries. It is shown that this alternate measure of length satisfies a convergence property from which convergence of solutions to the variational problem follows.

As in the previous section, for the discrete version of problem 2.1 with lattice spacing $\frac{1}{n}$, we restrict the boundaries to be composed of a union of closed lattice squares of $\frac{1}{n}\mathbf{Z}^2$. We now define an alternate cost for discrete boundaries (or rather discrete approximations to length for any boundary). Let $\Gamma \subset \mathbf{R}^2$. First, suppose that $P_n(\Gamma)$ (the discrete version on the lattice with spacing $1/n$) is connected. In this case, define $\mathcal{L}_n(\Gamma)$ by

$$\mathcal{L}_n(\Gamma) = \inf\{\mathcal{H}^1(\Lambda) : \Lambda \text{ connected and } P_n(\Lambda) = P_n(\Gamma)\}$$

In the general case, define $\mathcal{L}_n(\Gamma)$ by

$$\mathcal{L}_n(\Gamma) = \sum_i \mathcal{L}_n(G_i)$$

where the G_i are the connected components of $P_n(\Gamma)$.

The discrete measure of length assigned to a connected discrete boundary $P_n(\Gamma)$ is the minimum length of a connected curve that gives rise to the same discrete boundary. Related measures of length for discrete boundaries have been studied in [68, 85, 86]. Note that for bounded Γ, the inf could actually be replaced by min since the infimum is achieved. This follows from the compactness result of compact sets under the Hausdorff metric and the lower semicontinuity of Hausdorff measure. The sum in the extension to arbitrary boundaries is well defined since any discrete boundary has at most a countable number of connected components. In fact, a bounded discrete boundary has a finite number of connected components. Two boundaries that give rise to the same discrete boundary have the same discrete measure of length.

Using this discrete measure of length, we can define a partially discrete version of the original variational problem 2.1 with lattice spacing $\frac{1}{n}$. The boundaries are reconstructed only to within their $\frac{1}{n}$ discrete versions, but the observed and reconstructed images are still defined on continuous domains. Specifically, given an observed image $g \in L^\infty(\Omega)$ we reconstruct a discrete boundary $P_n(\Gamma)$ and an image f on $\Omega \setminus P_n(\Gamma)$ by minimizing

$$E_n(f, P_n(\Gamma)) = c_1 \iint_{\Omega \setminus P_n(\Gamma)} (f-g)^2 \, dx \, dy + c_2 \iint_{\Omega \setminus P_n(\Gamma)} \| \nabla f \|^2 \, dx \, dy + c_3 \mathcal{L}_n(P_n(\Gamma)) \tag{2.4}$$

We now discuss some properties of the discrete measure of length and the associated discrete variational problem.

Theorem 2.14 *For every fixed $n < \infty$, a minimizing solution $(f^*, P_n(\Gamma^*))$ for $E_n(\cdot, \cdot)$ exists.*

Proof: For a fixed boundary, the minimization of E_n is a standard variational problem for which a solution exists. Since Ω is assumed to be bounded, there are only a finite number of distinct discrete boundaries $P_n(\Gamma)$, and so the result follows.

□

As previously mentioned, common way to prove existence to variational problems is to show a compactness property and lower-semicontinuity of the cost function.

Due to the finite number of possible discrete boundaries, compactness and lower-semicontinuity over the set of discrete boundaries are immediate. However, $\mathcal{L}_n(\cdot)$ is not lower-semicontinuous over *all* Γ. For a simple example, take $n = 1$, and let Γ_k be the straight line joining $(\frac{1}{2}, \frac{1}{2})$ and $(2 - \frac{1}{k}, \frac{1}{2})$. Then $P_1(\Gamma_k) = S_{00} \cup S_{10}$ so that $\mathcal{L}_1(\Gamma_k) = 0$. Also, the Γ_k converge to the straight line Γ joining $(\frac{1}{2}, \frac{1}{2})$ and $(2, \frac{1}{2})$, so that $P_1(\Gamma) = S_{00} \cup S_{10} \cup S_{20}$. Therefore, $\mathcal{L}_1(\Gamma) = 1 > \liminf_{k\to\infty} \mathcal{L}_1(\Gamma_k) = 0$.

The following result shows that for a large class of boundaries the discrete measures of length converge to the usual notion of length as $n \to \infty$.

Theorem 2.15 *If Γ is a compact set with a finite number of connected components then $\lim_{n\to\infty} \mathcal{L}_n(\Gamma) = \mathcal{H}^1(\Gamma)$.*

Proof: If $G_1, \ldots, G_m$ are the connected components of Γ, then for sufficiently large n the $P_n(G_i)$ are disjoint for $i = 1, \ldots, m$. In that case, $\mathcal{L}_n(\Gamma) = \sum_{i=1}^m \mathcal{L}_n(G_i)$. Also, $\mathcal{H}^1(\Gamma) = \sum_{i=1}^m \mathcal{H}^1(G_i)$. The result will follow if it can be shown for each connected component. Therefore, suppose that Γ is connected.

By the definition of $\mathcal{L}_n(\Gamma)$, we have $\mathcal{L}_n(\Gamma) \leq \mathcal{H}^1(\Gamma)$ so that $\limsup_{n\to\infty} \mathcal{L}_n(\Gamma) \leq \mathcal{H}^1(\Gamma)$. On the other hand, for each n, there exists a compact and connected Λ_n with $P_n(\Lambda_n) = P_n(\Gamma)$ and $\mathcal{H}^1(\Lambda_n) = \mathcal{L}_n(\Gamma)$. Since $P_n(\Lambda_n) = P_n(\Gamma)$, $\Lambda_n \to \Gamma$ in Hausdorff metric as $n \to \infty$. By the lower-semicontinuity of $\mathcal{H}^1$ (see [41], Theorem 3.18) we have $\mathcal{H}^1(\Gamma) \leq \liminf_{n\to\infty} \mathcal{H}^1(\Lambda_n) = \liminf_{n\to\infty} \mathcal{L}_n(\Gamma)$. Thus, $\lim_{n\to\infty} \mathcal{L}_n(\Gamma)$ exists and equals $\mathcal{H}^1(\Gamma)$.

□

The following theorem shows that the discrete measures of length converge in another useful way to Hausdorff measure.

Theorem 2.16 *$\mathcal{L}_n(\cdot)$ is Γ-convergent to $\mathcal{H}^1(\cdot)$ on the space of compact subsets of $\mathbf{R}^2$ with a bounded number of connected components and with the topology induced by the Hausdorff metric. I.e., let $\Gamma \subset \mathbf{R}^2$ be compact with $\#(\Gamma) \leq M < \infty$. Then the following two conditions hold:*

(i) *For every sequence of compact sets $\Gamma_n \subset \mathbf{R}^2$ with $\Gamma_n \to \Gamma$ in Hausdorff metric and $\#(\Gamma_n) \leq M$ $\forall n$ we have*

$$\mathcal{H}^1(\Gamma) \leq \liminf_{n\to\infty} \mathcal{L}_n(\Gamma_n)$$

(ii) *There exists a sequence of compact sets* $\Gamma_n \subset \mathbf{R}^2$ *with* $\Gamma_n \to \Gamma$ *in Hausdorff metric and* $\#(\Gamma_n) \leq M \ \forall n$ *such that*

$$\mathcal{H}^1(\Gamma) \geq \limsup_{n\to\infty} \mathcal{L}_n(\Gamma_n)$$

Proof: To show (i), let Γ_n be any sequence of compact sets with $\#(\Gamma_n) \leq M$ and $\Gamma_n \to \Gamma$ in Hausdorff metric. Then $\#(P_n(\Gamma_n)) \leq \#(\Gamma_n) \leq M$. Therefore, for each n there exists Λ_n with $\#(\Lambda_n) = \#(P_n(\Gamma_n)) \leq M$, $P_n(\Lambda_n) = P_n(\Gamma_n)$, and $\mathcal{H}^1(\Lambda_n) = \mathcal{L}_n(\Gamma_n)$. Since $\Gamma_n \to \Gamma$, we have $P_n(\Gamma_n) \to \Gamma$ and also since $P_n(\Lambda_n) = P_n(\Gamma_n)$ we have $\Lambda_n \to \Gamma$. Hence, by the lower-semicontinuity of Hausdorff measure for a bounded number of connected components (as mentioned in Section 2.2 — see [96, 97]), we get $\mathcal{H}^1(\Gamma) \leq \liminf_{n\to\infty} \mathcal{H}^1(\Lambda_n) = \liminf_{n\to\infty} \mathcal{L}_n(\Gamma_n)$.

To show (ii), simply take $\Gamma_n = \Gamma$ for all n. Then by the definition of $\mathcal{L}_n(\Gamma)$, we have $\mathcal{H}^1(\Gamma) \geq \mathcal{L}_n(G) = \mathcal{L}_n(\Gamma_n)$ for all n and so $\mathcal{H}^1(\Gamma) \geq \limsup_{n\to\infty} \mathcal{L}_n(\Gamma_n)$.

□

The Γ-convergence property shown above is sufficient to show convergence of solutions to the discrete problem 2.4 as the lattice spacing $\frac{1}{n}$ goes to zero if the number of connected components of the admissible boundaries is uniformly bounded. I.e., following [96], we let the cost term for the boundaries be

$$\nu_n(\Gamma) = \mathcal{L}_n(\Gamma) + F(\#(\Gamma))$$

where $F(k) = 0$ for $k \leq M < \infty$ and $F(k) = \infty$ for $k > M$. Let E_n^M denote the cost functional with the above boundary term, and let E^M denote the cost functional whose boundary term is

$$\nu(\Gamma) = \mathcal{H}^1(\Gamma) + F(\#(\Gamma))$$

For these variational problems, we have the following convergence result, which essentially follows from the previous result on Γ-convergence.

Theorem 2.17 *Let* (f_n^*, Γ_n^*) *denote a minimizing pair for* E_n^M. *Then there exists a subsequence (still denoted* (f_n^*, Γ_n^*)*) and a pair* (f, Γ) *such that* $(f_n^*, \Gamma_n^*) \to (f, \Gamma)$ *and* (f, Γ) *minimizes* E^M.

Proof: The existence of a pair (f, Γ) with $(f_n^*, \Gamma_n^*) \to (f, \Gamma)$ follows from Lemma 3 of [96]. The fact that (f, Γ) minimizes E^M follows from the Γ-convergence of $\mathcal{L}_n$ to $\mathcal{H}^1$ (shown above) which implies the Γ-convergence of E_n^M to E^M.

□

2.4.3 Segmentation with Piecewise Linear Boundaries

In this section we formulate a modified version of the original variational problem 2.1 which eliminates the problems associated with computing the length of irregular curves. The modification consists of requiring the boundaries Γ to consist of a union of a bounded number of straight line segments. We propose a partially discrete version of the modified problem and show a convergence result for the solutions to the discrete problems.

Let $\mathcal{S}^M = \mathcal{S}^M(\Omega)$ be the set of all compact subsets of Ω that are the union of $M < \infty$ or fewer connected line segments contained in Ω. Consider the problem of minimizing 2.1 subject to the constraint that $\Gamma \in \mathcal{S}^M$. Alternatively, we can consider this problem as one of minimizing the cost function $E_{\mathcal{S}^M}$ whose boundary term is $\mathcal{H}^1(\Gamma)$ for $\Gamma \in \mathcal{S}^M$ and infinite otherwise.

Theorem 2.18 *For every fixed $M < \infty$ and Ω a bounded region of $\mathbf{R}^2$, $\mathcal{S}^M(\Omega)$ is compact with respect to the Hausdorff metric. I.e., for every infinite sequence $\Gamma_n \in \mathcal{S}^M(\Omega)$ there is a subsequence, still denoted Γ_n, and $\Gamma \in \mathcal{S}^M(\Omega)$ such that $\Gamma_n \to \Gamma$ in Hausdorff metric.*

Proof: First, each Γ_n is a compact subset of Ω. As previously mentioned, the set of compact subsets of Ω is compact with respect to the Hausdorff metric (e.g., see [41], Theorem 3.16). Hence, there is a subsequence, which we still denote Γ_n, and a compact set $\Gamma \subset \Omega$ such that $\Gamma_n \to \Gamma$ in Hausdorff metric. We need only show that $\Gamma \in \mathcal{S}^M(\Omega)$, i.e., that Γ is the union of at most M line segments.

Note that by a line segment we include the possibility of a single point or the empty set. It is straightforward to show the result for $M = 1$. I.e., given a sequence of single line segments any convergent subsequence converges to a single line segment. Now, each Γ_n is the union of at most M line segments so that $\Gamma_n = \cup_{i=1}^M A_{n,i}$ where each $A_{n,i}$ is a line segment (possibly a point or the empty set).

We can extract a subsequence (still indexed by n) so that $A_{n,1} \to A_1$ where A_1 is a line segment (again, possibly a point or the empty set). Similarly, we can extract sub-subsequences $M-1$ more times (still indexed by n) so that for each $i = 1, \ldots, M$, $A_{n,i} \to A_i$ where each A_i is a line segment.

Therefore, for the final sub-subsequence, $\Gamma_n = \cup_{i=1}^M A_{n,i}$ and for each $i = 1, \ldots, m$ we have $A_{n.i} \to A_i$ where A_i is a line segment. Hence, $\Gamma_n \to \Gamma = \cup_{i=1}^M A_i$ and so

$\Gamma \in \mathcal{S}^M(\Omega)$.

□

Theorem 2.19 *A minimizing solution* (f^*, Γ^*) *for* $E_{\mathcal{S}M}(\cdot,\cdot)$ *exists.*

Proof: This follows from the compactness of $\mathcal{S}^M$ shown above and the lower-semicontinuity of the cost function (e.g., see [96, 97]).

□

Now we consider partially discrete versions of the problem $E_{\mathcal{S}M}$. As before, only the boundary is discretized. For this problem we consider a different form of discrete representation for the boundaries, taking advantage of the fact that the boundaries are piecewise linear. Specifically, we consider the set of lattice points in $\frac{1}{n}\mathbf{Z}^2$, and require the endpoints of each line segment in the boundary to lie on these lattice points. Let $\mathcal{S}^{M,n} = \mathcal{S}^{M,n}(\Omega)$ denote the collection of all sets consisting of at most M straight line segments whose endpoints lie in $\frac{1}{n}\mathbf{Z}^2 \cap P_n(\Omega)$. To obtain a convergence result we will consider the boundary to be a dilation of the linear segments where the amount of dilation is related to the lattice spacing. For the discrete problem with lattice spacing $1/n$, the cost function is minimized over boundaries of the form $\Gamma^{(\delta_n)}$ where $\delta_n > \sqrt{2}/n$ with $\delta_n \to 0$ as $n \to \infty$ and $\Gamma \in \mathcal{S}^{M,n}(\Omega)$. The reconstructed image is then defined only on $\Omega \setminus \Gamma^{(\delta_n)}$. However, the cost of the dilated boundary $\Gamma^{(\delta_n)}$ is taken to be simply the the total length of the straight line segments comprising Γ. Let $E_{\mathcal{S}M,n}$ denote the discrete problem with lattice spacing $1/n$. For these discrete problems, we have the following convergence result.

Theorem 2.20 *Let* (f_n^*, Γ_n^*) *denote a minimizing pair for* $E_{\mathcal{S}M,n}$. *Then there exists a subsequence (still denoted* (f_n^*, Γ_n^*)*) and a pair* (f, Γ) *such that* $(f_n^*, \Gamma_n^*) \to (f, \Gamma)$ *and* (f, Γ) *minimizes* $E_{\mathcal{S}M}$.

Proof: The existence of a pair (f, Γ) with $(f_n^*, \Gamma_n^*) \to (f, \Gamma)$ follows from the compactness of $\mathcal{S}^M$ and Lemma 3 of [96]. We only need to show that (f, Γ) minimizes $E_{\mathcal{S}M}$.

Let (f^*, Γ^*) minimize $E_{\mathcal{S}M}$. For each n, let Λ_n be obtained from Γ^* by taking the best approximation to Γ^* using line segments whose endpoints lie in $\frac{1}{n}\mathbf{Z}^2 \cap P_n(\Omega)$, i.e. the best boundary which can be used in the discrete problem with lattice spacing $\frac{1}{n}$. This will be obtained by taking the line segments whose endpoints lie on the lattice

sites closest to the endpoints of Γ^*. Note that certainly $\Lambda_n^{(\delta_n)} \supset \Gamma^*$ since $\delta_n > \sqrt{2}/n$ and with lattice spacing $\frac{1}{n}$ there is always a lattice point within $\sqrt{2}/n$ of any point in Ω. Let h_n be the restriction of f^* to $\Omega \setminus \Lambda_n^{(\delta_n)}$. Then, we have

$$\begin{aligned} E_{SM}(f,\Gamma) &\leq \liminf_{n\to\infty} E_{SM,n}(f_n^*,\Gamma_n^*) \leq \liminf_{n\to\infty} E_{SM,n}(h_n,\Lambda_n) \\ &= \lim_{n\to\infty} E_{SM,n}(h_n,\Lambda_n) = E_{SM}(f^*,\Gamma^*) \end{aligned}$$

where the first inequality uses the lower-semicontinuity of E_{SM} (which follows from results from [96]), the second inequality follows from the optimality of (f_n^*,Γ_n^*), and the equalities follow from the continuity of the first two terms of E_{SM} with respect to dilations of the boundary (see [96]). Therefore, $E_{SM}(f,\Gamma) = E_{SM}(f^*,\Gamma^*)$ so that (f,Γ) minimizes E_{SM}.

□

2.5 Discussion and Open Problems

We obtained several new properties of Minkowski content, and gave new proofs for some results already known. Refinements for many of the results may be interesting to consider. For example, Lemma 2.1 shows that the boundary of a dilated set cannot be too irregular. We expect that much stronger statements concerning the regularity of the boundary could be made. For example, if A and P denote the area and perimeter of a δ-dilation, then we conjecture that $P/A \leq 2/\delta$ with equality iff the δ-dilation consists of a union of disjoint circles of radius δ. This is a kind of "reverse" isoperimetric inequality for δ-dilations. As a general direction, it may be worth investigating whether the results obtained for Minkowski content and dilations are applicable to some problems in mathematical morphology [107].

Regarding the segmentation problem, we provided some evidence suggesting that the standard discretization fails to approximate the continuous formulation (for completeness, it would be nice to actually prove this). Hence, analytical results concerning the continuous formulation cannot be applied to the standard discrete solutions (even for a very fine lattice spacing). It may be interesting to study the continuous variational principle which corresponds to the standard discretization. The natural conjecture is that in the corresponding continuous formulation the cost term for a curve is $\int_0^L (|\cos\theta(s)| + |\sin\theta(s)|)ds$ where s is arc length and $\theta(s)$ is the angle

between the tangent to curve and a fixed x-axis, rather than the length of the curve.

We presented three procedures for partially discretizing the variational formulation of the segmentation problem for which we proved convergence results in the continuum limit. Our results provide a rigorous justification for certain finite-difference-like approximation to a variational problem explicitly incorporating boundaries. We expect that the methodologies presented could be used to analyze discrete approximations to other "free-discontinuity" variational problems. In particular, the approach may be applicable to other problems in early vision which explicitly minimize over boundaries.

There are several reasons for considering a variety of discrete approximation methods such as those presented here. First, the proofs of the convergence results for two of the techniques presented here are much simpler since we were able to use some powerful results on Hausdorff measure, rather than deriving new results as we did for the method using Minkowski content. Second, certain discretization methods and cost terms may be easier to implement and computationally more advantageous than others. Finally, considering several alternative approaches might suggest certain general properties that are shared by all discretizations of the original problem.

Regarding the second and third points we have some specific ideas in mind. Both approaches presented here as well as the method using Minkowski content are computationally unattractive compared to the standard discretization. For example, in the standard discretization the cost term associated with a discrete boundary is obtained by simply taking the total length of the segments in the discrete boundary. On the other hand, with the digitizations proposed, the computation of the cost terms of Sections 2.4.1 and 2.4.2 is much more involved. In particular, the MRF corresponding to the standard discretization is very simple, requiring no interactions between the boundary sites, while implementing the discretizations proposed would require large neighborhood structures (growing unboundedly as the lattice spacing tends to zero) and complex potentials.

The distinction can also be formalized along the following lines. Consider a distributed implementation in which there is a processor at each lattice site. The state of a processor is either zero or one depending on whether or not the boundary passes through the associated lattice square. To compute the length terms of Sections 2.4.2 or 2.4.1, each processor must perform a computation depending on the state of a very large number (tending to infinity) of other processors as the lattice spacing tends to zero. On the other hand, for the cost term of the usual discretization, the contribu-

tion of a particular processor to the total cost depends on the state of the processor but is independent of the state of all other processors (regardless of the discretization level). Hence, if implemented in parallel architectures in the natural way, the two methods that possess the proper convergence properties require computations that are in some sense nonlocal as the lattice spacing tends to zero, while the usual discretization results in a local computation (independent of the discretization level) but fails to have the right convergence properties. Note that for the method using piecewise linear approximations, if implemented in the natural way, the computation can be done locally but each processor requires an unbounded number of states in the continuum limit (to indicate whether an endpoint of the line segment is present at that processor and, if so, at which processor the other endpoint lies).

A natural question is whether the computational difficulties discussed above can be circumvented by a clever discrete approximation. This problem is discussed in Chapter 3. As we will see, the results in Chapter 3 suggest that for rectangular lattices the difficulties are not merely due to a poor choice of discrete approximations, but are inherent difficulties associated with any discrete approximation to measures of length. This result probably holds for many other regular lattices as well (e.g., hexagonal). However, interestingly, the problems with nonlocal computation can be avoided for appropriate random and deterministic tesselations.

Also, as alluded to above, the nonlocal computations can likely be avoided if the processors are allowed to have infinitely many states. For example, this could correspond to associating a direction (or local tangent) to each boundary element in addition to just it presence or absence. Hence, in the MRF formulations this might correspond to coupled intensity and boundary fields both of which are real valued. Some work which may be related to this idea is contained in [93]. A somewhat different approach to having real valued boundary elements is suggested by an important result of Ambrosio [4, 5]. He obtained an interesting Γ-convergent approximation to the original variational problem. Specifically, he showed that the functional

$$E^h(f,v) = c_1 \iint_\Omega (f-g)^2 + c_2 \iint_\Omega (1-v^2)^h \parallel \nabla f \parallel^2 + c_3 \left(\iint_\Omega (1-v^2)^h \parallel \nabla v \parallel^2 + \frac{h^2 v^2}{16} \right)$$

Γ-converges to 2.1 as $h \to \infty$, so that minimizers of $E^h(f,v)$ converge to a minimizer of $E(f,\Gamma)$ as $h \to \infty$. Here f is as before and $v : \Omega \to [0,1]$ plays the role of the boundaries. For finite h, v represents a sort of smoothed version of Γ in the sense of having a value close to 1 near Γ and having a value of 0 away from Γ,

and varying continuously in between. This result suggests a natural digitization of 2.1 by taking a finite difference approximation to $E^h(f, v)$ as discussed in [97] and [45]. However, as far as we know, a proof of convergence for such finite difference approximations is lacking in this case. We expect that convergence should hold as long as $h \to \infty$ appropriately as the lattice spacing $1/n \to 0$, namely $h/n \to 0$. Such a conjecture is natural in light of the results of [69, 73] and was in fact stated in [97]. Furthermore, convergence issues aside, it is not clear that computational difficulties are avoided with these approaches. E.g., in Ambrosio's approximation, there may be some computational or numerical problems as $h \to \infty$. Further work needs to be done to understand whether any computational difficulties arise in this case.

Chapter 3

Local Versus Non-local Computation of Length From Discrete Approximations

3.1 Introduction

In Chapter 2, we discussed some discretizations of the segmentation problem for which we showed desirable convergence results in the continuum limit. These formulations have the advantage that solutions to the discrete versions converge to a solution of the continuous problem as the lattice spacing tends to zero. However, they have the disadvantage that for discrete boundaries the cost functional is considerably more difficult to evaluate than for the standard discretization. This has important implications as to the suitability of these methods for computation on parallel architectures.

As we discussed, the discrete formulations have a close relationship to problems arising from a probabilistic approach using Markov random fields (MRF's), which is attractive for a number of theoretical and practical reasons. One major reason for the attractiveness of MRF's is their local neighborhood structure. A very useful property of the standard discretization is its small neighborhood structure independent of the level of discretization. On the other hand, the discretizations discussed in the previous chapter require the neighborhood size to grow (unboundedly) as the lattice spacing tends to zero. For very fine discretizations, the neighborhood structure is highly nonlocal and the advantages of the MRF structure are essentially lost. This is due to the choice of the cost for the discrete boundaries, which were selected

for their convergence properties to the true length of the curves in the continuum limit. A natural question is whether one can retain such convergence properties with computations using local neighborhood structures.

Thus, we are led to consider the local versus non-local nature of computing the length of a curve from discrete approximations. Interestingly, this problem has connections with the learning paradigms discussed below (see Chapter 7). In this chapter, we formalize our notions of local and non-local computation, and provide some results for various tesselations.

3.2 Definition of Local Computation

In studying the question of local versus non-local computation of length, we restrict ourselves to a particular type of discrete representation for curves. First, for simplicity, we consider only curves contained in the unit square. For each n, we assume that the unit square is partitioned into a number of regions $s_{n,1}, ..., s_{n,k(n)}$ with $k(n) \to \infty$ as $n \to \infty$. We think of n as indicating the discretization level, so that $n \to \infty$ typically corresponds to finer and finer partitions. The discrete representation of a curve Γ on the partition of level n will consist of those regions $s_{n,i}$ which Γ passes through, i.e., for which $\Gamma \cap s_{n,i} \neq \emptyset$.

The notions of local versus global computation we use are similar to those considered by Minsky and Papert [84] (see also [1, 75]). For a lattice at level n, we imagine a processor in each region $s_{n,1}, \ldots, s_{n,k(n)}$. Let $p_{n,j}$ denote the location (x, y coordinates) of the processor in region $s_{n,j}$. For simplicity, we will also let $p_{n,j}$ refer to the processor itself (as well as its position).

We assume that each processor has information as to whether or not the curve Γ passes through its region. Furthermore, each processor $p_{n,j}$ has an associated neighborhood $N_{n,j}$ which is a set of other processors at level n which provide information to the processor $p_{n,j}$. That is, processor $p_{n,j}$ performs a computation depending only on the state of the pixels in its neighborhood, which will be denoted by $\Gamma|_{N_{n,j}}$. We assume that the outputs of each of the processors are combined linearly to produce the final computed value. Hence, the computed value $\hat{L}_n(\Gamma)$ for the length of the curve Γ from its discretization at level n is given by an expression of the form

$$\hat{L}_n(\Gamma) = \sum_{j=1}^{k(n)} \phi_{n,j}(\Gamma|_{N_{n,j}}) \tag{3.1}$$

The diameter of a neighborhood $N_{n,j}$ is the maximum distance between any two processors in $N_{n,j}$. The diameter d_n of the computation at discretization level n is the largest diameter over all neighborhoods $N_{n,j}$. Note that as $n \to \infty$, the processors necessarily get closer together, since they are all within the unit square. Since we are interested in computations in which each processor does not communicate to too many other processors, it is not sufficient simply to bound d_n as $n \to \infty$. Instead, we will bound the scaled diameter $\sqrt{k(n)}d_n$. A computation of the above form is said to be diameter limited (in the limit, or as the lattice spacing tends to zero) if $\sqrt{k(n)}d_n$ is uniformly bounded as a function of n, i.e., for some $d < \infty$ we have $\sqrt{k(n)}d_n \leq d$ for all n.

A diameter limited computation provides one notion of what we mean by a local computation. Following [84], another notion is that of an order limited computation. The order of a neighborhood $N_{n,j}$ is simply the number of processors in $N_{n,j}$, i.e. its cardinality. The order α_n of a computation at discretization level n is the maximum order over all j of $N_{n,j}$. Then, an order limited computation is one for which α_n is uniformly bounded as a function of n, so that there is some $\alpha < \infty$ such that $\alpha_n < \alpha$ for all n.

It is difficult to prove any results without imposing some additional structure on the computation. We consider the case of a translation invariant neighborhood structure and translation invariant processors meaning that for each n, $N_{n,j}$ and $\phi_{n,j}$ are independent of j. We also consider the case in which only those processors which are "on" can contribute to the computation. That is, we assume that the contribution $\phi_{n,j}(\Gamma|_{N_{n,j}})$ of processor $p_{n,j}$ is zero if Γ does not pass through the region associated with $p_{n,j}$. In the case of regular tesselations, these assumptions allow a simplification of the form of the computation in (3.1). Specifically, for an order or diameter limited computation there are a finite number K of distinct patterns for $\Gamma_{|N_{n,j}}$ (i.e., states of pixels in a neighborhood). Each processor which sees pattern i in its neighborhood contributes the same quantity $a_{n,i}$ to the total computation. Therefore, if we let t_n denote the total number of pixels which Γ passes through, and let $f_{n,i}(\Gamma)$ denote the frequency of occurrence of pattern i, then $\hat{L}_n$ is given by

$$\hat{L}_n(\Gamma) = t_n(\Gamma) \sum_{i=1}^{K} a_{n,i} f_{n,i}(\Gamma) \tag{3.2}$$

3.3 Local/Non-local Results for Various Tesselations

A rectangular digitization is the one most commonly used in image processing. In this case, at discretization level n the unit square is partitioned along the coordinate axes into n^2 square pixels of size $1/n$ by $1/n$. The pixels correspond to the closed lattice squares of $\frac{1}{n}\mathbf{Z}^2$. This is exactly the discretization defined in Section 2.4.1. The discrete version of a curve Γ is composed of the union of closed lattice squares of $\frac{1}{n}\mathbf{Z}^2$ which Γ passes through.

For such discretizations, we have the following result.

Theorem 3.1 *The length of a curve cannot be computed using a diameter limited computation from discrete approximations on a rectangular tesselation. In particular, if* $nd_n \leq d < \infty$ *then for some straight line* Γ, $\lim_{n\to\infty} \hat{L}_n(\Gamma) \neq L(\Gamma)$.

Proof: We will proceed by showing that any diameter limited computation fails to compute length appropriately in the limit on many straight lines. Consider a line segment of unit length and let θ be the angle that the extension of the segment makes with the x-axis. Since we are considering only the case where Γ is such a straight line segment, we will write (3.2) indexed by θ:

$$\hat{L}_n(\theta) = t_n(\theta) \sum_{i=1}^{K} a_{n,i} f_{n,i}(\theta) \tag{3.3}$$

For lines with small θ, as n gets large the digitization consists of long stretches of pixels in a row with occasional corners (or shifts) to different rows (see Figure 3-1). Now, suppose the computation is diameter limited with $nd_n \leq d$. Then for $0 \leq \theta < \tan^{-1}\frac{1}{d}$ the corners are sufficiently far apart so that the digitized pattern in the neighborhood of every processor contains either no corners or exactly one corner (see Figure 3-2), with the different locations of the corner in the neighborhood corresponding to different patterns.

Since we are concerned with the behavior of the computation as $n \to \infty$, we can ignore the effects at the ends of the line segment, and the effects of the offset of the line segment with respect to the digitization. For each $0 \leq \theta < \tan^{-1}\frac{1}{d}$, the frequencies of occurrence of all patterns which contain a corner are approximately the same for

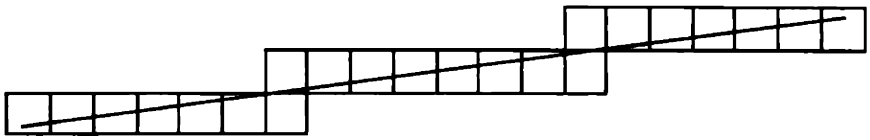

Figure 3-1: Digitized line segment with small slope.

large n. Hence, we can simplify (3.3) to

$$\hat{L}_n(\theta) = t_n(\theta)[a_{n,1} f_{n,1}(\theta) + a_{n,2} f_{n,2}(\theta)] \tag{3.4}$$

where $f_{n,1}$ and $f_{n,2}$ denote the frequency of occurrence of patterns without a corner and with a corner respectively.

For large n, there are approximately $n \sin\theta$ corners for a unit length segment at angle θ, and the segment passes through approximately $n \cos\theta$ columns. Hence, the digitized version of the line segment contains approximately $t_n(\theta) = n\cos\theta + n\sin\theta$ total pixels. Also, since the computation is diameter limited with $nd_n \leq d$, at most d pixels see a given corner, so that $nd\sin\theta$ pixels see some corner. Therefore, $f_{n,2}(\theta) = d\sin\theta/(\cos\theta + \sin\theta)$ and $f_{n,1}(\theta) = 1 - f_{n,2}(\theta)$. Substituting these expressions in (3.4) and letting $n \to \infty$ gives

$$\hat{L}(\theta) = a_1 \cos\theta + (a_1 + d(a_2 - a_1))\sin\theta \tag{3.5}$$

for $0 \leq \theta < \tan^{-1}\frac{1}{d}$. Since the line segments at all angles have unit length, in the interval $0 \leq \theta < \tan^{-1}\frac{1}{d}$, the computation is correct only for those θ for which $\hat{L}(\theta) = 1$. For finite d, from (3.5) $\hat{L}(\theta) = 1$ clearly cannot be satisfied for all $0 \leq \theta < \tan^{-1}\frac{1}{d}$ (in fact, it can be exactly satisfied for at most two values of θ in the desired interval).

□

We expect that similar results are true for the other standard regular tesselations (i.e., hexagonal and triangular). However, it is interesting that it is not true for all tesselations.

Specifically, we first consider random tesselations produced by a number of random straight lines. The lines will be drawn from the usual invariant measure which is uniform in both the angle and radial coordinate (up to some maximum radius) of the

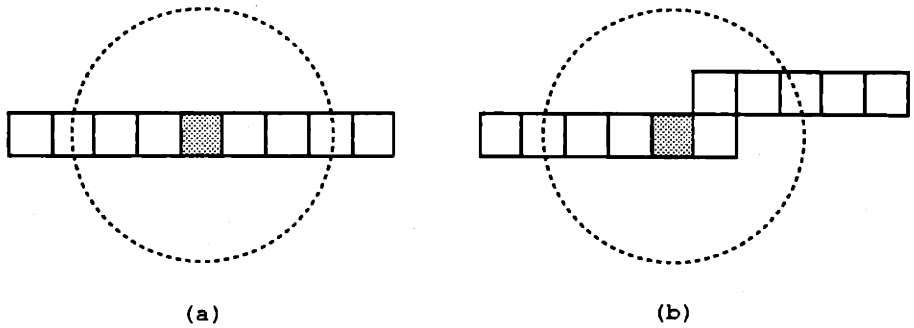

Figure 3-2: Two types of patterns in diameter limited neighborhood for line segments with sufficiently small slope.

polar coordinate representation of the line. The "lattice spacing" is related to the number of random lines drawn. For this tesselation we have the following result for computing the length of any straight line segment.

Theorem 3.2 *Let Γ be any straight line segment. There is a diameter limited computation on tesselations formed from independent lines drawn uniformly which converges to the length of Γ with probability one.*

Proof: For the computation on the tesselation at level n, we will let each processor which is "on" (i.e., which Γ passes through) contribute $2/n$ independent of the state of all other processors. Hence, the neighborhood of a given processor consists only of the processor itself, so that the computation is clearly diameter limited.

Now we need to show that as $n \to \infty$ this computation recovers the length of a straight line segment Γ. Let $\beta(n, \Gamma)$ denote the number of pixels comprising the digitized version of Γ on the lattice at level n. For a random line ℓ, let $m(\ell, \Gamma)$ denote the number of intersections between ℓ and Γ. Since Γ is a straight line segment, for almost all ℓ, $m(\ell, \Gamma)$ is either zero or one. Furthermore, since the pixels consist of regions formed by straight lines and Γ is a straight line segment, the number of pixels comprising the digitized version of Γ is just one plus the number of straight lines intersecting Γ. I.e., $\beta(n, \Gamma) = \sum_{i=1}^{n} m(\ell_i, \Gamma)$ so that

$$\hat{L}_n = 2 \cdot \frac{1}{n} \sum_{i=1}^{n} m(\ell_i, \Gamma)$$

From the law of large numbers, we have that as $n \to \infty$, $\hat{L}_n \to 2 \cdot Em(\ell, \Gamma)$ with probability one. A result from stochastic geometry states that if Γ is any rectifiable curve in the unit square and ℓ is a random line intersecting the unit square drawn uniformly then $Em(\ell, \Gamma) = \frac{1}{2}L(\Gamma)$ (e.g., see [105] and Section 7.2 for more discussion). Thus, using this result we have that $\hat{L}_n \to L(\Gamma)$ with probability one.

□

The intuitive idea of the results above is that the length of a line segment is twice the area of a corresponding subset of the cylinder, namely the area of the set of lines which intersect the line segment. With a rectangular tesselation, we obtain samples on the cylinder only for $\theta = 0$ and $\theta = \pi/2$. On the other hand, with the random tesselation, we obtain a random sampling of points on the cylinder from which we can easily estimate the desired area. This suggests that there is a tradeoff between the complexity of the sampling used and the complexity of the resulting computation. It also suggests that appropriate deterministic sampling strategies should allow a local computation of the length of a line segment. The theorem below shows that this is in fact the case for a particular deterministic sampling strategy. Specifically, we consider the tesselation at level n to be that formed by parallel lines with spacing $1/n$ taken at angles $2\pi j/n$ for $j = 0, \ldots, n-1$. Denote this tesselation by U_n. Moran [87] has obtained results on estimating the length of a curve by counting intersections with the straight lines forming the tesselation U_n. The following result on the local computation of length using the tesselation U_n follows from Moran's results.

Theorem 3.3 *Let Γ be any straight line segment. There is a local computation on the deterministic tesselation U_n which converges to the length of Γ.*

Proof: As in Theorem 3.2, for the computation on the tesselation at level n, we will let each processor which is "on" (i.e., which Γ passes through) contribute $2/n$ independent of the state of all other processors. Hence, the neighborhood of a given processor consists only of the processor itself, so that the computation is clearly diameter limited. To show that as $n \to \infty$ this computation recovers the length of a straight line segment Γ, note that (as in Theorem 3.2) since the pixels consist of regions formed by straight lines and Γ is a straight line segment, the number of pixels comprising the digitized version of Γ is just one plus the number of straight lines intersecting Γ. Hence, the theorem follows using the results of Moran [87].

□

3.4 Discussion and Open Problems

There are a number of interesting questions and directions to pursue along the lines of this chapter. It would be interesting to extend the local computation results for both the random and deterministic tesselations (Theorems 3.2 and 3.3) to general curves. The difficulty is that for general curves there isn't a simple correspondence between the number of regions intersected by the curve and the number of intersections the lines make with the curve. It would also be interesting to extend the local and non-local results to other tesselations. For example, we conjecture that non-local results similar to Theorem 3.1 hold for regular tesselations such as triangular or hexagonal. Likewise, we conjecture that results similar to Theorem 3.2 hold for other random tesselations such as Voronoi tesselations obtained from homogeneous planar Poisson point processes. (For work on random tesselations see for example [3, 114]). The results of [112] may be useful in proving results of this type. One difficulty in dealing with tesselations which are not formed by a number of straight lines is that the duality between intersections of the lines and sampling on the cylinder (manifold of straight lines) is lost. Perhaps there is a more general way in which to view the sampling which works for other tesselations.

Another direction to pursue is to try to relax some of the assumptions such as translation invariance, etc. However, it seems that proving results in these cases will be difficult. One extension that we feel should go through is to prove a non-local result like Theorem 3.1 for order limited computations as opposed to just diameter limited. Also, it would be useful to obtain error bounds in terms of the diameter (or order) of the computation, since it is likely that although an exact computation in the limit may be non-local, a good approximation can be obtained with a small diameter (order). It should be possible to use 3.5) to obtain lower bounds on the achievable error for diameter limited computations.

A natural question is whether the tesselations which allow local computation of length can be used to construct a discrete version of the segmentation problem which is local and yet possesses the appropriate convergence properties in the continuum limit. We expect that this can be done, although the corresponding MRF structure would be somewhat complicated due to the irregular placements of lattice sites.

Our results show that local lattice systems may inherently lack the ability to perform certain computations due to the arrangement and connections of the lattice sites. The results suggest that other lattice-type systems such as cellular automata or

spin systems in statistical mechanics may possess inherent computational limitations arising from their architecture. It may be interesting to study these questions.

The notion of local versus non-local computations appears to be of fundamental importance. The work presented in this chapter suggests many other general directions which may be interesting to pursue. It may be worthwhile investigating whether other computations (e.g., determining convexity or connectedness from discrete approximations of a set) can be done locally. One could consider questions of local/non-local computations using other discretizations or in which the processors have access to other types of data, as opposed to just data from a discretization on a tesselation as considered here. It might also be interesting to consider forms of computation other than just those of the type in 3.1, as well as to investigate other notions of local and non-local computations.

Chapter 4

Computational Limitations of Model Based Recognition

The problem of model based recognition can be informally described in the following way: given a library of modeled objects and a set of sensed data, identify and locate the objects from the library that are present in the data. The standard computational approach is to represent the modeled objects and the data in terms of discrete features so that the recognition can be solved as a search problem. These results indicate that by applying rigidity constraints in various ways, model based recognition can be efficiently applied to recognize a small number of objects even from partial views and in the presence of non-malicious noise. The relevant complexity parameter in such cases is the number of features that model each object.

In this chapter we analyze the case in which objects are represented by a small number of features. The relevant complexity parameter in this case is the number of objects. Instead of analyzing the performance of specific algorithms, our approach is to apply techniques from complexity theory to identify cases in which model based recognition appears to be inherently difficult. Specifically, we show that the problem is NP-complete, and thus, its complexity (modulo standard complexity assumptions, i.e., P $\neq$ NP) is super-polynomial in the size of the library.

Proving that a problem is NP-complete is a common technique in complexity analysis for identifying the problem as intrinsically difficult. In a (well defined) sense, an NP-complete problem is the most difficult problem in the class NP, which includes many difficult problems such as the traveling salesman. However, an NP-complete problem is not completely unapproachable; a standard method for coping with such problems is to identify easily solved sub-problems. In the case of model based recog-

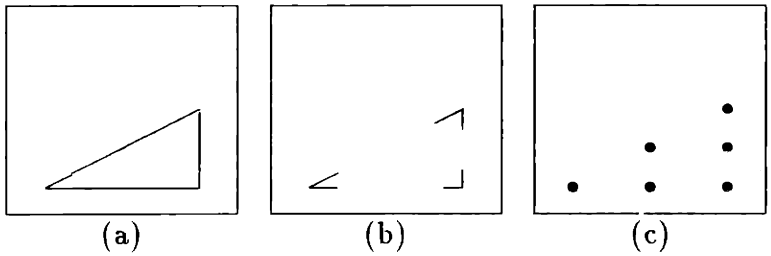

Figure 4-1: Examples of local features.

nition this might correspond to exploiting additional structure of the modeled objects and the way they are viewed. For more information on the theory of NP-complete problems see [44]. For applications of NP-completeness results to vision tasks see [64, 117].

The negative results of this chapter can be used to determine constraints that may simplify the problem of model based recognition. We will attempt to identify three types of constraints: constraints that leave the problem NP-complete, constraints that guarantee efficient (polynomial) algorithms, and constraints that make our NP-completeness proofs inapplicable, so that they may simplify the problem. The generic model based recognition problem that we consider is noise free and assumes no occlusion. An example of constraints of the first type is that each pair of local features can be found in at most three objects from the library. An example of constraints of the second type is that each pair of local features can be found in at most two objects from the library. An example of constraints of the third type is occlusion of convex objects.

4.1 Preliminary Definitions

We consider situations in which objects can be described in terms of sets of local features. A local feature is taken to be a simple geometric shape (although the results hold for arbitrary interpretations of "simple" and "local"), and an object is described by a set of local features and their location in space. Commonly used features are points, lines, angles, etc. An example is shown in Figure 4-1, where a triangle is described in terms of straight lines (a), corners (b), and points along its edges (c).

Definition: An object description by local features is a set of pairs

$$O = \{\langle f_1, X_1\rangle, \langle f_2, X_2\rangle, \ldots \langle f_t, X_t\rangle\}$$

where for $1 \leq i \leq t$, f_i is a local feature and X_i is its location is space relative to a fixed coordinate system.

Definition: A library is a set of object descriptions.

Definition: A picture is sensed data given as a set of local features and their location is space.

The problem of model based object recognition is:

> For a family of coordinate transformations Ψ, a library L, and a picture $P = \{\langle f_1, X_1\rangle, \ldots, \langle f_m, X_m\rangle\}$, determine a disjoint partition of P into objects from L, i.e., subsets $O_1, \ldots, O_q$ such that: (i) for $i \neq j$ $O_i \cap O_j = \emptyset$; (ii) $P = O_1 \cup \cdots \cup O_q$; (iii) for $1 \leq i \leq q$ there is $\psi_i \in \Psi$ that transforms an object from L into O_i.

Our main result is that the problem of model based recognition under translations, rotations, and perspective projections is NP-complete. The proofs are based on a reduction from *exact cover by 3 sets (X3C)* that is known to be NP-complete (see [44] page 221).

X3C: The following *exact cover by 3-sets* problem is NP-complete:
Instance: a set E of m elements and a collection C of 3-element subsets of E.
Question: does C contain an exact cover for E, i.e., a subcollection $C' \subset C$ such that every element of E occurs in exactly one member of C'?
Comment: X3C remains NP-complete even if no element occurs in more than three subsets in C, but is solvable in polynomial time if no element occurs in more than two subsets. A related problem, *exact cover by 2-sets*, is solvable in polynomial time.

4.2 The Case of Translation and Rotation

Theorem 4.1 *Let L be a library of objects and let P be a picture. The decision problem of whether P can be described as a disjoint union of translated and rotated objects from L is NP-complete. The problem remains NP-complete even if each object is described by 3 points.*

$$p_1 \quad \leftarrow \quad m^2+1 \quad \rightarrow \quad p_2 \quad \cdots \quad p_i \quad \leftarrow \quad m^2+i \quad \rightarrow \quad p_{i+1} \quad \cdots \quad p_m$$

Figure 4-2: The picture in the proof of Theorem 4.1.

$$p_2 \quad \leftarrow \quad 2m^2+5 \quad \rightarrow \quad p_4 \quad \leftarrow \quad m^2+4 \quad \rightarrow \quad p_5$$

Figure 4-3: A typical object in the proof of Theorem 4.1.

Proof: Membership in NP is obvious. To show that the problem is NP-complete we reduce X3C to it.

Let $\{E, C\}$ be an instance of the X3C problem. C is a collection of 3-element subsets of the m elements $e_1, \ldots, e_m \in E$. We begin by constructing a picture P of m points $p_1, \ldots, p_m$ on the x axis. The location of p_1 is at the origin, the point p_2 is at distance m^2+1 from p_1, the point p_3 is at distance m^2+2 from p_2, etc. See the illustration in Figure 4-2. Let $\phi : E \rightarrow P$ denote the mapping of elements in E to points in P. For $1 \leq i \leq m$ we have:

$$\phi(e_i) = \text{a point at } x = (i-1)m^2 + i(i-1)/2 \tag{4.1}$$

Clearly, ϕ is 1-1 and onto, so that the inverse mapping is well defined. We now create the library L from the 3-element subsets in C. For a 3-set composed of the elements e_α, e_β, e_γ we add to L an object described by the 3 points $\phi(e_\alpha)$, $\phi(e_\beta)$, $\phi(e_\gamma)$. The object generated by the elements e_2, e_4, e_5 is shown in Figure 4-3.

To prove the NP-complete result it remains to show that P is a disjoint union of rotated and translated objects from L if and only if C contains an exact cover of E. The proof is based on Lemma 4.1 which is proved at the end of this section.

Let $C' \subset C$ be an exact cover of E, where $q = m/3 = |C'|$. For $\{e_{i_1}, e_{i_2}, e_{i_3}\} \in C'$ define $O_i = \{\phi(e_{i_1}), \phi(e_{i_2}), \phi(e_{i_3})\}$, so that $O_i \in L$ for $1 \leq i \leq q$. Since C' is a cover of E and ϕ is onto, $P = \bigcup_{i=1}^{q} O_i$. Since C' is exact and ϕ is 1-1, $O_i \cap O_j = \emptyset$ for $i \neq j$.

Conversely, let Ψ be the family of coordinate translations and rotations, and assume $O_i \in L$, $\psi_i \in \Psi$ for $1 \leq i \leq q$ such that: (i) for $i \neq j$ $\psi_i(O_i) \cap \psi_j(O_j) = \emptyset$; (ii) $P = \bigcup_{i=1}^{q} \psi_i(O_i)$. From Lemma 4.1 it follows that ψ_i is the identity transformation ($\psi_i(O_i) = O_i$), so that $\psi_i(O_i) \in L$ for $1 \leq i \leq q$. Let $O_i = \{p_{i_1}, p_{i_2}, p_{i_3}\}$. Define $T_i = (\phi^{-1}(p_{i_1}), \phi^{-1}(p_{i_2}), \phi^{-1}(p_{i_3}))$, and $C' = \{T_i : 1 \leq i \leq q\}$. From (ii) and the fact that ϕ^{-1} is onto it follows that C' is a cover. From (i) and the fact that ϕ^{-1} is 1-1 it

follows that C' is an exact cover.

□

Lemma 4.1 *Let O be an object from the library defined in the proof of Theorem 4.1, and let O' be an object defined by 3 points from the picture in the proof of Theorem 4.1. If O can be mapped by translation and rotation to O' then $O = O'$.*

Proof: Without loss of generality let O be described by the points p_{i_1}, p_{i_2}, p_{i_3} and O' by the points p_{j_1}, p_{j_2}, p_{j_3}, where $i_1 < i_2 < i_3$ and $j_1 < j_2 < j_3$. Since the objects are 1-dimensional, a transformation taking O to O' involves either zero rotation or a 180° rotation. We show that the transformation must be with zero rotation and zero translation.

First, suppose the transformation involves no rotation, then the distance between p_{i_1} and p_{i_2} is the same as the distance between p_{j_1} and p_{j_2}. From Equation (4.1) we have

$$(j_2 - j_1)m^2 + \frac{j_2(j_2-1) - j_1(j_1-1)}{2} = (i_2 - i_1)m^2 + \frac{i_2(i_2-1) - i_1(i_1-1)}{2}$$

Let $s(i,j) = (j(j-1) - i(i-1))/2$, so that the above equation can be written as

$$[(j_2 - j_1) - (i_2 - i_1)]m^2 = s(i_1, i_2) - s(j_1, j_2). \tag{4.2}$$

Clearly, $0 < s(i,j) < m^2$ for $1 \leq i < j \leq m$, and $|s(i_1,i_2) - s(j_1,j_2)| < m^2$. But since the right hand side of Equation (4.2) is divisible by m^2 it must equal 0, and we have

$$\begin{aligned} s(i_1, i_2) &= s(j_2, j_2) \\ j_2 - j_1 &= i_2 - i_1 \end{aligned} \tag{4.3}$$

The unique solution to the system (4.3) with j_1, j_2 as the unknowns is $j_1 = i_1$ and $j_2 = i_2$. Since in pure translation the distance between p_{i_1} and p_{i_3} is the same as the the distance between p_{j_1} and p_{j_3} the same derivation gives $j_3 = i_3$, so that $O = O'$.

It remains to show that a transformation taking O to O' cannot involve rotation. Suppose, on the contrary, that O is mapped to O' by a transformation involving nonzero rotation. As mentioned above, this rotation must be 180°. But then the distance between p_{i_1} and p_{i_2} is the same as the distance between p_{j_3} and p_{j_2}, and the distance between p_{i_2} and p_{i_3} is the same as the distance between p_{j_2} and p_{j_1}. Using the same derivation as above we get $j_1 = i_3$, $j_2 = i_2$, and $j_3 = i_1$. But since $j_1 < j_3$ and $i_1 < i_3$ we have a contradiction.

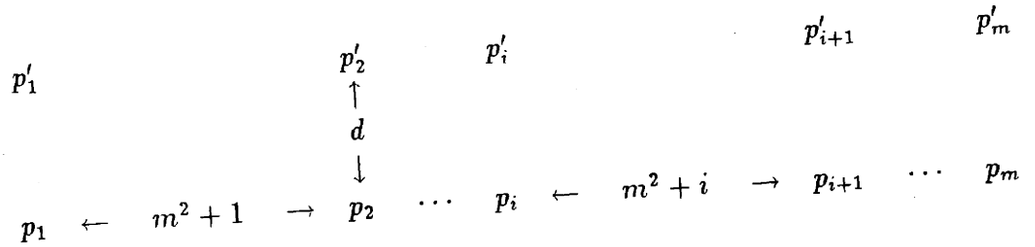

Figure 4-4: The picture in the proof of Theorem 4.2.

□

4.3 Translation, Rotation, and Scaling

Theorem 4.2 *Let L be a library of objects and let P be a picture. The decision problem of whether P can be described as a disjoint union of translated, rotated, and scaled objects from L is NP-complete. The problem remains NP-complete even if each object is described by 6 points.*

Proof: Membership in NP is obvious. To show that the problem is NP-complete we reduce X3C to it.

Let $\{E, C\}$ be an instance of the X3C problem. We begin by constructing a 2D picture Q as a disjoint union of two pictures: $Q = P \cup P'$. The pictures are defined by the two 1-1 and onto mappings: $\phi : E \to P$ and $\theta : E \to P'$.

$$\begin{aligned} \phi(e_i) &= \text{a point at } x = (i-1)m^2 + i(i-1)/2, \quad y = 0 \\ \theta(e_i) &= \text{a point at } x = (i-1)m^2 + i(i-1)/2, \quad y = d \end{aligned} \tag{4.4}$$

See the illustration in Figure 4-4. We now create the library L from the 3-element subsets of C. For $(e_\alpha, e_\beta, e_\gamma)$ we add to L an object described by the 6 points: $\theta(e_\alpha)$, $\theta(e_\beta)$, $\theta(e_\gamma)$, $\phi(e_\alpha)$, $\phi(e_\beta)$, $\phi(e_\gamma)$. The object generated by the elements e_2, e_4, e_5 is shown in Figure 4-5.

To complete the proof it remains to show that Q is a disjoint union of translated, rotated, and scaled objects from L if and only if C contains an exact cover of E. The proof is based on Lemma 4.2 which will be proved at the end of this section.

Let $C' \subset C$ be an exact cover of E, with $q = |C'|$. For $\{e_{i_1}, e_{i_2}, e_{i_3}\} \in C'$ define $O_i = \{\theta(e_{i_1}), \theta(e_{i_2}), \theta(e_{i_3}), \phi(e_{i_1}), \phi(e_{i_2}), \phi(e_{i_3})\}$, so that $O_i \in L$ for $1 \le i \le q$. Since C' is a cover of E, and ϕ, θ are onto P and P' respectively, $Q = P \cup P' = \bigcup_{i=1}^{q} O_i$. Since C' is exact and ϕ, θ are 1-1, $O_i \cap O_j = \emptyset$ for $i \neq j$.

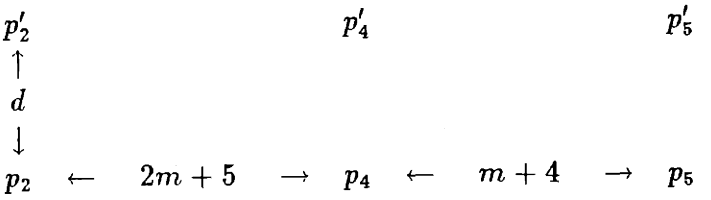

Figure 4-5: A typical object in the proof of Theorem 4.2.

Conversely, let Ψ be the family of coordinate translations rotations and scaling and assume $O_i \in L$, $\psi_i \in \Psi$ for $1 \leq i \leq q$ such that: (i) for $i \neq j$, $\psi_i(O_i) \cap \psi_j(O_j) = \emptyset$; (ii) $Q = \bigcup_{i=1}^{q} \psi_i(O_i)$. From Lemma 4.2 it follows that ψ_i is the identity transformation, so that $\psi_i(O_i) \in L$ for $1 \leq i \leq q$. Let $O_i = \{p_{i_1}, p_{i_2}, p_{i_3}, p'_{i_1}, p'_{i_2}, p'_{i_3}\}$, where we assume without loss of generality that $p_{i_1}, p_{i_2}, p_{i_3}$ have zero y coordinates. Define $T_i = \{\phi^{-1}(p_{i_1}), \phi^{-1}(p_{i_2}), \phi^{-1}(p_{i_3})\}$, and $C' = \{T_i : 1 \leq i \leq q\}$. From (ii) and the fact that ϕ^{-1} is onto E it follows that C' is a cover. From (i) and the fact that ϕ^{-1} is 1-1 it follows that C' is an exact cover.

□

Lemma 4.2 *Let O be an object from the library defined in the proof of Theorem 4.2, and let O' be an object defined by 6 points from the picture in the proof of Theorem 4.2. If O can be mapped by translation, rotation, and scaling to O' then $O = O'$.*

Proof: Let O be generated by $e_{i_1}, e_{i_2}, e_{i_3}$. Let u_1, u_2, u_3, be the points of O' that are mapped to $\theta(e_{i_1}), \theta(e_{i_2}), \theta(e_{i_3})$ respectively, then u_1, u_2, u_3 are collinear. Similarly, let v_1, v_2, v_3, be the points of O' that are mapped to $\phi(e_{i_1}), \phi(e_{i_2}), \phi(e_{i_3})$ respectively, then v_1, v_2, v_3 are collinear. Since $\theta(e_{i_1})$, $\theta(e_{i_2})$, $\phi(e_{i_1})$ form a right triangle, u_1, u_2, v_1 form a right triangle, so that the triplets u_1, u_2, u_3, and v_1, v_2, v_3 are not on the same line in the picture. Therefore, it must be that one triplet lies on the line $y = 0$, and the other on the line $y = d$, and since the distance between the lines in the library object is d, the transformation involves no scaling.

It remains to show that the transformation involves no translation and rotation and this follows from Lemma 4.1 when applied to the points u_1, u_2, u_3 and the library of objects defined by the triplets of points $\{\theta(e_{i_1}), \theta(e_{i_2}), \theta(e_{i_3})\}$ for $1 \leq i \leq q$.

□

4.4 The Case of Perspective Projection

A perspective projection is the mapping $\pi : \mathcal{R}^3 \rightarrow \mathcal{R}^2$ given by

$$x = f\frac{X}{Z} \quad ; \quad y = f\frac{Y}{Z} \tag{4.5}$$

Here it is assumed that the camera is at the origin and pointed directly down the Z axis. The reference frame is oriented as the image plane, which is located at distance f from the origin (see [53]). Unlike translation, rotation, and scaling, perspective projection may destroy geometric properties by merging lines and points. In the extreme case, any object far enough from the image plane is projected into a single point in a finite resolution picture. To eliminate degenerate cases we consider only *stable* perspective projections.

Definition: A stable perspective projection has the following properties: (i) Distinct 3D feature points are mapped into distinct 2D feature points. (ii) Non-collinear 3D feature points are mapped into non-collinear 2D feature points.

Notice that a small perturbation of the viewing point of an unstable perspective projection always gives a stable perspective projection.

Theorem 4.3 *Let L be a library of 3D objects, and let P be a 2D picture given as a set of local features and their 2D location. The decision problem of whether P can be described as a stable perspective projection of a disjoint union of translated and rotated objects from L is NP-complete. The problem remains NP-complete even if each object is described by 12 points.*

Proof: Membership in NP is obvious. To show that the problem is NP-complete we reduce X3C to it.

Let $\{E, C\}$ be an instance of the X3C problem. We begin by constructing the 2D picture $Q = P_1 \cup P_2 \cup P_3 \cup P_4$, where

$$P_j = \{\phi_j(e_i); 1 \leq i \leq m\} \text{ for } 1 \leq j \leq 4$$

$$\begin{aligned}
\phi_1(e_i) &= \text{a point at } x = (i-1)m^2 + i(i-1)/2, \quad y = 0 \\
\phi_2(e_i) &= \text{a point at } x = (i-1)m^2 + i(i-1)/2, \quad y = m^3 \\
\phi_3(e_i) &= \text{a point at } y = (i-1)m^2 + i(i-1)/2, \quad x = -1 \\
\phi_4(e_i) &= \text{a point at } y = (i-1)m^2 + i(i-1)/2, \quad x = m^3 + 1
\end{aligned}$$

Thus, the points are on the edges of a planar rectangle.

We now create the library L from the 3-element subsets of C. For $(e_{i_1}, e_{i_2}, e_{i_3})$ we add to L an object described by the twelve 3D points: $\{\phi_j^z(e_{i_t}); 1 \leq j \leq 4,\ 1 \leq t \leq 3\}$, where:

$$\begin{aligned}
\phi_1^z(e_i) &= \text{a point at } X = (i-1)m^2 + i(i-1)/2,\ Y = 0,\ \ Z = f \\
\phi_2^z(e_i) &= \text{a point at } X = (i-1)m^2 + i(i-1)/2,\ Y = m^3,\ Z = f \\
\phi_3^z(e_i) &= \text{a point at } Y = (i-1)m^2 + i(i-1)/2,\ X = -1,\ Z = f \\
\phi_4^z(e_i) &= \text{a point at } Y = (i-1)m^2 + i(i-1)/2,\ X = m^3 + 1,\ Z = f
\end{aligned}$$

Observe that $\pi(\phi_j^z(e_i)) = \phi_j(e_i)$ for $1 \leq j \leq 4$. It remains to show that Q is a stable perspective projection of a disjoint union of translated and rotated objects from L if and only if C contains an exact cover of E. The proof is based on Lemma 4.3 which will be proved at the end of this section.

Let $C' \subset C$ be an exact cover of E, with $q = |C'|$. For $\{e_{i_1}, e_{i_2}, e_{i_3}\} \in C'$ define O_i as the 3D object described by the twelve 3D points: $\{\phi_j^z(e_{i_t}); 1 \leq j \leq 4,\ 1 \leq t \leq 3\}$, so that $O_i \in L$ for $1 \leq i \leq q$. Since C' is a cover of E, and ϕ_j are onto P_j respectively, $Q = \bigcup_{i=1}^{q} \pi(O_i)$. Since C' is exact and ϕ_j are 1-1, $\pi(O_i) \cap \pi(O_j) = \emptyset$ for $i \neq j$.

Conversely, let Ψ be the family of coordinate translations and rotations and assume $O_i \in L$ and $\psi_i \in \Psi$ for $1 \leq i \leq q$, such that: (i) for $i \neq j$, $\pi(\psi_i(O_i)) \cap \pi(\psi_j(O_j)) = \emptyset$; (ii) $Q = \bigcup_{i=1}^{q} \pi(\psi_i(O_i))$. From (ii) and Lemma 4.3 it follows that ψ_i is the identity transformation, so that $\psi_i(O_i) \in L$ for $1 \leq i \leq q$. Let $O_i = \{p_{i_1}^j, p_{i_2}^j, p_{i_3}^j\}$ for $1 \leq j \leq 4$, where we assume without loss of generality that $p_{i_t}^j$ were generated by ϕ_j^z. Define $T_i = \{(\phi_j^z)^{-1}(p_{i_t}^j) : 1 \leq j \leq 4, 1 \leq t \leq 3\}$, and $C' = \{T_i : 1 \leq i \leq q\}$. From (ii) and the fact that $(\phi_j^z)^{-1}$ is onto E it follows that C' is a cover. From (i) and the fact that $(\phi_j^z)^{-1}$ is 1-1 it follows that C' is an exact cover.

□

Lemma 4.3 *Let O be a 3D object from the library defined in the proof of Theorem 4.3, and let O' be an object defined by 12 points from the picture in the proof of Theorem 4.3. If O can be mapped by translation rotation and stable perspective projection to O' then the mapping is with zero translation and rotation.*

Proof: We use the following properties of perspective projection (see [53], Chapter 13): (a) Collinear 3D points are projected into collinear 2D points. (b) If the projection of parallel 3D lines is parallel 2D lines then the 3D lines are parallel to the image plane.

Let O be generated by $e_{i_1}, e_{i_2}, e_{i_3}$. Let L_j be the 3D line of the rotated and translated points $\phi_j^z(e_{i_1})$, $\phi_j^z(e_{i_2})$, $\phi_j^z(e_{i_3})$ for $1 \leq j \leq 4$, so that L_1 is parallel to L_2 and L_3 is parallel to L_4. Let u_j^1, u_j^2, u_j^3 be the points of O' that are mapped to $\phi_j^z(e_{i_1})$, $\phi_j^z(e_{i_2})$, $\phi_j^z(e_{i_3})$ respectively, then u_j^1, u_j^2, u_j^3 are collinear for $1 \leq j \leq 4$, and since the projection is stable, the 4 triplets are on 4 different lines in the picture. The picture has exactly four lines with at least 3 points. These lines are: $y = 0$, $y = m^3$, $x = -1$, and $x = m^3 + 1$. Therefore, the 4 triplets come from these 4 lines.

Let l_j be the projection of L_j for $1 \leq j \leq 4$. l_1 intersects with two lines from $\{l_2, l_3, l_4\}$, and is parallel to the third. Since L_1 intersects with L_3 and L_4, l_1 intersects with l_3, l_4, and is parallel to l_2. Thus, we have two parallel lines L_1, L_2 that are projected into parallel lines. Therefore, both L_1 and L_2 must be parallel to the image plane; let Z_1 and Z_2 be their depth. From the same arguments the lines L_3, L_4 are parallel to the image plane; let Z_3, Z_4 be their depth respectively. But since L_3 intersects with both L_1 and L_2 we have $Z_1 = Z_2 = Z_3 = Z_4$.

We conclude that all the points of the translated and rotated object O have the same distance from the image plane. From Equation (4.5) it follows that in this case the distance from the image plane has the effect of scaling the object. Thus, Lemma 4.3 follows from Lemma 4.2 when applied to the library of objects defined by $\phi_j(e_{i_1})$, $\phi_j(e_{i_2})$, $\phi_j(e_{i_3})$ and the 6 points u_j^1, u_j^2, u_j^3 for $1 \leq j \leq 2$.

□

4.5 Discussion and Open Problems

We have shown that the problem of model based recognition is NP-complete. Thus, there is little hope for a performance guaranteed algorithm that can solve the problem efficiently. However, it is still possible that easy sub-classes of the problem can be characterized by additional structure of the modeled objects (e.g., convexity) and the way they are viewed (e.g., occlusion). Our results can help determine what constraints are potentially useful. For example, we can identify some constraints that may potentially simplify model based recognition, and other constraints that leave the problem NP-complete.

Local features other than a point: With no additional structure this can only make the problem more difficult. However, with additional structure of the local features the problem may become polynomial. For example, straight lines may have an additional constraint that their ends meet (see Figure 4-1).

Occlusion: Without additional structure this can only make the problem more difficult. However, with additional constraints such as convexity this makes our NP-completeness proofs inapplicable, so that it may potentially simplify the problem.
A small number of feature points: If each object is described by 2 points the problem is polynomially solvable by matching techniques.
A large number of feature points: Without additional structure this can only make the problem more difficult. However, if it is assumed that small subsets of these points determine a unique object from the library then the problem is polynomially solvable. (This is the essential assumption in geometric hashing [74]).
Almost distinct subsets: If the distance between every pair of feature points uniquely determines two (or less) objects, the problem is polynomially solvable. If this distance determines three (or more) objects the problem is still NP-complete. This follows from the comment in the definition of X3C.
Dimensionality: Notice that the results of Theorem 4.1 hold also for translation and rotation in 2 and 3 dimensions. Similarly, the results of Theorem 4.2 hold also for 3 dimensions.

It would be interesting to study further the complexity of model based recognition under different conditions such as those mentioned above. Some other interesting directions may be to consider the complexity of formulations in which we require only "approximate" solutions and/or to consider stochastic formulations of model based recognition (including noisy observations) in which one can consider, for example, the average case complexity.

Chapter 5

Metric Entropy, VC Dimension, and Learnability for a Class of Distributions

Recently, there has been a great deal of work on the Probably Approximately Correct (or PAC) learning model, which is a precise framework attempting to capture the notion of what we mean by 'learning from examples'. The essential idea consists of approximating an unknown 'concept' from a finite number of positive and negative 'examples' of the concept. For example, the concept might be some unknown geometric figure in the plane, and the positive and negative examples are points inside and outside the figure, respectively. The goal is to approximate the figure from a finite number of such points. The examples are assumed to be drawn according to some probability distribution, and the same distribution is used to evaluate how well a concept is learned. However, no assumptions are made about which particular distribution is used. That is, the learning is required to take place for every distribution.

The earliest work and many fundamental results related to PAC-like models was done by Vapnik [121]. Many important results relevant to the PAC model have been obtained in the probability and statistics literature [119, 120, 36, 94]. Recently, Valiant [118] independently proposed a similar model in the computer science community, and his work spawned a large amount of work analyzing and extending variations of the PAC model. More recently, Haussler [56] has formulated a very general framework refining and consolidating much of the previous work on the PAC model.

Blumer et al. [24], based on results of [119] gave a characterization of PAC learnability for the distribution-free framework in terms of a combinatorial parameter which measures the 'size' of a concept class. Benedek and Itai [17] studied a variation of the PAC model in which the examples are assumed to be drawn from a fixed and known distribution. They gave a characterization of learnability in this case in terms of a different measure of the size of a concept class.

In Section 5.1, we give some definitions, a precise description of the learning framework, and some previous results from [24] and [17]. The definitions and notation used are essentially those from [24], which are a slight variation from those originally given in [118]. The result of [24] states that a concept class is learnable for every distribution iff it has finite Vapnik-Chervonenkis (VC) dimension. An analogous result of [17] characterizes learnability for a fixed distribution. We point out that the characterization of [17] is identical to that of finite metric entropy, which has been studied in other contexts. The results characterizing learnability suggest that there may be relationships between the VC dimension of a concept class and its metric entropy with respect to various distributions. Some such relationships are known, and in Section 5.2 we summarize some known results of others and prove some new results. In Section 5.3 we consider learnability for a class of distributions, which is a natural extension of learnability for a fixed distribution. Benedek and Itai [17] posed the characterization of learnability in this case as an open problem. They conjectured that a concept class is learnable with respect to a class of distributions iff the metric entropy of the concept class with respect to each distribution is uniformly bounded over the class of distributions. We prove some partial results for this problem. Although the results we prove are far from verifying the conjecture in general, they are consistent with it. Furthermore, they provide some indication of conditions when power is gained by requiring learnability only for a class of distributions rather than for all distributions.

5.1 The PAC Learning Model and Previous Results Characterizing Learnability

In this section, we describe the formal model of learning introduced by Valiant [118] (learnability for all distributions) and a variant (learnability for a fixed distribution), and we state previous results characterizing learnability in these cases. The result of

Blumer et al. [24] characterizes learnability for all distributions in terms of a quantity known as the VC dimension. The result of Benedek and Itai [17] characterizes learnability for a fixed distribution in terms of a quantity known metric entropy.

Informally, Valiant's learning framework can be described as follows. The *learner* wishes to learn a concept unknown to him. The *teacher* provides the learner with random positive and negative examples of the concept drawn according to some probability distribution. From a finite set of examples, the learner outputs a hypothesis which is his current estimate of the concept. The error of the estimate is taken as the probability that the hypothesis will incorrectly classify the next randomly chosen example. The learner cannot be expected to exactly identify the concept since only a finite number of examples are seen. Also, since the examples are randomly chosen, there is some chance that the hypothesis will be very far off (due to poor examples). Hence, the learner is only required to closely approximate the concept with sufficiently high probability from some finite number of examples. Furthermore, the number of examples required for a given accuracy and confidence should be bounded independent of the distribution from which the examples are drawn. Below, we will describe this framework precisely, following closely the notation of [24].

Let X be a set which is assumed to be fixed and known. X is sometimes called the *instance space*. Typically, X is taken to be either $\mathbf{R}^n$ (especially $\mathbf{R}^2$) or the set of binary n-vectors. A *concept* will refer to a subset of X, and a collection of concepts $C \subseteq 2^X$ will be called a *concept class*. An element $x \in X$ will be called a *sample*, and a pair $\langle x, a \rangle$ with $x \in X$ and $a \in \{0,1\}$ will be called a *labeled sample*. Likewise, $\overline{x} = (x_1, \ldots, x_m) \in X^m$ is called an *m-sample*, and a *labeled m-sample* is an m-tuple $(\langle x_1, a_1 \rangle, \ldots, \langle x_m, a_m \rangle)$ where $a_i = a_j$ if $x_i = x_j$. For $\overline{x} = (x_1, \ldots, x_m) \in X^m$ and $c \in C$, the *labeled m-sample of c generated by* $\overline{x}$ is given by $sam_c(\overline{x}) = (\langle x_1, I_c(x_1) \rangle, \ldots, \langle x_m, I_c(x_m) \rangle)$ where $I_c(\cdot)$ is the indicator function for the set c. The *sample space* of C is denoted by S_C and consists of all labeled m-samples for all $c \in C$, all $\overline{x} \in X^m$, and all $m \geq 1$.

Let H be a collection of subsets of X. H is called the *hypothesis class*, and the elements of H are called *hypotheses*. Let F_{CH} be the set of all functions $f : S_C \to H$. A function $f \in F_{CH}$ is called *consistent* if it always produces a hypothesis which agrees with the samples, i.e. whenever $h = f(\langle x_1, a_1 \rangle, \ldots, \langle x_m, a_m \rangle)$ we have $I_h(x_i) = a_i$ for $i = 1, \ldots, m$. Given a probability distribution P on X, the *error* of f with respect to P for a concept $c \in C$ and sample $\overline{x}$ is defined as $error_{f,c,P}(\overline{x}) = P(c \Delta h)$ where $h = f(sam_c(\overline{x}))$ and $c \Delta h$ denotes the symmetric difference of the sets c and h.

Finally, in the definition of learnability to be given below, the samples used in forming a hypothesis will be drawn from X independently according to the same probability measure P. Hence, an m-sample will be drawn from X^m according to the product measure P^m.

We can now state the following definition of learnability for every distribution, which is the version from Blumer et al. [24] of Valiant's [118] original definition (without restrictions on computability – see below).

Definition 5.1 (Learnability for Every Distribution) *The pair* (C, H) *is* learnable *if there exists a function* $f \in F_{CH}$ *such that for every* $\epsilon, \delta > 0$ *there is a* $0 < m < \infty$ *such that for every probability measure* P *and every* $c \in C$*, if* $\overline{x} \in X^m$ *is chosen at random according to* P^m *then the probability that* $error_{f,c,P}(\overline{x}) < \epsilon$ *is greater than* $1 - \delta$.

Several comments concerning this definition are in order. First, learnability depends on both the concept class C and the hypothesis class H, which is why we defined learnability in terms of the pair (C, H). However, in the literature the case $H \supseteq C$ is often considered, in which case, for convenience, we may speak of learnability of C in place of (C, C). Second, the sample size m is clearly a function of ϵ and δ but a fixed $m = m(\epsilon, \delta)$ must work uniformly for every distribution P and concept $c \in C$. Because of this, the term *distribution-free learning* is often used to describe this learning framework. Finally, ϵ can be thought of as an accuracy parameter while δ can be thought of as a confidence parameter. The definition requires that the learning algorithm f output a hypothesis that with high probability (greater than $1 - \delta$) is approximately correct (to within ϵ). Angluin and Laird [7] used the term *probably approximately correct* (PAC) learning to describe this definition.

A somewhat more general and useful definition of learnability was actually used by Valiant in [118] and later by others. This definition incorporates both a notion of the size or complexity of concepts and the central idea that the learning algorithm (i.e., the function which produces a hypothesis from labeled samples) should have polynomial complexity in the various parameters. Other variations of this definition, such as seeing positive examples only, or having the choice of positive or negative examples, have also been considered. Some equivalences among the various definitions learnability were shown in [54]. In this report, we will not consider these variations. Also, we will be considering the case that $H \supseteq C$ throughout, so that we will simply speak of learnability of C rather than learnability of (C, H).

A fundamental result of Blumer et al. [24] relates learnability for every distribution to the Vapnik-Chervonenkis (VC) dimension of the concept class to be learned. The notion of VC dimension was introduced in [119] and has been studied and used in [36, 124, 57]. Many interesting concept classes have been shown to have finite VC dimension.

Definition 5.2 (Vapnik-Chervonenkis Dimension) *Let $C \subseteq 2^X$. For any finite set $S \subseteq X$, let $\Pi_C(S) = \{S \cap c : c \in C\}$. S is said to be* shattered *by C if $\Pi_C(S) = 2^S$. The* Vapnik-Chervonenkis dimension *of C is defined to be the largest integer d for which there exists a set $S \subseteq X$ of cardinality d such that S is shattered by C. If no such largest integer exists then the VC dimension of C is infinite.*

A concept class C will be called *trivial* if C contains only one concept or two disjoint concepts. In [24], a definition was also given for what was called a *well-behaved* concept class, which involves the measurability of certain sets used in the proof of their theorem. We will not concern ourselves with the definition here. The following theorem is stated exactly from [24] and was their main result.

Theorem 5.1 *For any nontrivial, well-behaved concept class C, the following are equivalent:*

(i) The VC dimension of C is finite.

(ii) C is learnable.

(iii) If d is the VC dimension of C then

(a) for sample size greater than $\max(\frac{4}{\epsilon}\log\frac{2}{\delta}, \frac{8d}{\epsilon}\log\frac{8d}{\epsilon})$, any consistent function $f \in F_{CH}$ is a learning algorithm for C, and

(b) for $\epsilon < \frac{1}{2}$ and sample size less than $\max(\frac{1}{2\epsilon}\log\frac{1}{\delta}, d(1-2(\epsilon+\delta-\epsilon\delta)))$, no function $f \in F_{CH}$ where $C \subseteq H$ is a learning algorithm for C.

A definition of learnability similar to that of Definition 5.1 can be given for the case of a single, fixed, and known probability measure.

Definition 5.3 (Learnability for a Fixed Distribution) *Let P be a fixed and known probability measure. The pair (C, H) is said to be* learnable with respect to P *if there exists a function $f \in F_{CH}$ such that for every $\epsilon, \delta > 0$ there is a $0 < m < \infty$ such that for every $c \in C$, if $\overline{x} \in X^m$ is chosen at random according to P^m then the probability that $error_{f,c,P}(\overline{x}) < \epsilon$ is greater than $1 - \delta$.*

Conditions for learnability in this case were studied by Benedek and Itai [17]. They introduced the notion of what they called a 'finite cover' for a concept class with respect to a distribution and were able to show that finite coverability characterizes learnability for a fixed distribution. It turns out that their definition of finite coverability is identical to the notion of metric entropy, which has been studied in other literature. Specifically, the measure of error between two concepts with respect to a distribution is a pseudo-metric (or semi-metric). The notion of finite coverability is identical to the notion of finite metric entropy with respect to the pseudo-metric induced by the distribution P.

We define metric entropy below, but first show that P induces a pseudo-metric on the concept class. Define $d_P(c_1, c_2) = P(c_1 \Delta c_2)$ for $c_1, c_2 \subseteq X$ and measurable with respect to P. For $c_1, c_2 \in C$, $d_P(c_1, c_2)$ just represents the error between c_1 and c_2 that has been used throughout. In the following proposition we prove that $d_P(\cdot, \cdot)$ defines a pseudo-metric on the set of all subsets of X measurable with respect to P, and hence defines a pseudo-metric on the concept class C.

Proposition 5.1 *For any probability measure P, $d_P(c_1, c_2) = P(c_1 \Delta c_2)$ is a pseudo-metric on the σ-algebra $\mathcal{S}$ of subsets of X measurable with respect to P. I.e., for all $c_1, c_2, c_3 \in \mathcal{S}$*

(i) $d_P(c_1, c_2) \geq 0$

(ii) $d_P(c_1, c_2) = d_P(c_2, c_1)$

(iii) $d_P(c_1, c_3) \leq d_P(c_1, c_2) + d_P(c_2, c_3)$

Proof: (i) is true since P is a probability measure, and (ii) is true since $c_1 \Delta c_2 = c_2 \Delta c_1$. (iii) follows from subadditivity and the fact that $c_1 \Delta c_3 \subseteq (c_1 \Delta c_2) \cup (c_2 \Delta c_3)$.

□

Note that $d_P(\cdot, \cdot)$ is only a pseudo-metric since it does not usually satisfy the requirement of a metric that $d_P(c_1, c_2) = 0$ iff $c_1 = c_2$. That is, c_1 and c_2 may be unequal but may differ on a set of measure zero with respect to P, so that $d_P(c_1, c_2) = 0$.

We now define metric entropy.

Definition 5.4 (Metric Entropy) *Let (Y, ρ) be a metric space. Define $N(\epsilon) \equiv N(\epsilon, Y, \rho)$ to be the smallest integer n such that there exists $y_1, \ldots, y_n \in Y$ with*

$Y = \cup_{i=1}^{n} B_\epsilon(y_i)$ where $B_\epsilon(y_i)$ is the open ball of radius ϵ centered at y_i. If no such n exists, then $N(\epsilon, Y, \rho) = \infty$. The metric entropy *of Y (often called the* ϵ-entropy*) is defined to be* $\log_2 N(\epsilon)$.

$N(\epsilon)$ represents the smallest number of balls of radius ϵ which are required to cover Y. For another interpretation, suppose we wish to approximate Y by a finite set of points so that every element of Y is within ϵ of at least one member of the finite set. Then $N(\epsilon)$ is the smallest number of points possible in such a finite approximation of Y. The notion of metric entropy for various metric spaces has been studied and used by a number of authors (e.g. see [36, 37, 58, 65, 67, 115]).

For convenience, if P is a distribution we will use the notation $N(\epsilon, C, P)$ (instead of $N(\epsilon, C, d_P)$), and we will speak of the metric entropy of C with respect to P, with the understanding that the metric being used is $d_P(\cdot, \cdot)$. Benedek and Itai [17] proved that a concept class C is learnable for a fixed distribution P iff C has finite metric entropy with respect to P. We state their results formally in the following theorem, which we have written in a form analogous to Theorem 5.1.

Theorem 5.2 *Let C be a concept class and P be a fixed and known probability measure. The following are equivalent:*

(i) The metric entropy of C with respect to P is finite for all $\epsilon > 0$.

(ii) C is learnable with respect to P.

(iii) If $N(\epsilon) = N(\epsilon, C, P)$ is the size of a minimal ϵ-approximation of C with respect to P and $C^{(\epsilon/2)} = \{y_1, \ldots, y_{N(\epsilon/2)}\}$ is an $\frac{\epsilon}{2}$-approximation to C then

(a) for sample size greater than $(32/\epsilon)\ln(N(\epsilon/2)/\delta)$ any function $f : S_C \to C^{(\epsilon/2)}$ which minimizes the number of disagreements on the samples is a learning algorithm for C, and

(b) for sample size less than $\log_2[(1-\delta)N(2\epsilon)]$ no function $f \in F_{CH}$ is a learning algorithm for C.

Note that in condition (iii)(a), only functions whose range was a finite $\frac{\epsilon}{2}$-approximation to C were considered. As noted in [17], a function that simply returns some concept consistent with the samples does not necessarily learn. In fact, in [17] it is claimed that there are examples where for every finite sample there are concepts ϵ-far from the target concept (even with $\epsilon = 1$) that are still consistent with

the samples. The following is a simple example which substantiates their claim. Let $X = [0,1]$, P be the uniform distribution on X, and C be the concept class containing all finite sets of points and the entire unit interval. That is,

$$C = \{\{x_1, \ldots, x_r\} : 1 \leq r < \infty \text{ and } x_i \in [0,1],\ i = 1, \ldots, r\} \cup \{[0,1]\}$$

If the target concept is $[0,1]$ then for every finite sample there are many concepts which are consistent with the sample but are ϵ-far (with $\epsilon = 1$) from $[0,1]$. Namely, any finite set of points which contains the points of the sample is a concept with this property.

5.2 Relationships Between Metric Entropy and the Vapnik-Chervonenkis Dimension

The results of the previous section naturally suggest that there may be some relationships between the VC dimension of a concept class and its metric entropy with respect to various distributions. This is indeed the case. The known relationships essentially provide upper and lower bounds to $\sup_P N(\epsilon, C, P)$ in terms of the VC dimension of the concept class. Upper bounds are more difficult to obtain since these require a uniform bound on the metric entropy over all distributions. The lower bounds result from statements of the form that there exists a distribution P (typically a uniform distribution over some finite set of points) for which $N(\epsilon, C, P)$ is greater than some function of the VC dimension.

Here we prove two new lower bounds which improve on previous known lower bounds. Benedek and Itai [17] have shown that if C is a concept class of finite VC dimension $d \geq 1$ then (i) there is a distribution P such that $\lfloor \log_2 d \rfloor \leq N(\frac{1}{4}, C, P)$, and (ii) if $\epsilon < \frac{1}{2d}$ then there is a distribution P such that $2^d \leq N(\epsilon, C, P)$. First, some comments on these relationships are in order.

Regarding relation (i), we note that if the VC dimension of C is infinite then we can find a sequence of distributions P_n for $n = 1, 2, \ldots$ such that $\lim_{n \to \infty} N(\frac{1}{4}, C, P_n) = \infty$. Relation (i) is proved by considering the uniform distribution on a finite set of d points shattered by C. If the VC dimension of C is infinite, we can find a sequence of distributions under which C unboundedly large metric entropy by taking P_n to be the uniform distribution over n points shattered by C and using (the proof of) relation (i) for each $n = 1, 2, \ldots$. However, for a concept class of infinite VC dimension,

in general we may not necessarily be able to find a particular distribution P for which $N(\epsilon, C, P) = \infty$, but will only be able to approach infinite metric entropy by a sequence of distributions. Nevertheless, in some cases we can achieve infinite metric entropy as shown by the following example. Let $X = [0,1]$ and let C be the set of all Borel sets. Then taking P to be the uniform distribution, we have $N(\frac{1}{4}, C, P) = \infty$ since the infinite collection of sets corresponding to the Haar basis functions (i.e., $c_n = \{x \in [0,1] :$ *the* nth *digit in the binary expansion of* x *is* $1\}$) are pairwise a distance $\frac{1}{2}$ apart with respect to P.

Regarding (ii), a more general statement can be made which does not depend on the VC dimension of C. Specifically, let $x_1, \ldots, x_n \in X$ be distinct points and let $c_1, \ldots, c_k \in C$ be concepts whose intersection with $\{x_1, \ldots, x_n\}$ gives rise to distinct subsets, i.e. $c_i \cap \{x_1, \ldots, x_n\} \neq c_j \cap \{x_1, \ldots, x_n\}$ for $i \neq j$. Note that we must necessarily have $k \leq 2^n$. If we take P to be the uniform distribution on $\{x_1, \ldots, x_n\}$ then we obtain $N(\epsilon, C, P) \geq k$ for $\epsilon < \frac{1}{2n}$. This reduces to (ii) if C has VC dimension d and $c_1, \ldots, c_{2^d}$ are concepts which shatter the set of points $\{x_1, \ldots, x_d\}$. However, our statement is more general since, regardless of the VC dimension of C, it may be possible to find n concepts which give rise to n distinct subsets of $\{x_1, \ldots, x_n\}$ so that $N(\epsilon, C, P) \geq n$ for $\epsilon < \frac{1}{2n}$. This is essentially the basis for Lemma 5.2 below.

First, we prove a result which has a larger range of applicability than (ii) and gives a stronger dependence on d than (i) for $\epsilon < \frac{1}{2}$. Although the bound of (ii) is exponential in d, it is valid only for $\epsilon < \frac{1}{2d}$, so that the range of applicability goes to zero as $d \to \infty$. On the other hand, (i) is valid for a fixed ϵ independent of d (namely $\epsilon = \frac{1}{4}$) but gives only logarithmic dependence on d. The following bound gives exponential dependence on d for a fixed range of applicability ($\epsilon < \frac{1}{2}$).

Lemma 5.1 *If C is a concept class of finite dimension $d \geq 1$ then there is a probability measure P such that*

$$e^{2(\frac{1}{2}-\epsilon)^2 d} \leq N(\epsilon, C, P)$$

for all $\epsilon \leq \frac{1}{2}$.

Proof: Let $\{x_1, \ldots, x_d\}$ be a set of d points that is shattered by C, and let p be the uniform distribution on $\{x_1, \ldots, x_d\}$, i.e. $P(x_i) = \frac{1}{d}$ for $i = 1, \ldots, d$. For this distribution, the only relevant property of a concept c are those x_i which are contained in c. Hence, we can represent c by a d bit binary string with a one in position i indicating that $x_i \in c$, and we can identify the concept class C with the set of all d bit binary strings.

Given two concepts c_1, c_2 represented as binary strings, $d_P(c_1, c_2) = \frac{k}{d}$ where k is the number of bits on which c_1 and c_2 differ, and so $d_P(c_1, c_2) \leq \epsilon$ iff c_2 differs from c_1 on $k \leq \epsilon d$ bits. The number of binary strings that differ on k bits from a given string is $\binom{d}{k}$. Therefore, the number of concepts in an ϵ-ball around a given concept is $\sum_{0 \leq k \leq \epsilon d} \binom{d}{k}$. Since the total number of concepts is 2^d, we need at least

$$2^d / \sum_{0 \leq k \leq \epsilon d} \binom{d}{k}$$

concepts in an ϵ-cover, so that

$$N(\epsilon, C, P) \geq 2^d / \sum_{0 \leq k \leq \epsilon d} \binom{d}{k}$$

Now, Dudley [36] states the Chernoff-Okamoto inequality

$$\sum_{0 \leq k \leq m} \binom{n}{k} p^k (1-p)^{n-k} \leq e^{-(np-m)^2/[2np(1-p)]}$$

for $p \leq \frac{1}{2}$ and $m \leq np$, which can be obtained from a more general inequality (for sums of bounded random variables) of Hoeffding [60]. Taking $n = d$, $p = \frac{1}{2}$, and $m = \epsilon d$ we obtain

$$\sum_{0 \leq k \leq \epsilon d} \binom{d}{k} \leq 2^d e^{-2(\frac{1}{2} - \epsilon)^2 d}$$

for $\epsilon \leq \frac{1}{2}$. Using this in our earlier bound on $N(\epsilon, C, P)$, we get

$$N(\epsilon, C, P) \geq e^{2(\frac{1}{2} - \epsilon)^2 d}$$

for $\epsilon \leq \frac{1}{2}$ which is the desired inequality.

□

Most lower bounds (including the one above) are not particularly useful for small ϵ — i.e., the bounds remain finite as $\epsilon \to 0$. This is the best that can be obtained assuming only that C has VC dimension $d < \infty$, since C itself could be finite. The following result assumes that C is infinite but makes no assumption about the VC dimension of C.

Lemma 5.2 *Let C be a concept class with an infinite number of distinct concepts. Then for each $\epsilon > 0$ there is a probability distribution P such that $N(\epsilon, C, P) > \frac{1}{2\epsilon}$.*

Proof: First, we show by induction that given n distinct concepts, $n-1$ points $x_1, \ldots, x_{n-1}$ can be found such that the n concepts give rise to distinct subsets of $x_1, \ldots, x_{n-1}$. This is clearly true for $n = 2$. Suppose it is true for $n = k$. Then for $n = k+1$ concepts $c_1, \ldots, c_{k+1}$ apply the induction hypothesis to $c_1, \ldots, c_k$ to get $x_1, \ldots, x_{k-1}$ which distinguish $c_1, \ldots, c_k$. c_{k+1} can agree with at most one of $c_1, \ldots, c_k$. Then another point x_k can be chosen to distinguish these two.

Now, let $\epsilon > 0$ and set $n = \lfloor \frac{1}{2\epsilon} \rfloor$. Let $c_1, \ldots, c_n$ be n distinct concepts in C, and let $x_1, \ldots, x_{n-1}$ be $n-1$ points that distinguish $c_1, \ldots, c_n$. Let P be the uniform distribution on $x_1, \ldots, x_{n-1}$. Since the c_i are distinguished by the x_i, $d_P(c_i, c_j) \geq 1/(n-1) = 1/(\lfloor \frac{1}{2\epsilon} \rfloor - 1) > 2\epsilon$. Hence, every concept is within ϵ of at most one of $c_1, \ldots, c_n$ so that $N(\epsilon, C, P) \geq n = \lfloor \frac{1}{2\epsilon} \rfloor$.

□

The following theorem summarizes the result of the two lemmas given above and previous upper and lower bounds obtained by others.

Theorem 5.3 *Let C be a concept class with infinitely many concepts and let $1 \leq d < \infty$ be the VC dimension of C. For $\epsilon \leq 1/2$,*

$$\sup_P \log_2 N(\epsilon, C, P) \geq \max(2d(1/2 - 2\epsilon)^2 \log_2 e, \quad \log_2 \frac{1}{2\epsilon})$$

and for $\epsilon \leq 1/2d$,

$$\sup_P \log_2 N(\epsilon, C, P) \leq d \log_2 (\frac{2e}{\epsilon} \ln \frac{2e}{\epsilon}) + 1$$

The first and second terms of the lower bound follow from Lemmas 5.1 and 5.2 respectively. The upper bound is from [56] which is a refinement of a result from [94] using techniques originally from [36]. A weaker upper bound was also given in [17]. Note that the condition that C contain infinitely many concepts is required only for the second term of the lower bound. Also, note that as an immediate corollary of the theorem we have the result that $\sup_P N(\epsilon, C, P) < \infty$ iff the VC dimension of C is finite.

5.3 Partial Results on Learnability for a Class of Distributions

In this section we prove some partial results regarding learnability for a class of distributions. The definition of learnability in this case is completely analogous to the definitions given earlier, but for completeness we state it formally.

Definition 5.5 (Learnability for a Class of Distributions) *Let $\mathcal{P}$ be a fixed and known collection of probability measures. The pair (C, H) is said to be* learnable with respect to $\mathcal{P}$ *if there exists a function $f \in F_{CH}$ such that for every $\epsilon, \delta > 0$ there is a $0 < m < \infty$ such that for every probability measure $P \in \mathcal{P}$ and every $c \in C$, if $\overline{x} \in X^m$ is chosen at random according to P^m then the probability that $error_{f,c,P}(\overline{x}) < \epsilon$ is greater than $1 - \delta$.*

Benedek and Itai [17] posed the problem of characterizing learnability for a class of distributions as an open problem, and they made the following conjecture.

Conjecture 5.1 *A concept class C is learnable with respect to a class of distributions $\mathcal{P}$ iff for every $\epsilon > 0$,*

$$N(\epsilon, C, \mathcal{P}) \equiv \sup_{P \in \mathcal{P}} N(\epsilon, C, P) < \infty$$

The notation defined in the statement of the conjecture will be used throughout. Namely, if $\mathcal{P}$ is any class of distributions, then $N(\epsilon, C, \mathcal{P})$ is defined by $N(\epsilon, C, \mathcal{P}) = \sup_{P \in \mathcal{P}} N(\epsilon, C, P)$.

For a single distribution, the conjecture reduces immediately to the known result of [17] (stated in Section 5.1). For every distribution, the results of Section 5.2 imply that the condition $\sup_{all\ P} N(\epsilon, C, P) < \infty\ \forall \epsilon > 0$ is equivalent to the condition that C have finite VC dimension. Hence, the conjecture in this case reduces to the known result of [24] (also stated in Section 5.1). As pointed out in [17], the case where $\mathcal{P}$ is finite is similar to the case of a single distribution, and the case where $\mathcal{P}$ contains all discrete distributions is similar to the case of all distributions. The result for all discrete distributions follows again from Section 5.2 since $\sup_{discrete\ P} N(\epsilon, C, P) < \infty\ \forall \epsilon > 0$ iff the VC dimension of C is finite.

We now prove some results for more general classes of distributions. Although our results are far from verifying the conjecture completely, the partial results we

obtain are consistent with it. Furthermore, our results provide some indication of whether prior knowledge of the distribution or attempts at density estimation can help in terms of learnability.

One natural extension to considering a single distribution P_0 is to consider the class of all distributions sufficiently close to P_0. One measure of proximity of distributions is the total variation defined as follows. First, we assume that we are working with some fixed σ-algebra $\mathcal{S}$ of X. Let $\mathcal{P}^*$ denote the set of all probability measures defined on $\mathcal{S}$. For $P_1, P_2 \in \mathcal{P}^*$, the *total variation* between P_1 and P_2 is defined as

$$\|P_1 - P_2\| = \sup_{A \in \mathcal{S}} |P_1(A) - P_2(A)|$$

For a given distribution P_0 and $0 \leq \lambda \leq 1$ define

$$\mathcal{P}_v(P_0, \lambda) = \{P \in \mathcal{P}^* : \|P - P_0\| \leq \lambda\}$$

$\mathcal{P}_v(P_0, \lambda)$ represents all probability measures which are within λ of P_0 in total variation. For $\lambda = 0$, $\mathcal{P}_v(P_0, 0)$ contains only the distribution P_0, and for $\lambda = 1$, $\mathcal{P}_v(P_0, 1)$ contains all distributions.

Another possibility for generating a class of distributions from P_0 utilizes the property that a convex combination of two probability measures is also a probability measure. Specifically, if P_1 and P_2 are probability measures then $\lambda P_1 + (1-\lambda)P_2$ is also a probability measure for $0 \leq \lambda \leq 1$. One interpretation of this convex combination is that with probability λ a point is drawn according to P_1, and with probability $1-\lambda$ the point is drawn according to P_2. Given a distribution P_0 and $0 \leq \lambda \leq 1$, define

$$\mathcal{P}_l(P_0, \lambda) = \{(1-\eta)P_0 + \eta P : \eta \leq \lambda, P \in \mathcal{P}^*\}$$

The distributions in $\mathcal{P}_l(P_0, \lambda)$ can be thought of as those obtained by using P_0 with probability greater than or equal to $1-\lambda$ and using an arbitrary distribution otherwise. Note that, as with $\mathcal{P}_v(P_0, \lambda)$, we have $\mathcal{P}_l(P_0, 0) = \{P_0\}$ and $\mathcal{P}_l(P_0, 1) = \mathcal{P}^*$.

Both $\mathcal{P}_l(P_0, \lambda)$ and $\mathcal{P}_v(P_0, \lambda)$ can be thought as 'spheres' of distributions centered at P_0, i.e. all distributions sufficiently 'close' to P_0 in an appropriate sense. The following proposition verifies the conjecture for $\mathcal{P}_l(P_0, \lambda)$ and $\mathcal{P}_v(P_0, \lambda)$ and shows that a concept class is learnable for $\mathcal{P}_l(P_0, \lambda)$ or $\mathcal{P}_v(P_0, \lambda)$ with $\lambda > 0$ iff it is learnable for all distributions.

Proposition 5.2 *Let C be a concept class and be P_0 a fixed distribution. Then for each fixed and $0 < \lambda \leq 1$, the following are equivalent:*

(i) $N(\epsilon, C, \mathcal{P}_l(P_0, \lambda)) < \infty$ for all $\epsilon > 0$

(ii) C has finite VC dimension

(iii) C is learnable for $\mathcal{P}_l(P_0, \lambda)$

Furthermore, $\mathcal{P}_l(P_0, \lambda) \subseteq \mathcal{P}_v(P_0, \lambda)$ so that the above are equivalent for $\mathcal{P}_v(P_0, \lambda)$ as well.

Proof: **(ii) $\Rightarrow$ (iii)** This follows from the results of [24] (what we have called Theorem 5.1). Namely, (ii) implies learnability for all distributions which implies learnability for $\mathcal{P}_l(P_0, \lambda) \subseteq \mathcal{P}^*$.

(iii) $\Rightarrow$ (i) We prove this by showing that (i) is not true then (iii) is not true. If $N(\epsilon, C, \mathcal{P}_l(P_0, \lambda)) = \infty$ for some $\epsilon > 0$, then for every $M < \infty$ there exists $P_M \in \mathcal{P}_l(P_0, \lambda)$ such that $N(\epsilon, C, P_M) > M$. But then by the results of [17] (what we have called Theorem 5.2), more than $\log_2 N(\epsilon, C, P_M) \geq \log_2(1-\delta)M$ samples are required to learn for P_M. Since M is arbitrary, letting $M \to \infty$ contradicts the fact that C is learnable for $\mathcal{P}_l(P_0, \lambda)$. Thus, $N(\epsilon, C, \mathcal{P}_l(P_0, \lambda)) < \infty$ for all $\epsilon > 0$.

(i) $\Rightarrow$ (ii) For every $P \in \mathcal{P}^*$, let $Q = (1-\lambda)P_0 + \lambda P \in \mathcal{P}_l(P_0, \lambda)$. If $c_1, c_2 \subseteq X$ are any measurable sets, then

$$\begin{aligned} d_Q(c_1, c_2) &= Q(c_1 \Delta c_2) = (1-\lambda)P_0(c_1 \Delta c_2) + \lambda P(c_1 \Delta c_2) \\ &\geq \lambda P(c_1 \Delta c_2) = \lambda d_P(c_1 \Delta c_2) \end{aligned}$$

Therefore, $N(\lambda\epsilon, C, Q) \geq N(\epsilon, C, P)$ and so

$$\begin{aligned} N(\epsilon, C, \mathcal{P}^*) &= \sup_{P \in \mathcal{P}^*} N(\epsilon, C, P) \leq \sup_{P \in \mathcal{P}^*} N(\lambda\epsilon, C, (1-\lambda)P_0 + \lambda P) \\ &= \sup_{Q \in \mathcal{P}_l(P_0, \lambda)} N(\lambda\epsilon, C, Q) < \infty \end{aligned}$$

Hence, from the results of Section 5.2, C has finite VC dimension.

Finally, to show $\mathcal{P}_l(P_0, \lambda) \subseteq \mathcal{P}_v(P_0, \lambda)$, let $Q \in \mathcal{P}_l(P_0, \lambda)$. Then $Q = (1-\eta)P_0 + \eta P$ for some $P \in \mathcal{P}^*$ and $\eta \leq \lambda$. For every $A \in \mathcal{S}$, we have

$$\begin{aligned} |Q(A) - P_0(A)| &= |(1-\eta)P_0(A) + \eta P(A) - P_0(A)| \\ &= \eta |P(A) - P_0(A)| \leq \eta \leq \lambda \end{aligned}$$

Therefore, $\|Q - P_0\| \leq \lambda$ so that $Q \in \mathcal{P}_v(P_0, \lambda)$.

□

The following result shows that learnability of a concept class is retained under finite unions of distribution classes. That is, if a concept class C is learnable for a finite number of sets of distributions $\mathcal{P}_1, \ldots, \mathcal{P}_n$ then it is learnable with respect to their union $\mathcal{P} = \cup_{i=1}^n \mathcal{P}_i$. This is to be expected if the conjecture is true since $N(\epsilon, C, \mathcal{P}) = \max_i N(\epsilon, C, \mathcal{P}_i) < \infty$ iff $N(\epsilon, C, \mathcal{P}_i) < \infty$ for $i = 1, \ldots, n$.

Proposition 5.3 *Let C be a concept class, and let $\mathcal{P}_1, \ldots, \mathcal{P}_n$ be n sets of distributions. If C is learnable with respect to $\mathcal{P}_i$ for $i = 1, \ldots, n$ then C is learnable with respect $\cup_{i=1}^n \mathcal{P}_i$.*

Proof: Let f_i be an algorithm which learns C with respect to $\mathcal{P}_i$, and let $m_i(\epsilon, \delta)$ be the number of samples required by f_i to learn with accuracy ϵ and confidence δ. Define an algorithm f as follows. Ask for

$$m(\epsilon, \delta) = \max_{1 \leq i \leq n} m_i(\frac{\epsilon}{2}, \frac{\delta}{2}) + \frac{32}{\epsilon} \ln \frac{n}{\delta/2}$$

samples. Using the first $\max_i m_i(\frac{\epsilon}{2}, \frac{\delta}{2})$ samples, form hypotheses $h_1, \ldots, h_n$ using algorithms $f_1, \ldots, f_n$ respectively. Then, using the last $\frac{32}{\epsilon} \ln \frac{n}{\delta/2}$ samples, let f output the hypothesis h_i which is inconsistent with the fewest number of this second group of samples. We claim that f is a learning algorithm for C with respect to $\cup_{i=1}^n \mathcal{P}_i$.

Let $P \in \cup_{i=1}^n P_i$, and let $c \in C$. Then $P \in \mathcal{P}_k$ for some k. Since the f_i are learning algorithms with respect to the $\mathcal{P}_i$, at least one h_i (namely h_k) is within $\frac{\epsilon}{2}$ of c with probability (with respect to product measures of P) greater than $1 - \frac{\delta}{2}$. Given that h_i is within $\frac{\epsilon}{2}$ of c for some i, the proof of Lemma 4 from [17] shows that the most consistent hypothesis (on the second group of samples) is within ϵ of c with probability greater than $1 - \frac{\delta}{2}$. Therefore, if A denotes the event that at least one h_i is within $\frac{\epsilon}{2}$ of c then

$$\begin{aligned} Pr\{d_P(f(sam_c(\overline{x})), c) < \epsilon\} &= Pr\{d_P(f(sam_c(\overline{x})), c) < \epsilon \mid A\} \cdot Pr\{A\} \\ &\geq (1 - \frac{\delta}{2})(1 - \frac{\delta}{2}) > 1 - \delta \end{aligned}$$

Thus, f is a learning algorithm for C with respect to $\cup_{i=1}^n \mathcal{P}_i$ using $m(\epsilon, \delta)$ samples.

□

Note that the above result is not true in general for an infinite number of classes of distributions since the sample complexity of the corresponding algorithms may be unbounded (i.e. we may have $\sup_i N(\epsilon, C, \mathcal{P}_i) = \infty$). However, even if $N(\epsilon, C, \mathcal{P}_i)$ is uniformly bounded the proof above does not go through since the application of Lemma 4 from [17] requires finitely many hypotheses. This is essentially the difficulty encountered in attempting to prove the conjecture directly.

For a finite number of distributions $P_1, \ldots, P_n$, define their *convex hull*, denoted by $conv(P_1, \ldots, \mathcal{P}_n)$, as the set of distributions that can be written as a convex combination of $P_1, \ldots, \mathcal{P}_n$. That is,

$$conv(P_1, \ldots, P_n) = \{\lambda_1 P_1 + \cdots + \lambda_n P_n : 0 \leq \lambda_i \leq 1 \text{ and } \lambda_1 + \cdots + \lambda_n = 1\}$$

We now prove the following proposition.

Proposition 5.4 *Let C be a concept class and let $P_1, \ldots, P_n$ be probability measures. The following are equivalent:*

(i) C is learnable with respect to P_i for each $i = 1, \ldots, n$.

(ii) $N(\epsilon, C, conv(P_1, \ldots, P_n)) < \infty$ for all $\epsilon > 0$.

(iii) C is learnable with respect to $conv(P_1, \ldots, P_n)$.

Proof: (iii) $\Rightarrow$ (i) This is immediate.

(i) $\Rightarrow$ (ii) Since C is learnable with respect to P_i for each i, by Theorem 5.2 we have $N(\epsilon, C, P_i) < \infty$ for all $\epsilon > 0$ and $i = 1, \ldots, n$. Let $N_i(\epsilon) = N(\epsilon, C, P_i)$ and let $c_{i,1}, \ldots, c_{i,N_i(\epsilon/2)}$ be an $\frac{\epsilon}{2}$-approximation of C with respect to d_{P_i}. For each $i = 1, \ldots, n$, let $C_{i,j} = \{c \in C : d_{P_i}(c, c_{i,j}) \leq \frac{\epsilon}{2}\}$ for $j = 1, \ldots, N_i(\frac{\epsilon}{2})$. We have $C = \cup_{j=1}^{N_i(\epsilon/2)} C_{i,j}$ for all $i = 1, \ldots, n$. Let

$$C_{k_1,\ldots,k_n} = \bigcap_{i=1}^{n} C_{i,k_i}$$

for $1 \leq k_i \leq N_i(\frac{\epsilon}{2})$, $i = 1, \ldots, n$. Clearly,

$$C = \bigcup_{all\ (k_1,\ldots,k_n)} C_{k_1,\ldots,k_n}$$

Also, by construction the 'diameter' of each $C_{k_1,\ldots,k_n}$ with respect to d_{P_i} is less than or equal to ϵ for all $i = 1, \ldots, n$, i.e. for each $i = 1, \ldots, n$ we have

$$\sup_{c_1,c_2 \in C_{k_1,\ldots,k_n}} d_{P_i}(c_1, c_2) \leq \epsilon$$

Hence, if we define a metric $\rho(\cdot, \cdot)$ by

$$\rho(c_1, c_2) = \max_{1 \leq i \leq n} d_{P_i}(c_1, c_2)$$

then $N(\epsilon, C, \rho) \leq \prod_{i=1}^{n} N_i(\frac{\epsilon}{2}) < \infty$ since we can form an ϵ-approximation of C with respect to ρ by simply taking any point from each $C_{k_1,\ldots,k_n}$ that is nonempty.

Now, if $Q \in conv(P_1, \ldots, P_n)$ then $Q = \sum_{i=1}^{n} \lambda_i P_i$ for some $0 \leq \lambda_i \leq 1$ with $\sum_{i=1}^{n} \lambda_i = 1$. For any measurable $c_1, c_2 \subseteq X$, we have

$$\begin{aligned} d_Q(c_1, c_2) &= \sum_{i=1}^{n} \lambda_i d_{P_i}(c_1, c_2) \\ &\leq \left(\sum_{i=1}^{n} \lambda_i \right) \max_{1 \leq i \leq n} d_{P_i}(c_1, c_2) = \rho(c_1, c_2) \end{aligned}$$

so that $N(\epsilon, C, Q) \leq N(\epsilon, C, \rho)$. Thus,

$$N(\epsilon, C, conv(P_1, \ldots, P_n)) = \sup_{Q \in conv(P_1,\ldots,P_n)} N(\epsilon, C, Q) \leq \prod_{i=1}^{n} N_i(\frac{\epsilon}{2}) < \infty$$

(ii) $\Rightarrow$ (iii) If $N(\epsilon, C, conv(P_1, \ldots, P_n)) < \infty$ for all $\epsilon > 0$, then, in particular, $N(\epsilon, C, P_i) < \infty$ for $i = 1, \ldots, n$ and $\epsilon > 0$. Therefore, we can employ the construction used above in proving that (i) implies (ii) to get a finite $\frac{\epsilon}{2}$-approximation of C uniformly for all $Q \in conv(P_1, \ldots, P_n)$. As shown above, such an approximation can be found with less than or equal to $\prod_{i=1}^{n} N_i(\frac{\epsilon}{4})$ elements where $N_i(\frac{\epsilon}{4}) = N(\frac{\epsilon}{4}, C, P_i)$. Thus, using the proof of Lemma 4 from [17], the algorithm which takes $\frac{32}{\epsilon}\left(\ln \frac{1}{\delta} + \sum_{i=1}^{n} \ln N_i(\frac{\epsilon}{4})\right)$ samples and outputs an element of the $\frac{\epsilon}{2}$-approximation with the smallest number of inconsistent samples is a learning algorithm for C with respect to $conv(P_1, \ldots, P_n)$.

□

The above proposition verifies the conjecture for classes of distributions which are 'convex polyhedra with finitely many sides' in the space of all distributions. In fact, combined with the previous proposition, the conjecture is verified for all finite unions of such polyhedra.

5.4 Discussion and Open Problems

It was first pointed out that the condition for learnability with respect to a fixed distribution obtained in [17] is identical to the notion of finite metric entropy. In considering relationships between the VC dimension of a concept class and its metric entropy, we provided two new lower bounds. Finally, we proved some partial results concerning learnability with respect to a class of distributions. These results are consistent with a conjecture in [17]. Specifically, it was shown that the conjecture holds for any 'sphere' of distributions and for any set of distributions which is a finite union of 'convex polyhedra with finitely many sides'. In addition to verifying the conjecture in these cases, the results indicate some limitations of attempting to enlarge the set of learnable concept classes by requiring learnability only for a class of distributions as opposed to all distributions.

As far as we know, the conjecture on learnability for a class of distributions has not been solved. However, some results regarding learnability with respect to a general class of distributions are known. First, it is easy to show that the condition $\sup_{P\in\mathcal{P}} N(\epsilon, C, P) < \infty$ is necessary for learnability. Natarajan [90, 91] has shown that a somewhat different condition is sufficient. As before, given a finite set of points $x_1, \ldots, x_l$ let $\Pi_C(x_1, \ldots, x_l)$ denote the subsets of $x_1, \ldots, x_l$ generated by intersection with a member of C. Let $|\Pi_C(x_1, \ldots, x_l)|$ denote the cardinality of $\Pi_C(x_1, \ldots, x_l)$, and let $H_{P,l}(C) = \log_2 E_P|\Pi_C(x_1, \ldots, x_l)|$ where E_P denotes the expected value with the points $x_1, \ldots, x_l$ drawn independently according to P. The quantity $H_{P,l}(C)$ was introduced by Vapnik and Chervonenkis [119] and was referred to as a notion of entropy. Based on the results of Vapnik and Chervonenkis, Natarajan has shown that the condition $\lim_{l\to\infty} \sup_{P\in\mathcal{P}} H_{P,l}(C)/l = 0$ is sufficient for learnability of C with respect to P. A simple example shows that this condition is not necessary. Let $X = [0, 1]$, P be the uniform distribution on X, and let C consist of the interval $[0, 1]$ itself together with all finite sets of points. Then C is learnable since every concept is a distance 0 away from either the empty set or the entire interval, so that letting the hypothesis be either $\emptyset$ or $[0, 1]$ based on a single sample is a learning algorithm. On the other hand, $H_{P,l}(C) = 1$ for all l.

Using a distinction considered in [15] and [91] we can refine the conjecture. A concept class will be called strongly learnable if the learner can output any concept consistent with the examples and still have a learning algorithm. The term learnable by itself will still refer to the case where some hypothesis (that need not be con-

sistent) is guaranteed to be close to the target concept. (Note that, unfortunately, the terminology is not particularly good as the term "weak learnability" has been used by others to denote a different idea [106].) The result of Natarajan shows that $\lim_{l\to\infty} \sup_{P\in\mathcal{P}} H_{P,l}(C)/l = 0$ is sufficient for strong learnability since the condition is equivalent to a uniform convergence property of the empirical measures [119] which in turn implies strong learnability. However, it is not necessary since the concept class consisting of all finite subsets of $[0,1]$ is strongly learnable under the uniform distribution (all concepts are a distance zero apart) but the condition (and hence uniform convergence) is clearly violated. As far as we know, it is an open problem to find a single condition which is both necessary and sufficient for strong learnability. And, as mentioned above, for (not necessarily strong) learnability, the condition $\sup_{P\in\mathcal{P}} N(\epsilon, C, P) < \infty$ is necessary and is conjectured to be sufficient.

Hence, there are gaps in the known results characterizing of learnability for a class of distributions. Results on the uniform convergence of empirical measures are related but are not directly helpful. These questions seem somewhat fundamental but quite difficult.

Chapter 6

Active Learning Using Arbitrary Binary Valued Queries

In the original PAC model the examples provided to the learner are obtained from some probability distribution over which the learner has no control. In this sense, the model assumes a purely passive learner. There has been quite a bit of work done on increasing the power of the learner's information gathering mechanism. For example, Angluin [8, 9] has studied a variety of oracles and their effect on learning, Amsterdam [6] considered a model which gives the learner some control over the choice of examples by allowing the learner to focus attention on some chosen region of the instance space, and Eisenberg and Rivest [38] studied the effect on sample complexity of allowing membership queries in addition to random examples.

In this chapter, we study the limits of what can be gained by allowing the most general set of binary valued learner-environment interactions, and giving the learner complete control over the information gathering. Specifically, we consider completely 'active' learning in that the the learner is allowed to ask arbitrary yes/no (i.e., binary valued) questions, and these questions need not be decided on beforehand. That is, the questions the learner asks can depend on previous answers and can also be generated randomly. Many of the oracles previously considered in the literature are simply particular types of yes/no questions (although those oracles that provide counterexamples are not). Both active learning with respect to a fixed distribution and distribution-free active learning are considered. Since we are concerned with active learning, the probability distribution is not used to generate the examples, but is used only to measure the distance between concepts.

Definitions of passive and active learning are provided in Section 6.1. In Sec-

tion 6.2, active learning with respect to a fixed distribution is considered. A simple information theoretic argument shows that active learning does not enlarge the set of learnable concept classes, but as expected can reduce the sample complexity of learning. In Section 6.3, distribution-free active learning is considered. In this case, active learning can take place only in the degenerate situation of a finite concept class. We also consider a form of distribution-free learning in which we assume that the learner knows the distribution being used, so that 'distribution-free' refers only to the requirement that a bound can be obtained on the number of yes/no questions required independent of the distribution used to measure distance between concepts. However, even in this case active learning surprisingly does not enlarge the set of learnable concept classes, but does reduce the sample complexity as expected.

6.1 Definition of Active Learnability

By active learning we will mean that the learner is allowed to ask arbitrary yes/no questions. We will consider only the case $H = C$ throughout, and so we define active learnability in this case only. For a fixed distribution, the only object unknown to the learner is the chosen concept. In this case, an arbitrary binary question provides information of the type $c \in C_0$ where C_0 is some subset of C. That is, all binary questions can be reduced to partitioning C into two pieces and asking to which of the two pieces does c belong. For distribution-free learning (or more generally, learning for a class of distributions) the distribution P is also unknown. In this case, every binary question can be reduced to the form "Is $(c, P) \in q$?" where q is an arbitrary subset of $C \times \mathcal{P}$, so that C and $\mathcal{P}$ can be simultaneously and dependently partitioned. This follows by letting q be the set of (c, P) pairs for which the answer to the binary question is "yes." Thus, the information the active learner obtains is of the form $(\langle q_1, a_1\rangle, \ldots, \langle q_m, a_m\rangle)$ where $q_i \subseteq C \times \mathcal{P}$ and $a_i = 1$ if $(c, P) \in q_i$ and $a_i = 0$ otherwise. The q_i correspond to the binary valued (i.e., yes/no) questions and a_i denotes the answer to the question q_i when the true concept and probability measure are c and P respectively.

In general, q_i can be generated randomly or deterministically and can depend on all previous questions and answers $\langle q_1, a_1\rangle, \ldots, \langle q_{i-1}, a_{i-1}\rangle$. The q_i are not allowed to depend explicitly on the true concept c and probability measure P, but can depend on them implicitly through answers to previous questions. Let $\overline{q} = (q_1, \ldots, q_m)$ denote a set of m questions generated in such a manner, and let $sam_{c,P}(\overline{q})$ denote the set of m

question and answer pairs when the true concept and probability measure are c and P respectively. Let $S_{C,\mathcal{P}}$ denote all sets of m question and answer pairs generated in such a manner for all $c \in C$, $P \in \mathcal{P}$, and $m \geq 1$. By an active learning algorithm we mean an algorithm A for selecting $q_1, \ldots, q_m$ together with a mapping $f : S_{C,\mathcal{P}} \to C$ for generating a hypothesis from $sam_{c,P}(\overline{q})$. In general, A and/or f may be nondeterministic, which results in probabilistic active learning algorithms. If both A and f are deterministic we have a deterministic active learning algorithm. Note that if the distribution P is known then with a probabilistic algorithm an active learner can simulate the information received by a passive learner by simply generating random examples and asking whether they are elements of the unknown concept.

Definition 6.1 (Active Learnability for a Class of Distributions) *Let $\mathcal{P}$ be a fixed and known collection of probability measures. C is said to be* actively learnable with respect to $\mathcal{P}$ *if there exists a function $f : S_{C,\mathcal{P}} \to C$ such that for every $\epsilon, \delta > 0$ there is a $0 < m(\epsilon, \delta) < \infty$ such that for every probability measure $P \in \mathcal{P}$ and every $c \in C$, if $h = f(sam(c, P))$ then the probability (with respect to any randomness in A and f) that $P(h \Delta c) < \epsilon$ is greater than $1 - \delta$.*

6.2 Active Learning for a Fixed Distribution

In this section, we consider active learning with respect to a fixed and known probability distribution. That is, $\mathcal{P}$ consists of a single distribution P that is known to the learner. As we mentioned in Chapter 5, Benedek and Itai [17] have obtained conditions for passive learnability in this case in terms of the metric entropy of C. Specifically, they showed that any passive learning algorithm requires at least $\log_2(1 - \delta)N(2\epsilon, C, P)$ samples and that $(32/\epsilon)\ln(N(\epsilon/2)/\delta)$ samples is sufficient.

The following result shows that the same condition of finite metric entropy is required in the case of active learning. In active learning, the learner wants to encode the concept class to an accuracy ϵ with a binary alphabet, so that the situation is essentially an elementary problem in source coding from information theory [43]. However, the learner wants to minimize the length of the longest codeword rather than the mean codeword length.

Theorem 6.1 *A concept class C is actively learnable with respect to a distribution P iff $N(\epsilon, C, P) < \infty$ for all $\epsilon > 0$. Furthermore, $\lceil \log_2(1 - \delta)N(2\epsilon, C, P) \rceil$ queries are necessary, and $\lceil \log_2(1 - \delta)N(\epsilon, C, P) \rceil$ queries are sufficient. For deterministic*

learning algorithms, $\lceil \log_2 N(\epsilon, C, P) \rceil$ *queries are both necessary and sufficient.*

Proof: First consider $\delta = 0$. $\lceil \log_2 N(\epsilon, C, P) \rceil$ questions are sufficient since one can construct an ϵ-approximation to C with $N(\epsilon, C, P)$ concepts, then ask $\lceil \log_2 N(\epsilon, C, P) \rceil$ questions to identify one of these $N(\epsilon, C, P)$ concepts that is within ϵ of the true concept. $\lceil \log_2 N(\epsilon, C, P) \rceil$ questions are necessary since by definition every ϵ-approximation to C has at least $N(\epsilon, C, P)$ elements. Hence, with any fewer questions there is necessarily a concept in C which is not ϵ-close to any concept the learner might output.

The essential idea of the argument above is that the learner must be able to encode $N(\epsilon, C, P)$ distinct possibilities and to do so requires $\lceil \log_2 N(\epsilon, C, P) \rceil$ questions. Now, for $\delta > 0$, the learner is allowed to make a mistake with probability δ. In this case, it is sufficient that the learner be able to encode $(1 - \delta)N(\epsilon, C, P)$ possibilities since the learner could first randomly select $(1 - \delta)N(\epsilon, C, P)$ concepts from an ϵ-approximation of $N(\epsilon, C, P)$ concepts (each with equal probability) and then ask questions to select one of the $(1 - \delta)N(\epsilon, C, P)$ concepts, if there is one, that is ϵ-close to the true concept. To show the lower bound, first note that we can find $N(2\epsilon) = N(2\epsilon, C, P)$ concepts $c_1, \ldots, c_{N(2\epsilon)}$ which are pairwise at least 2ϵ apart since at least $N(2\epsilon)$ balls of radius 2ϵ are required to cover C. Then the balls $B_\epsilon(c_i)$ of radius ϵ centered at these c_i are disjoint. For each i, if c_i is the true concept then the learning algorithm must output a hypothesis $h \in B_\epsilon(c_i)$ with probability greater than $1 - \delta$. Hence, if k queries are asked, then

$$
\begin{aligned}
&(1 - \delta)N(2\epsilon, C, P) \\
&\qquad \leq \sum_{i=1}^{N(2\epsilon)} \Pr(h \in B_\epsilon(c_i) | c = c_i) \\
&\qquad = \sum_{i=1}^{N(2\epsilon)} \int \Pr(h \in B_\epsilon(c_i) | c = c_i,\ q_1, \ldots, q_k) dA(q_1, \ldots, q_k) \\
&\qquad = \sum_{i=1}^{N(2\epsilon)} \int \sum_{a_1, \ldots, a_k} \Pr(h \in B_\epsilon(c_i) | c = c_i,\ \langle q_1, a_1 \rangle, \ldots, \langle q_k, a_k \rangle) dA(q_1, \ldots, q_k) \\
&\qquad = \int \sum_{a_1, \ldots, a_k} \sum_{i=1}^{N(2\epsilon)} \Pr(h \in B_\epsilon(c_i) | c = c_i,\ \langle q_1, a_1 \rangle, \ldots, \langle q_k, a_k \rangle) dA(q_1, \ldots, q_k) \\
&\qquad = \int \sum_{a_1, \ldots, a_k} \sum_{i=1}^{N(2\epsilon)} \Pr(h \in B_\epsilon(c_i) | \langle q_1, a_1 \rangle, \ldots, \langle q_k, a_k \rangle) dA(q_1, \ldots, q_k) \\
&\qquad \leq \int \sum_{a_1, \ldots, a_k} 1\ dA(q_1, \ldots, q_k)
\end{aligned}
$$

$$= \int 2^k dA(q_1, \ldots, q_k)$$
$$= 2^k$$

where the integral is with respect to any randomness in the questions, the fourth equality (i.e. where conditioning on $c = c_i$ is dropped) follows the fact that the hypothesis generated by the learner is independent of the true concept given the queries and answers, and the second inequality follows from the fact that the $B_\epsilon(c_i)$ are disjoint. Thus, since k is an integer, $k \geq \lceil \log_2(1-\delta)N(2\epsilon, C, P) \rceil$.

Finally, if fewer than $N(\epsilon, C, P)$ possibilities are encoded, then some type of probabilistic algorithm must necessarily be used, since otherwise there would be some concept which the learner would always fail to learn to within ϵ.

□

Thus, compared with passive learning for a fixed distribution, active learning does not enlarge the set of learnable concept classes, but as expected, fewer queries are required in general. However, only a factor of $1/\epsilon$, some constants, and a factor of $1/\delta$ in the logarithm are gained by allowing active learning, which may or may not be significant depending on the behavior of $N(\epsilon, C, P)$ as a function of ϵ.

Note that in active learning very little is gained by allowing the learner to make mistakes with probability δ. That is, there is a very weak dependence on δ in the sample size bounds. In fact for any $\delta \leq 1/2$, we have $\log_2(1-\delta)N(2\epsilon, C, P) = \log_2 N(2\epsilon, C, P) - \log_2 1/(1-\delta) \geq \log_2 N(2\epsilon, C, P) - 1$, so that even allowing the learner to make mistakes half the time results in the lower bound differing from the upper bound and the bound for $\delta = 0$ essentially by only the term 2ϵ versus ϵ in the metric entropy. Also, note that Theorem 6.1 is true for learnability with respect to an arbitrary metric and not just those induced by probability measures.

6.3 Distribution-Free Active Learning

Distribution-free learning (active or passive) corresponds to the case where $\mathcal{P}$ is the set of all probability measures $\mathcal{P}^*$ over, say, the Borel σ-algebra. As mentioned in Chapter 5, Blumer et al. [24] characterized passive learnability for all distributions (i.e., distribution-free) in terms of the VC dimension of C. Specifically, if C has VC dimension $d < \infty$ (and satisfies certain measurability conditions that we will not concern ourselves with) they showed that $\max(\frac{1}{2\epsilon}\log\frac{1}{\delta}, d(1-2(\epsilon+\delta-\epsilon\delta)))$ samples

are necessary and $\max(\frac{4}{\epsilon}\log\frac{2}{\delta}, \frac{8d}{\epsilon}\log\frac{8d}{\epsilon})$ samples are sufficient, although since their work some refinements have been made in these bounds.

The case of distribution-free active learnability is a little more subtle than active learnability for a fixed distribution. For both active and passive learning, the requirement that the learning be distribution-free imposes two difficulties. The first is that there must exist a uniform bound on the number of examples or queries over all distributions — i.e., a bound independent of the underlying distribution. The second is that the distribution is unknown to the learner, so that the learner does not know how to evaluate distances between concepts. Hence, since the metric is unknown, the learner cannot simply replace the concept class with a finite ϵ-approximation as in the case of a fixed and known distribution.

For passive learnability, the requirement that the concept class have finite VC dimension is necessary and sufficient to overcome both of these difficulties. However, for active learning the second difficulty is severe enough that no learning can take place as long as the concept class is infinite.

Lemma 6.1 *Let C be an infinite set of concepts. If $c_1, \ldots, c_n \in C$ is any finite set of concepts in C then there exists $c_{n+1} \in C$ and a distribution P such that $d_P(c_{n+1}, c_i) \geq 1/2$ for $i = 1, \ldots, n$.*

Proof: Consider all sets of the form $b_1 \cap b_2 \cap \cdots \cap b_n$ where b_i is either c_i or $\overline{c_i}$. There are at most 2^n distinct sets $B_1, \ldots, B_{2^n}$ of this form. Note that the B_i are disjoint, their union is X, and each c_i for $i = 1, \ldots, n$ consists of a union of certain B_i. Since C is infinite, there is some set $c_{n+1} \in C$ that is not equal to a union of of any subset of $B_1, \ldots, B_{2^n}$. Then, for such a $c_{n+1} \in C$, we have that for some nonempty B_k, $c_{n+1} \cap B_k$ is nonempty and $c_{n+1} \cap B_k \neq B_k$. Hence, there exist points $x_1, x_2 \in X$ with $x_1 \in c_{n+1} \cap B_k$ and $x_2 \in B_k \setminus c_{n+1}$. Let P be the probability measure which assigns probability $1/2$ to x_1 and $1/2$ to x_2. For each $i = 1, \ldots, n$, either $B_k \subset c_i$ or $B_k \cap c_i = \emptyset$. Thus, in either case $c_{n+1} \Delta c_i$ contains exactly one of x_1 or x_2 so that $d_P(c_{n+1}, c_i) = 1/2$ for $i = 1, \ldots, n$.

□

Theorem 6.2 *C is actively learnable for all distributions iff C is finite.*

Proof: If C is finite it is clearly actively learnable since the learner need only ask $\lceil \log_2 |C| \rceil$ questions where $|C|$ is the cardinality of C to decide which concept is the correct one.

If C is infinite we will show that C is not actively learnable by showing that after finitely many questions an adversary could give answers so that there are still infinitely many candidate concepts which are far apart under infinitely many remaining probability distributions. Since C is infinite, we can repeatedly apply the lemma above to obtain an infinite sequence of concepts $c_1, c_2, \ldots$ and an associated sequence of probability measures $P_1, P_2, \ldots$ such that under the distribution P_i, the concept c_i is a distance $1/2$ away from all preceding concepts. I.e., for each i $d_{P_i}(c_i, c_j) = 1/2$ for $j = 1, \ldots, i-1$.

Now, any question that the active learner can ask is of the form "Is $(c, P) \in q$?" where q is a subset of $C \times \mathcal{P}$. Consider the pairs $(c_1, P_1), (c_2, P_2), \ldots$. Either q or $\overline{q}$ (or both) contain an infinite number of the pairs (c_i, P_i). Thus, an adversary could always give an answer such that an infinite number of pairs (c_i, P_i) remain as candidates for the true concept and probability measure. Similarly, after any finite number of questions an infinite number of (c_i, P_i) pairs remain as candidates. Thus, by the property that $d_{P_i}(c_i, c_j) = 1/2$ for $j = 1, \ldots, i-1$, it follows that for any $\epsilon < 1/2$ the active learner cannot learn the target concept.

□

Essentially, if the distribution is unknown, then the active learner has no idea about 'where' to seek information about the concept. On the other hand, in passive learnability the examples are provided according to the underlying distribution, so that information is obtained in regions of importance. Hence, in the distribution-free case, random samples (from the distribution used to evaluate performance) are indispensable.

Suppose that we remove the second difficulty by assuming that the learner has knowledge of the underlying distribution. Then the learner knows the metric being used and so can form a finite approximation to the concept class. In this case, the distribution-free requirement plays a part only in forcing a uniform bound on the number of queries needed. Certainly, the active learner can learn any concept class that is learnable by a passive learner since the active learner could simply ask queries according to the known distribution to simulate a passive learner. However, the following theorem shows that active learning, even with the side information as to the distribution being used, does not enlarge the set of learnable concept classes.

Theorem 6.3 *If the learner knows the underlying probability distribution then C is actively learnable for all distributions iff C has finite VC dimension. Furthermore,*

$\lceil \sup_P \log_2(1-\delta)N(2\epsilon, C, P)\rceil$ *questions are necessary and* $\lceil \sup_P \log_2(1-\delta)N(\epsilon, C, P)\rceil$ *questions are sufficient. For deterministic algorithms* $\lceil \sup_P \log N(\epsilon, C, P)\rceil$ *questions are both necessary and sufficient.*

Proof: If the distribution is known to the learner, then the result of Theorem 6.1 applies for each distribution. Learnability for all distributions then simply imposes the uniform (upper and lower) bounds requiring the supremum over all distributions for both general (i.e., probabilistic) active learning algorithms and for deterministic algorithms. For the first part of the theorem, we need the following result (mentioned in Section 5.2) relating the VC dimension of a concept class to its metric entropy: the VC dimension of C is finite iff $\sup_P N(\epsilon, C, P) < \infty$ for all $\epsilon > 0$. The first part of the theorem follows immediately from this result.

□

Thus, even with this extra 'side' information, the set of learnable concept classes is not enlarged by allowing an active learner. However, as before one would expect an improvement in the number of samples required. A direct comparison is not immediate since the bounds for passive learnability involve the VC dimension, while the results above are in terms of the metric entropy. A comparison can be made using bounds relating the VC dimension of a concept class to its metric entropy with respect to various distributions as discussed in Section 5.2. Specifically, Theorem 5.3 provides upper and lower bounds to $\sup_P N(\epsilon, C, P)$. This theorem gives bounds on the number of questions needed in distribution-free active learning (with the side information) directly in terms of ϵ, δ and the VC dimension of C. The bounds as stated in Theorem 5.3 are directly applicable to deterministic active learning algorithms or for active learning with $\delta = 0$. For probabilistic algorithms with $\delta > 0$ the quantity $\log_2 1/(1-\delta)$ needs to be subtracted from both the lower and upper bounds.

6.4 Discussion and Open Problems

In this chapter, we considered the effect on PAC learnability of allowing a rich set of learner-environment interactions. Previous work along these lines has provided the learner with access to various types of oracles. Many of the oracles considered in the literature answer queries which are special cases of yes/no questions (although those oracles that provide counterexamples are not of this type). As expected, the use of oracles can often aid in the learning process. To understand the limits of how

much could be gained through oracles, we have considered an active learning model in which the learner chooses the information received by asking arbitrary yes/no questions about the unknown concept and/or probability distribution.

For a fixed distribution, active learning does not enlarge the set of learnable concept classes, but it does have lower sample complexity than passive learning. For distribution-free active learning, the set of learnable concept classes is drastically reduced to the degenerate case of finite concept classes. Furthermore, even if the learner is told the distribution but is still required to learn uniformly over all distributions, a concept class is actively learnable iff it has finite VC dimension.

For completeness, we mention that results can also be obtained if the learner is provided with 'noisy' answers to the queries. The effects of various types of noise in passive learning have been studied [7, 62, 111]. For active learning, two natural noise models are random noise in which the answer to a query is incorrect with some probability $\eta < 1/2$ independent of other queries, and malicious noise in which an adversary gets to choose a certain number of queries to receive incorrect answers. For random noise, the problem is essentially equivalent to communication through a binary symmetric channel, so that standard results from information theory on the capacity and coding for such channels [43] can be applied. For malicious noise, some results on binary searching with these types of errors [101] can be applied. For both noise models, the conditions for fixed distribution and distribution-free learnability are the same as the noise-free case, but with a larger sample complexity. However, the more interesting aspects of our results are the indications of the limitations of active learning, and the noise-free case makes stronger negative statements.

Finally, an open problem that may be interesting to pursue is to study the reduction in sample complexity of distribution-free learning if the learner has access to both random examples and arbitrary yes/no questions. This is similar to the problem considered in [38], but there the learner could choose only examples to be labeled rather than ask arbitrary questions. Our result for the case where the learner knows the distribution being used provides a lower bound, but if the distribution is not known then we expect that for certain concept classes much stronger lower bounds would hold. In particular, we conjecture that results analogous to those in [38] hold in the case of arbitrary binary valued questions, so that, for example, asking yes/no questions could reduce the sample complexity to learn a dense-in-itself concept class (as defined in [38]) by only a constant factor.

Chapter 7

Learning with Generalized Samples and an Application to Stochastic Geometry

In this chapter, we introduce an extension of the PAC model in which the learner may receive information from more general types of samples. By a "generalized sample" we will mean essentially a functional assigning a real number to each concept, where the number assigned may not necessarily be the value of the unknown concept at a point, but could be some other attribute of the unknown concept (e.g., the integral over a region, or the derivative at a given point, etc.). The model is defined for the general case in which the concepts are real valued functions, and is applicable to both distribution-free and fixed distribution learnability. The idea is simply to transform learning with generalized samples to a problem of learning with standard samples over a new instance space and concept class. The PAC learning criteria over the original space is induced by the corresponding standard PAC criteria over the transformed space. Thus, the criteria for learnability and sample size bounds are the usual ones involving metric entropy and a generalization of VC dimension for functions (in the fixed distribution and distribution-free cases respectively).

We consider a particular example of learning from generalized samples that is related to a classical result from stochastic geometry. Namely, we take X to be the unit square in the plane, and consider concept classes which are collections of curves contained in X. For example, one simple concept class of interest is the set of straight line segments contained in X. A much more general concept class we consider is the set of curves in X with bounded length and bounded turn (total absolute curvature).

The samples observed by the learner consist of randomly chosen straight lines labeled as to the number of intersections the random line makes with the target concept (i.e., the unknown curve). We consider learnability with respect to a fixed distribution, where the distribution is the uniform distribution on the set of lines intersecting X. A learnability result is obtained by providing metric entropy bounds for the class of curves under consideration.

The example of learning a curve is closely related to a result from stochastic geometry which states that the expected number of intersections a random line makes with an arbitrary rectifiable curve is proportional to the length of the curve. This result suggests that the length of a curve can be estimated (or "learned") from a set of generalized samples. In fact, this idea has been studied, although primarily from the point of view of deterministic sampling [113, 87]. The learnability result makes the much stronger statement that for certain classes of curves, from just knowing the number of intersections with a set of random lines, the curve itself can be learned (from which the length can then be estimated). Also, for these classes of curves, the learning result guarantees uniform convergence of empirical estimates of length to true length, which does not follow directly from the stochastic geometry result.

Finally, we discuss a number of open problems and directions for further work. We believe the framework presented here can be applied to a number of problems in signal/image processing, geometric reconstruction, and stereology, to provide sample size bounds under a PAC criterion. Some specific problems that may be approachable with these ideas include tomographic reconstruction using random ray or projection sampling and convex set reconstruction from support line or other types of measurements.

7.1 PAC Learning with Generalized Samples

In the PAC learning model discussed in Chapters 5 and 6, the unknown concept was a subset of the instance space X (i.e., and indicator function on X). A natural generalization of this model is to consider the learning of functions as opposed to just sets (i.e., binary valued functions). A very general framework for learning functions was formulated by Haussler [56], building on some fundamental work by Vapnik and Chervonenkis [119, 120, 121], Dudley [36], and Pollard [94]. In this framework, the concept class (hypotheses), denoted by F, is a collection functions from a domain X to a range Y. The samples are drawn according to a distribution on $X \times Y$ from some

class of distributions. A loss function is defined on $Y \times Y$, and the goal of the learner is to produce a hypothesis from F which is close to the optimal one in the sense of minimizing the expected loss between the hypothesis and the random samples.

Learning from generalized samples can be formulated as an extension of the framework in [56] as briefly described below. However, for simplicity of the presentation we consider a restricted formulation which is sufficiently general to treat the example of learning a curve discussed in this paper. We now define more carefully what we mean by learning from generalized samples. Let X be the original instance space as before, and let the concept class F be a collection of real valued functions on X. In the usual model, the information one gets are samples $(x, f(x))$ where $x \in X$ and where $f \in F$ is the target concept. We can view this as obtaining a functional δ_x and applying this functional to the target concept f to obtain the sample $(\delta_x, \delta_x(f)) = (\delta_x, f(x))$. The functional in this case simply evaluates f at the point x, and is chosen randomly from the class of all such "impulse" functionals. Instead, we now assume we get generalized samples in the sense that we obtain a more general functional $\tilde{x}$, which is some mapping from F to $\mathbf{R}$. The observed labeled sample is then $(\tilde{x}, \tilde{x}(f))$ consisting of the functional and the real number obtained by applying this functional to the target concept f. We assume the functional $\tilde{x}$ is chosen randomly from some collection of functionals $\tilde{X}$. Thus, $\tilde{X}$ is the instance space for the generalized samples, and the distribution P is a probability measure on $\tilde{X}$. Let S_F denote the set of labeled m-samples for each $m \geq 1$, for each $\tilde{x} \in \tilde{X}$, and each $f \in F$.

Given P, we can define an error criterion (i.e., notion of distance between concepts) with respect to P as

$$d_P(f_1, f_2) = E|\tilde{x}(f_1) - \tilde{x}(f_2)|$$

This is simply the average absolute difference of real numbers produced by generalized samples on the two concepts. Note that this notion of distance reduces exactly to the notion of distance used in Chapter 5 for the case where the concepts are sets (indicator functions) and the samples are the standard samples. Also, note we could define the framework with more general loss criteria as in [56], but for the example considered in this paper we use the criterion above.

Definition 7.1 (Learning From Generalized Samples) *Let $\mathcal{P}$ be a fixed and known collection of probability measures. Let F be a collection of functions from the instance space X into $\mathbf{R}$, and let $\tilde{X}$ be the instance space of generalized samples for F. F is said to be* learnable with respect to $\mathcal{P}$ from the generalized samples $\tilde{X}$ *if*

there is a mapping $\mathcal{A} : S_F \rightarrow F$ *for producing a hypothesis* h *from a set of labeled samples such that for every* $\epsilon, \delta > 0$ *there is a* $0 < m = m(\epsilon, \delta) < \infty$ *such that for every probability measure* $P \in \mathcal{P}$ *and every* $f \in F$*, if* h *is the hypothesis produced from a labeled* m*-sample drawn according to* P^m *then the probability that* $d_P(f, h) < \epsilon$ *is greater than* $1 - \delta$*.*

If $\mathcal{P}$ is the set of all distributions over some σ-algebra of $\tilde{X}$ then this corresponds to distribution-free learning from generalized samples. If $\mathcal{P}$ consists of a single distribution P then this corresponds to fixed distribution learning from generalized samples. This is a direct extension of the usual definition of PAC learnability (see for example [24]) to learning functions from generalized samples over a class of distributions. In the definition we have assumed that there is an underlying target concept f. As with the restrictions mentioned earlier, this could be removed following the framework of [56].

Learning with generalized samples can be easily transformed into an equivalent problem of PAC learning from standard samples. The concept class F on X corresponds naturally to a concept class $\tilde{F}$ on $\tilde{X}$ as follows. For a fixed $f \in F$, each functional $\tilde{x} \in \tilde{X}$ produces a real number when applied to f. Therefore, f induces a real valued function on $\tilde{X}$ in a natural way. The real valued function on $\tilde{X}$ induced by f will be denoted by $\tilde{f}$, and is defined by

$$\tilde{f}(\tilde{x}) = \tilde{x}(f)$$

The concept class $\tilde{F}$ is the collection of all functions on $\tilde{X}$ obtained in this way as f ranges through F.

We are now in the standard PAC framework with instance space $\tilde{X}$, concept class $\tilde{F}$, and distribution P on $\tilde{X}$. Hence, as usual, P induces a learning criterion or metric (actually only a pseudo-metric in general) on $\tilde{F}$, and as a result of the correspondence between F and $\tilde{F}$, this metric is the equivalent to the (pseudo-)metric d_P induced by P on F defined above. This metric will be denoted by d_P over both F and $\tilde{F}$, and is given by

$$d_P(\tilde{f}_1, \tilde{f}_2) = E|\tilde{f}_1 - \tilde{f}_2| = E|\tilde{x}(f_1) - \tilde{x}(f_2)| = d_P(f_1, f_2)$$

Distribution-free and fixed distribution learnability are defined in the usual way for $\tilde{X}$ and $\tilde{F}$. Thus, a generalized notion of VC dimension for functions (called pseudo dimension in [56]) and metric entropy of $\tilde{F}$ characterize the learnability of $\tilde{F}$ in the distribution-free and fixed distribution cases respectively. These same quantities for

$\tilde{F}$ then also characterize the learnability of F with respect to d_P.

Using results from [56] (based on results from [94]), we have the following result for learning from generalized samples with respect to a fixed distribution. This result is a direct generalization of Theorem 1 from [56] which allows the use of generalized samples.

Theorem 7.1 *F is learnable from generalized samples (or equivalently, $\tilde{F}$ is learnable) with respect to a distribution P if for each $\epsilon > 0$ there is a finite ϵ-cover $\tilde{F}^{(\epsilon)}$ for $\tilde{F}$ (with respect to d_P) such that $0 \leq \tilde{f}_i \leq M(\epsilon)$ for each $\tilde{f}_i \in \tilde{F}^{(\epsilon)}$. Furthermore, a sample size*

$$m(\epsilon, \delta) \geq \frac{2M^2(\epsilon/2)}{\epsilon^2} \ln \frac{2|\tilde{F}^{(\epsilon/2)}|}{\delta}$$

is sufficient for ϵ, δ learnability.

Proof: Let $\tilde{F}^{(\epsilon/2)}$ be an $\frac{\epsilon}{2}$-cover with $0 \leq \tilde{f}_i \leq M(\epsilon/2)$ for each $\tilde{f}_i \in \tilde{F}^{(\epsilon/2)}$. Let $F^{(\epsilon/2)}$ be obtained from $\tilde{F}^{(\epsilon/2)}$ using the correspondence between F and $\tilde{F}$. After seeing $m(\epsilon, \delta)$ samples, let the learning algorithm output a hypothesis $h \in F^{(\epsilon/2)}$ which is most consistent with the data, i.e., which minimizes

$$\frac{1}{m(\epsilon, \delta)} \sum_{i=1}^{m(\epsilon,\delta)} |\tilde{x}_i(h) - y_i|$$

where $(\tilde{x}_i, y_i)$ are the observed generalized samples. Then using Theorem 1 of [56], it follows that with probability greater than $1 - \delta$ we have $d_P(f, h) \leq \epsilon$.

□

Although we will not use distribution-free learning in the example of learning a curve, for completeness we give a result for this case.

Definition 7.2 (Pseudo Dimension) *Let F be a collection of functions from a set Y to $\mathbf{R}$. For any set of points $\overline{y} = (y_1, \ldots, y_d)$ from Y, let $F_{|\overline{y}} = \{(f(y_1), \ldots, f(y_d)) : f \in F\}$. $F_{|\overline{y}}$ is a set of points in $\mathbf{R}^d$. If there is some translation of $F_{|\overline{y}}$ which intersects all of the 2^d orthants of $\mathbf{R}^d$ then $\overline{y}$ is said to be* shattered *by F. Following terminology from [56], the* pseudo dimension *of F, which we denote* $\dim(F)$*, is the largest integer d such that there exists a set of d points in Y that is shattered by F. If no such largest integer exists then* $\dim(F)$ *is infinite.*

We have the following result for distribution-free learning from generalized samples, again using results from [56].

Theorem 7.2 *F is distribution-free learnable from generalized samples (or equivalently, $\tilde{F}$ is distribution-free learnable) if for some $M < \infty$ we have $0 \leq \tilde{f} \leq M$ for every $\tilde{f} \in \tilde{F}$ and if $\dim(\tilde{F}) = d$ for some $1 \leq d < \infty$. Furthermore, a sample size*

$$m(\epsilon, \delta) \geq \frac{64M^2}{\epsilon^2}\left(2d \ln \frac{16eM}{\epsilon} + \ln \frac{8}{\delta}\right)$$

is sufficient for ϵ, δ distribution-free learnability.

Proof: The result follows from a direct application of Corollary 2 from [56], together with the correspondence between F and $\tilde{F}$ and the fact that $d_P(f_1, f_2) = d_P(\tilde{f}_1, \tilde{f}_2)$.

□

Note that the metric entropy of $\tilde{F}$ is identical to the metric entropy of F (since both are with respect to d_P), so that the metric entropy of F characterizes learnability for a fixed distribution as well. However, the pseudo dimension of F with respect to X does *not* characterize distribution-free learnability. This quantity can be very different from the pseudo dimension of $\tilde{F}$ with respect to $\tilde{X}$.

As mentioned above, for simplicity we have defined the concepts to be real valued functions, have chosen the generalized samples to return real values, and have selected a particular form for the learning criterion or metric d_P. Our ideas can easily be formulated in the much more general framework considered by Haussler [56]. Specifically, one could take F to be a family of functions with domain X and range Y. The generalized samples $\tilde{X}$ would be a collection of mappings from F to $\tilde{Y}$. A family of functions $\tilde{F}$ mapping $\tilde{X}$ to $\tilde{Y}$ would be obtained from F by assigning to each $f \in F$ an $\tilde{f} \in \tilde{F}$ defined by $\tilde{f}(\tilde{x}) = \tilde{x}(f)$. As in [56], the distributions would be defined on $\tilde{X} \times \tilde{Y}$, a loss function L would be defined on $\tilde{Y} \times \tilde{Y}$, and for each $\tilde{f} \in \tilde{F}$ the error of the hypothesis $\tilde{f}$ with respect to a distribution would be $EL(\tilde{f}(\tilde{x}), \tilde{y})$ where the expectation is over the distribution on $(\tilde{x}, \tilde{y})$.

Although learning with generalized samples is in essence simply a transformation to a different standard learning problem, it allows the learning framework and results to be applied to a broad range of problems. To show the variety in the type of observations that are available, we briefly mention some types of generalized samples that may be of interest in certain applications. In the case where the concepts are subsets of X (i.e., binary valued functions), some interesting generalized samples might be to draw random (parameterized) subsets (e.g., disks, lines, or other parameterized curves) of X labeled as to whether or not the random set intersects or is contained in

the target concept. Alternatively, the random set could be labeled as to the number of intersections (or length, area, or volume of the intersection, as appropriate). In the case where the concepts are real valued functions, one might consider generalized samples consisting of choosing certain random sets and returning the integral of the concept over these sets. For example, choosing random lines would correspond to tomographic type problems with random ray sampling. Other possibilities might be to return weighted integrals of the concept where the weighting function is selected randomly from a suitable set (e.g., an orthonormal basis), or to sample derivatives of the concept at random points.

7.2 A Result From Stochastic Geometry

In this section we state an interesting and well known result from stochastic geometry. This result will be used and is related to a specific example of learning from generalized samples discussed in the next section.

To state the result, we first need to describe the notion of choosing a "random" straight line, i.e., a uniform distribution for the set of straight lines intersecting a bounded domain. A line in the plane will be parameterized by the polar coordinates r, θ of the point on the line which is closest to the origin, where $r \geq 0$ and $0 \leq \theta \leq 2\pi$. The set (manifold) of all lines in the plane parameterized in this way corresponds to a semi-infinite cylinder.

A well known result from stochastic geometry states that the unique measure (up to a scale factor) on the set of lines which is invariant to rigid transformations of the plane (translation, rotation) is $drd\theta$, i.e., uniform density in r and θ. This measure is thus independent of the choice of coordinate system, and is referred to as the uniform measure (or density) for the set of straight lines in the plane. This measure corresponds precisely to the surface area measure on the cylinder.

From this measure, a uniform probability measure can be obtained for the set of all straight lines intersecting a bounded domain. Specifically, the set of straight lines intersecting a bounded domain X, which we will denote by $\tilde{X}$, is a bounded subset of the cylinder. The uniform probability measure on $\tilde{X}$ is then just the surface area measure of the cylinder suitably normalized (i.e., by the area of $\tilde{X}$).

We can now state the following classic result from stochastic geometry (see e.g. [105, 11]).

Theorem 7.3 *Let X be a bounded convex subset of $\mathbf{R}^2$, and let $c \subset X$ be a rectifiable curve. Suppose lines intersecting X are chosen uniformly, and let $n(\tilde{x}, c)$ denote the number of intersections of the random line $\tilde{x}$ with the curve c. Then*

$$E\, n(\tilde{x}, c) = \frac{2}{A}\mathcal{L}(c)$$

where $\mathcal{L}(c)$ denotes the length of the curve c and A is the perimeter of X.

In the next section, for simplicity we will take X to be the unit square. In this case, the theorem reduces simply to $En(\tilde{x}, c) = \frac{1}{2}\mathcal{L}(c)$.

A surprising (and powerful) aspect of this theorem is that the expected number of intersections a random line makes with the curve c depends only on the length of c but is independent of any other geometric properties of c. In fact, the expression on the left hand side (suitably normalized) can be used as a definition for the length (or one-dimensional measure) of general sets in the plane [109].

An interesting implication of Theorem 7.3 is that the length of an unknown curve can be estimated or "learned" if one is told the number of intersections between the unknown curve and a collection of lines chosen randomly (from the uniform distribution). In fact, deterministic versions of this idea have been studied [113, 87].

7.3 Learning a Curve by Counting Intersections with Lines

In this section, we consider a particular example of learning from generalized samples — that of learning a curve by counting intersections with straight lines. For concreteness we take X to be the unit square in R^2, although our results easily extend to the case where X is any bounded convex domain in $\mathbf{R}^2$. We will consider concept classes C which are collections of curves contained in X. For example, one particular concept class of interest will be the set of straight line segments contained in X. Other concept classes will consist of more general curves in X satisfying certain regularity constraints. The samples observed by the learner consist of randomly chosen straight lines labeled as to the number of intersections the random line makes with the target concept (i.e., the unknown curve). Recall, that with the r, θ parameterization, the set of lines intersecting X, which is the instance space $\tilde{X}$, is a bounded subset of the semi-infinite cylinder. We consider learnability with respect to a fixed distribution,

where the distribution P is the uniform distribution on $\tilde{X}$.

7.3.1 Learning a Line Segment

Consider the case where C is the set of straight line segments in X. In this case, given a concept $c \in C$, every straight line (except for a set of measure zero) intersects c either exactly once or not at all. Thus, $\tilde{C}$ consists of subsets (i.e., binary valued functions) of $\tilde{X}$, where each $\tilde{c} \in \tilde{C}$ contains exactly those straight lines $\tilde{x} \in \tilde{X}$ which intersect the corresponding $c \in C$.

The metric d_P on C and $\tilde{C}$ induced by P is given by

$$d_P(c_1, c_2) = d_P(\tilde{c}_1, \tilde{c}_2) = E|n(\tilde{x}, c_1) - n(\tilde{x}, c_2)| = P(\tilde{c}_1 \Delta \tilde{c}_2)$$

where, as in the previous section, $n(\tilde{x}, c)$ is the number of intersections the line x makes with c. In the case of line segments $n(\tilde{x}, c)$ is either one or zero, i.e. $\tilde{c}$ is binary valued, so that

$$d_P(c_1, c_2) = d_P(\tilde{c}_1, \tilde{c}_2) = P(c_2 \Delta c_2)$$

where $\tilde{c}_1 \Delta \tilde{c}_2$ is the usual symmetric difference of $\tilde{c}_1$ and $\tilde{c}_2$. Hence, $d_P(c_1, c_2)$ is the probability that a random line intersects exactly one of the two line segments c_1 and c_2.

In the case of line segments, a simple bound on the d_P distance between two segments can be obtained in terms of the distances between the endpoints of the segments.

Lemma 7.1 *Let c_1, c_2 be two line segments, and let a_1, b_1 and a_2, b_2 be the endpoints of c_1 and c_2 respectively. Then*

$$d_P(c_1, c_2) \leq \frac{1}{2}\left(||a_1 - a_2|| + ||b_1 - b2||\right)$$

Proof: Since c_1, c_2 are line segments, the distance $d_P(c_1, c_2)$ between c_1 and c_2 is the probability that a random line intersects exactly one of c_1 and c_2. Any line that intersects exactly one of c_1, c_2 must intersect one of the segments $\overline{a_1 a_2}$ or $\overline{b_1 b_2}$ joining the endpoints of c_1 and c_2. Therefore,

$$d_P(c_1, c_2) \leq P(\tilde{x} \cap \overline{a_1 a_2} \neq \emptyset \text{ or } \tilde{x} \cap \overline{b_1 b_2} \neq \emptyset) \leq P(\tilde{x} \cap \overline{a_1 a_2} \neq \emptyset) + P(\tilde{x} \cap \overline{b_1 b_2} \neq \emptyset)$$

Using Theorem 7.3, the probability that a random line intersects a line segment in

the unit square is simply half the length of the line segment, from which the result follows.

□

Using the previous lemma, we can bound the metric entropy of C (and hence $\tilde{C}$) with respect to the metric induced by P.

Lemma 7.2 *Let C be the set of line segments contained in the unit square X, and let P be the uniform distribution on the set of lines intersecting X. Then*

$$N(\epsilon, \tilde{C}, P) = N(\epsilon, C, P) \leq \frac{1}{4\epsilon^4}$$

Proof: We construct an ϵ-cover for C as follows. Consider a rectangular grid of points in X with spacing $\sqrt{2}\epsilon$. Let $C^{(\epsilon)}$ be the set of all line segments with endpoints on this grid. There are $\frac{1}{2\epsilon^2}$ points in the grid, so that there are $\frac{1}{4\epsilon^4}$ line segments in $C^{(\epsilon)}$. (Some of these segments are actually just points and could be eliminated from $C^{(\epsilon)}$ since they are d_P-distance zero from each other and the empty set. However, we ignore this fact since there are just $\frac{1}{2\epsilon^2}$ such points.) For any $c \in C$, there is a $c' \in C^{(\epsilon)}$ such that each endpoint of c' is within ϵ of an endpoint of c. Hence, from Lemma 7.1 $d_P(c, c') \leq \frac{1}{2}(\epsilon + \epsilon) = \epsilon$ so that $C^{(\epsilon)}$ is an ϵ-cover for C with $\frac{1}{4\epsilon^4}$ elements.

□

The construction of the previous lemma allows us to obtain the following learning result for straight line segments.

Theorem 7.4 *Let C be the set of line segments in the unit square X. Then C is learnable by counting intersections with straight lines chosen uniformly using*

$$m(\epsilon, \delta) = \frac{2}{\epsilon^2} \ln \frac{8}{\epsilon^4 \delta}$$

samples.

Proof: Let $\tilde{C}$ be the concept class over $\tilde{X}$ corresponding to C. Then $\tilde{c} \in \tilde{C}$ is defined by $\tilde{c}(\tilde{x}) = n(\tilde{x}, c)$, i.e., $\tilde{c}(\tilde{x})$ is the number of intersections of the line $\tilde{x}$ with c. Clearly, $0 \leq \tilde{c} \leq 1$ (except for a set of measure zero) for every $\tilde{c} \in \tilde{C}$ Using the construction of Lemma 7.2, we have an $\frac{\epsilon}{2}$-cover of $\tilde{C}$ with $4/\epsilon^4$ elements. Hence, the result follows from Theorem 7.1.

□

7.3.2 Learning Curves of Bounded Turn and Length

Now we consider the learnability of a much more general class of curves. First we need some preliminary definitions. We will consider rectifiable curves parameterized by arc length s, so that a curve c of length L is given by

$$c = \{(x_1(s), x_2(s)) | 0 \leq s \leq L\}$$

where $x_1(\cdot)$ and $x_2(\cdot)$ are continuous functions from $[0, L]$ to $\mathbf{R}$ such that $\sqrt{\dot{x}_1^2 + \dot{x}_2^2}$ is defined and equal to unity almost everywhere. If x_1 and x_2 are twice-differentiable at s, then the curvature of c at s, $\kappa(s)$, is defined as the rate of change of the direction of the tangent to the curve at s, and is given by $\kappa(s) = \ddot{x}_2\dot{x}_1 - \ddot{x}_1\dot{x}_2$ The total absolute curvature of c will be denoted by $\kappa(c)$ and is defined by $\kappa(c) = \int_0^L |\kappa(s)|\, ds$

Alexandrov and Reshetnyak [2] have developed an interesting theory for irregular curves. Among other things, they study the notion of the "turn" of a curve, which is a generalization of total absolute curvature to curves which are not necessarily twice-differentiable. For example, for a piecewise linear curve the turn is simply the sum of the absolute angles between adjacent segments. The turn for more general curves can be obtained by piecewise linear approximations. As expected, their notion of turn reduces to the total absolute curvature of a curve for which the latter quantity is defined. We will use the generalized notion of turn presented in [2] throughout, so that our results will apply to curves which are not twice-differentiable (e.g., piecewise linear curves). We let $\kappa(c)$ denote the turn of the curve c.

We will consider classes of curves of bounded length and bounded turn. Specifically, let $C_{K,L}$ be the set of all curves contained in the unit square whose length is less than or equal to L and whose turn is less than or equal to K. Note that for curves contained in a bounded domain, the length of a curve can be bounded in terms of the turn of the curve and the diameter of the domain (Theorem 5.6.1 from [2]; for differentiable curves see for example [105] p. 35). Hence, we really need only consider classes of curves with a bound on the turn. However, for convenience we will carry both parameters K and L explicitly.

As before, the samples will be random lines chosen according to the uniform distribution P on $\tilde{X}$, labeled as to the number of intersections the line makes with the unknown curve c. However, with curves in $C_{K,L}$ the number of intersections with a given line can be any positive integer (as opposed to just zero or one for straight line segments). Thus, the class $\tilde{C}_{K,L}$ consists of a collection of integer valued functions on

$\tilde{X}$ as opposed to just subsets of $\tilde{X}$ as in the previous section.

Also, as before, the results on learning for the set curves will be with respect to the metric d_P induced by the measure P. That is the d_P distance between two curves c_1 and c_2 or their corresponding functions $\tilde{c}_1$, $\tilde{c}_2$ is given by

$$d_P(c_1, c_2) = d_P(\tilde{c}_1, \tilde{c}_2) = E|n(\tilde{x}, c_1) - n(\tilde{x}, c_2)|$$

where the expectation is taken over the random line $\tilde{x}$ with respect to the uniform measure P. This notion of distance between curves has been studied previously (e.g., see [113] and [105] p.38). For example, it is known that d_P is in fact a metric on the set of rectifiable curves, so that d_P satisfies the triangle inequality and $d_P(c_1, c_2) = 0$ implies $c_1 = c_2$. (Note that in the references [105, 113] the notion of distance used is actually $\frac{1}{2}d_P$, but this makes no difference in the metric properties.)

To obtain a learning result for $C_{K,L}$ we will show that each curve in $C_{K,L}$ can be approximated (with respect to d_P) by a bounded number of straight line segments. The metric entropy computation for a single straight line segment can be extended to provide a metric entropy bound for curves consisting of a bounded number of straight line segments. Thus, by combining these two ideas we can obtain a metric entropy bound for $C_{K,L}$ which yields the desired learning result.

First, we need several properties of the d_P metric for curves of bounded turn.

Lemma 7.3 *If c_1, c_2 are curves with a common endpoint (so that $c_1 \cup c_2$ is a curve) and similarly for c_1', c_2' then*

$$d_P(c_1 \cup c_2, c_1' \cup c_2') \leq d_P(c_1, c_1') + d_P(c_2, c_2')$$

Proof: For any line $\tilde{x}$ (except for a set of measure zero), $n(\tilde{x}, c_1 \cup c_2) = n(\tilde{x}, c_1) + n(\tilde{x}, c_2)$ and similarly for c_1', c_2'. Therefore,

$$\begin{aligned} d_P(c_1 \cup c_2, c_1' \cup c_2') &= \frac{1}{2}E|n(\tilde{x}, c_1 \cup c_2) - n(\tilde{x}, c_1' \cup c_2')| \\ &= \frac{1}{2}E|n(\tilde{x}, c_1) - n(\tilde{x}, c_1') + n(\tilde{x}, c_2) - n(\tilde{x}, c_2')| \\ &\leq \frac{1}{2}E|n(\tilde{x}, c_1) - n(\tilde{x}, c_1')| + \frac{1}{2}E|n(\tilde{x}, c_2) - n(\tilde{x}, c_2')| \\ &= d_P(c_1, c_1') + d_P(c_2, c_2') \end{aligned}$$

□

By induction, this result can clearly be extended to unions of any finite number of curves. The case of a finite number of curves will be used in Lemma 7.6 below.

Lemma 7.4 *If c is a curve and $\hat{c}$ is the line segment joining the endpoints of c, then*

$$d_P(c,\hat{c}) = \frac{1}{2}\left(\mathcal{L}(c) - \mathcal{L}(\hat{c})\right)$$

Proof: Each line can intersect $\hat{c}$ at most once, and every line intersecting $\hat{c}$ also intersects c. Therefore, $n(\tilde{x},c) \geq n(\tilde{x},\hat{c})$ so that $|n(\tilde{x},c) - n(\tilde{x},\hat{c})| = n(\tilde{x},c) - n(\tilde{x},\hat{c})$ for all lines $\tilde{x}$ (except a set of measure zero). Hence,

$$d_P(c,\hat{c}) = E|n(\tilde{x},c) - n(\tilde{x},\hat{c})| = E\left(n(\tilde{x},c) - n(\tilde{x},\hat{c})\right) = \frac{1}{2}\mathcal{L}(c) - \frac{1}{2}\mathcal{L}(\hat{c})$$

where the last inequality follows from the stochastic geometry result (Theorem 7.3).

□

We will make use of the following result from [2].

Theorem 7.5 (Alexandrov and Reshetnyak) *Let c be a curve in $\mathbf{R}^n$ with $\kappa(c) < \pi$, and let α be the distance between its endpoints. Then*

$$\mathcal{L}(c) \leq \frac{\alpha}{\cos\frac{\kappa(c)}{2}}$$

Equality is obtained iff c consists of two line segments of equal length.

Lemma 7.5 *For $0 \leq \alpha \leq \pi/6$, $(1/\cos\alpha) - 1 \leq \alpha^2$ so that if c is a curve with turn $\kappa(c) \leq \pi/6$ and $\hat{c}$ is the line connecting the endpoints of c then*

$$d_P(c,\hat{c}) \leq \frac{\mathcal{L}(\hat{c})}{8}\kappa^2(c)$$

Proof: Let $g(\alpha) = 1/\cos\alpha$ and $h(\alpha) = 1 + \alpha^2$. For $0 \leq \alpha \leq \pi/6$, $\sin\alpha \leq 1/2$ and $\cos\alpha \geq \sqrt{3}/2$ so that $\ddot{g}(\alpha) = 2\sin^2\alpha/\cos^3\alpha + 1/\cos\alpha \leq \frac{4}{3\sqrt{3}} + \frac{2}{\sqrt{3}} = \frac{10\sqrt{3}}{9} < 2 = \ddot{h}(\alpha)$. Combining $\ddot{g}(\alpha) < \ddot{h}(\alpha)$ with the fact that $g(0) = h(0)$ and $\dot{g}(0) = \dot{h}(0)$ gives $g(\alpha) \leq h(\alpha)$ and so $1/\cos\alpha - 1 \leq \alpha^2$ for $0 \leq \alpha \leq \pi/6$.

Now, using the above result, Lemma 7.4, and Theorem 7.5 we have

$$d_P(c,\hat{c}) = \frac{1}{2}\left(\mathcal{L}(c) - \mathcal{L}(\hat{c})\right) \leq \frac{1}{2}\mathcal{L}(\hat{c})\left(\frac{1}{\cos\kappa(c)/2} - 1\right) \leq \frac{\mathcal{L}(\hat{c})}{8}\kappa^2(c)$$

□

Lemma 7.6 *If $c \in C_{K,L}$ then for each $\epsilon > 0$ the curve c can be approximated to within ϵ by an inscribed piecewise linear curve with at most $\frac{K^2 L}{8\epsilon}$ segments.*

Proof: As usual, let s denote arc length along c. Since $\kappa(c) \leq K$, for any $\alpha > 0$, we can find a decomposition of c into at most $\lceil K/\alpha \rceil$ pieces $\ell_1, \ldots, \ell_{\lceil K/\alpha \rceil}$ such that $\kappa(\ell_i) \leq \alpha$ for each i. For example, let $s_0 = 0$ and let

$$s_i = \sup\{s_{i-1} \leq s \leq L | \kappa(c(s_{i-1}, s)) \leq \alpha\}$$

where $c(s_{i-1}, s)$ is the part of the curve c between arc length s_{i-1} and s inclusive. Then, let $\ell_i = c(s_{i-1}, s_i)$. By definition, $\kappa(\ell_i) \leq \alpha$. The turn of a curve satisfies $\kappa(c(s,s')) \geq \kappa(c(s,t)) + \kappa(c(t,s'))$ for any $s \leq t \leq s'$ and $\kappa(c(s,s')) \to 0$ as $s' \to s$ from the right ([2], Corollaries 2 and 3, p. 121). From these properties it follows that if $s_i < L$ then for any $\eta > 0$, $\kappa(c(0, s_i + \eta)) \geq i\alpha$. Since $\kappa(c) \leq K$ we must have $s_i = L$ for some $i \leq \lceil K/\alpha \rceil$.

Now, let $\hat{\ell}_i$ be the line segment joining the ends of ℓ_i. Clearly, the union of the $\hat{\ell}_i$ form a piecewise linear curve inscribed in c (i.e., with endpoints of the segments lying on c). From Lemmas 7.3 and 7.5, and the fact that $\mathcal{L}(\hat{\ell}_i) \leq \mathcal{L}(c) \leq L$, we have

$$d_P(c, \cup_{i=1}^{K/\alpha} \hat{\ell}_i) \leq \sum_{i=1}^{K/\alpha} d_P(\ell_i, \hat{\ell}_i) \leq \frac{K}{\alpha} \cdot \frac{L}{8}\alpha^2 = \frac{KL}{8}\alpha$$

Thus, for $\alpha \leq \frac{8\epsilon}{KL}$, $d_P(c, \cup_{i=1}^{K/\alpha} \hat{\ell}_i) \leq \epsilon$ so that $\frac{K}{\alpha} = \frac{K^2 L}{8\epsilon}$ segments suffice for an ϵ-approximation to c by an inscribed piecewise linear curve.

□

Theorem 7.6 *Let $C_{K,L}$ be the set of all curves in the unit square with turn bounded by K and length bounded by L. Let P be the uniform distribution on the set of lines intersecting the unit square, and let d_P be the metric on $C_{K,L}$ defined by $d_P(c_1, c_2) = E|n(\tilde{x}, c_1) - n(\tilde{x}, c_2)|$. Then the metric entropy of $C_{K,L}$ with respect to d_P satisfies*

$$N(\epsilon, C_{K,L}, P) \leq \left(\frac{K^4 L^2}{8\epsilon^4}\right)^{1 + \frac{K^2 L}{4\epsilon}}$$

Proof: We construct an ϵ-cover for C as follows. Consider a rectangular grid of points in the unit square with spacing $\frac{2\sqrt{2}\epsilon^2}{K^2 L}$. Let $C_{K,L}^{(\epsilon)}$ be the set of all piecewise

linear curves with at most $\frac{K^2L}{4\epsilon}$ segments the endpoints of which all lie on this grid. There are $K^4L^2/8\epsilon^4$ points in the grid, so that there are at most $(K^4L^2/8\epsilon^4)^{1+K^2L/4\epsilon}$ distinct curves in $C^{(\epsilon)}_{K,L}$.

To show that $C^{(\epsilon)}_{K,L}$ is an ϵ-cover for $C_{K,L}$, let $c \in C_{K,L}$. By Lemma 7.6 there is a piecewise linear curve $\hat{c}$ with at most $\frac{K^2L}{4\epsilon}$ segments such that $d_P(c,\hat{c}) < \epsilon/2$. We can find a curve $c' \in C^{(\epsilon)}_{K,L}$ close to $\hat{c}$ by finding a point on the grid within $\frac{2\epsilon^2}{K^2L}$ of each endpoint of a segment in $\hat{c}$. By Lemma 7.1 each line segment of c' is a distance at most $\frac{2\epsilon^2}{K^2L}$ (with respect to d_P) from the corresponding line segment of $\hat{c}$. Since $\hat{c}, c'$ consist of at most $\frac{K^2L}{4\epsilon}$ segments, applying Lemma 7.3 we get $d_P(\hat{c},c') \leq \epsilon/2$. Hence, by the triangle inequality $d_P(c,c') \leq \epsilon$.

□

We can now prove a learning result for curves of bounded turn and length.

Theorem 7.7 *Let $C_{K,L}$ be the set of all curves in the unit square with turn bounded by K and length bounded by L. Then $C_{K,L}$ is learnable by counting intersections with straight lines chosen uniformly using*

$$m(\epsilon,\delta) = \frac{K^4L^2}{2\epsilon^4}\ln\frac{2}{\delta}\left(\frac{2K^4L^2}{\epsilon^4}\right)^{1+\frac{K^2L}{2\epsilon}}$$

Proof: Let $\tilde{C}_{K,L}$ be the concept class over $\tilde{X}$ corresponding to $C_{K,L}$. Then $\tilde{c} \in \tilde{C}$ is defined by $\tilde{c}(\tilde{x}) = n(\tilde{x},c)$, i.e., $\tilde{c}(\tilde{x})$ is the number of intersections of the line $\tilde{x}$ with c.

Using the construction of Theorem 7.6, we have an $\frac{\epsilon}{2}$-cover, $\tilde{C}^{(\epsilon/2)}$, of $\tilde{C}_{K,L}$ with $(2K^4L^2/\epsilon^4)^{1+K^2L/2\epsilon}$ elements. Furthermore, each element of the $\frac{\epsilon}{2}$-cover consists of at most $\frac{K^2L}{2\epsilon}$ line segments. Since a line $\tilde{x}$ can intersect each segment at most once, we have $0 \leq \tilde{c}_i(\tilde{x}) \leq \frac{K^2L}{2\epsilon}$ for every $\tilde{c}_i \in \tilde{C}^{(\epsilon/2)}$. Hence, the result follows from Theorem 7.1.

□

It is interesting to note that $\tilde{C}_{K,L}$ has infinite pseudo dimension (generalized VC dimension), so that one would not expect $C_{K,L}$ to be distribution-free learnable. That the pseudo dimension is infinite can be seen as follows. First, assume that $K, L \geq 2\pi$. For each k, let $\tilde{x}_1,\ldots,\tilde{x}_k$ be the set of lines corresponding to the sides of a k-gon inscribed in the unit circle. For any subset G of these k lines, we can find a curve $c_G \in C_{K,L}$ so that $n(\tilde{x}_i,c_G) = 2$ for $\tilde{x}_i \in G$ and $n(\tilde{x}_i,c_G) = 0$ for $x_i \notin G$. Such a

curve can be obtained by taking a point on the unit circle in each arc corresponding to $\tilde{x}_i \in G$, and taking c_G to be the boundary of the convex hull of these points. Then, $\kappa(c_G) = 2\pi$ and $\mathcal{L}(c_G) < 2\pi$ so that $c_G \in C_{K,L}$. Thus, the set $\tilde{x}_1, \ldots, \tilde{x}_k$ is shattered by $\tilde{C}_{K,L}$, and since k is arbitrary the pseudo dimension of $\tilde{C}_{K,L}$ is infinite. For $K, L < 2\pi$ we can apply essentially the same construction over an arc of the unit circle and without taking c_G to be a closed curve.

7.3.3 Connections With the Stochastic Geometry Result

For the class of curves whose length and curvature are bounded by constants, the learnability result of Theorem 7.7 can be thought of as a refinement of the stochastic geometry result. First, using the expression for the expected number of intersections, one can estimate or "learn" the length of c from a set of generalized samples. The learnability result makes the much stronger statement that the curve c itself can be learned (from which the length can then be estimated). However, we emphasize that the learning of the curve is with respect to the metric d_P. To show that the length can be estimated, we need only note that

$$|\mathcal{L}(c_1) - \mathcal{L}(c_2)| = |\frac{1}{2}E(n(y, c_1) - n(y, c_2))| \leq \frac{1}{2}E|n(y, c_1) - n(y, c_2)| = \frac{1}{2}d_P(c_1, c_2)$$

so that if we learn c to within ϵ then the length of c can be obtained to within $\epsilon/2$.

Second, for the class of curves considered, we have a *uniform* learning result. Hence, this refines the stochastic geometry result by guaranteeing uniform convergence of empirical estimates of length to true length for the class of curves considered.

7.4 Discussion and Open Problems

We introduced a model of learning from generalized samples, and considered an application of this model to a problem of reconstructing a curve by counting intersections with random lines. The curve reconstruction problem is closely related to a well known result from stochastic geometry. The stochastic geometry result (Theorem 7.3) suggests that the length of a curve can be estimated by counting the number of intersections with an appropriate set of lines, and this has been studied by others. Our results show that for certain classes of curves the curve itself can be learned from such information. Furthermore, over these classes of curves the estimates of length from a random sample converge uniformly to the true length of a curve.

The learning result for the curves is in terms of a metric induced by the uniform measure on the set of lines. Although some properties of this metric are known, to better understand the implications of the learning result, it would be useful to obtain further properties of this metric. One approach might be to obtain relationships between this metric and other metrics on sets of curves (e.g., Hausdorff metric, d_H). For example, we conjecture that over the class $C_{K,L}$

$$\inf_{\{c_1,c_2|d_H(c_1,c_2)>\epsilon\}} d_P(c_1,c_2) > 0$$

This result combined with the learning result with respect to d_P would immediately imply a learning result with respect to d_H.

The stochastic geometry result holds for any bounded convex subset of the plane, and as we mentioned before, our results can be extended to this case as well. Furthermore, results analogous to Theorem 7.3 can be shown in higher dimensions and in some non-Euclidean spaces [105]. Some results on curves of bounded turn analogous to those we needed also can be obtained more generally [2]. Hence, learning results should be obtainable for these cases.

Regarding other possible extensions for the problem of learning a curve, note that the stochastic geometry result is not true for distributions other than the uniform distribution. Also, we are not aware of any generalizations to cases where parameterized curves other than lines are chosen randomly. However, learnability results likely hold true for some other distributions and perhaps for other randomly chosen parameterized curves, although the metric entropy computations may be difficult.

There is an interesting connection between the example of learning a curve discussed here and a problem of computing the length of curves from discrete approximations. In particular, it was shown in Chapter 3 that computing the length of a curve from its digitization on a rectangular grid requires a nonlocal computation (even for just straight line segments), although computing the length of a line segment from discrete approximations on a random tesselation can be done locally. The construction is essentially a learning problem with intersection samples from random straight lines. Furthermore, the construction provides insight as to why local computation fails for a rectangular digitization and suggests that appropriate deterministic digitizations would still allow local computations. This is related to the work in [87].

We considered here only one particular example of learning from generalized samples. However, we expect that this framework can be applied to a number of problems

in signal/image processing, geometric reconstruction, stereology, stochastic geometry, etc., to provide learnability results and sample size bounds under a PAC criterion. As previously mentioned, learning with generalized samples is in essence simply a transformation to a different standard learning problem, although the variety available in choosing this transformation (i.e., the form of the generalized samples) should allow the learning framework and results to be applied a broad range of problems.

For example, the generalized samples could consist of choosing certain random sets and returning the integral of the concept over these sets. Other possibilities might be to return weighted integrals of the concept where the weighting function is selected randomly from a suitable set (e.g., an orthonormal basis), or to sample derivatives of the concept at random points. One interesting application would be to problems in tomographic reconstruction. In these problems, one is interested in reconstructing a function from a set of projections of the function onto lower dimensional subspaces. One could have the generalized samples consist of choosing random lines labeled according to the integral of the unknown function along the line. This would correspond to a problem in tomographic reconstruction with random ray sampling. Alternatively, as previously mentioned, one could combine the general framework discussed by Haussler [56] with generalized samples, and consider an application to tomography where the generalized samples consist of entire projections. This would be more in line with standard problems in tomography, but with the directions of the projections being chosen randomly.

For more geometric problems in which the concepts are subsets of X, some interesting generalized samples might be to draw random (parameterized) subsets (e.g., disks, lines, or other parameterized curves) of X labeled as to whether or not the random set intersects or is contained in the target concept. Other possibilities might be to label the random set as to the number of intersections (or length, area, or volume of the intersection, as appropriate) with the unknown concept. One interesting application to consider would be the reconstruction of a convex set from various types of data (e.g., see [61, 76, 110, 95]). For example, the generalized samples could be random lines labeled as to whether or not they intersect the convex set (which would provide bounds on the support function). This is actually just a special case of learning a curve which is closed and convex, although tighter bounds should be obtainable due to the added restrictions. Alternatively, the lines could be labeled as to the length of the intersection (which is like the tomography problem with random ray sampling in the case of binary objects). A third possibility (which is actually just

learning from standard samples) would be to obtain samples of the support function.

Formulating learning from generalized samples in the general framework of Haussler [56] allows issues such as noisy samples to be treated in a unified framework. Application of the framework to a particular problem reduces the question of estimation/learning under a PAC criterion to a metric entropy (or generalized VC dimension) computation. This is not meant to imply that such a computation is easy. On the contrary, the metric entropy computation is the essence of the problem and can be quite difficult. Another problem which can be difficult is interpreting the learning criterion on the original space induced by the distribution on the generalized samples. The induced metric is a natural one given the type of information available, but it may be difficult to understand the properties it endows on the original concept class. Finally, although this approach may provide sample size bounds for a variety of problems, it leaves wide open the question of finding good algorithms.

Chapter 8

Can One Decide the Type of the Mean from the Empirical Measure?

8.1 Introduction

Consider the following hypothesis testing problem: Let $x_1, x_2, \cdots$ denote a sequence of i.i.d. random variables with marginal law P_T, with support $[0,1]$. The mean of P_T, denoted $\overline{\mu}_T$, belongs either to a (known) set A which has measure 0 or to its complement $B = A^c$. We want to decide, based on the observation sequence $x_1, x_2, \cdots, x_n$ whether $\overline{\mu}_T \in A$ or not.

This problem was considered by Cover in [28], where he treated the case of $A = \mathcal{Q}_{[0,1]}$, the set of rationals in $[0,1]$, and more generally the case of countable A. He proposed there a test which, for any measure with $\overline{\mu}_T \in A$, will make (a.s.) only a finite number of mistakes, and for measures with $\overline{\mu}_T \in B \backslash N$, the test makes (a.s.) only a finite number of mistakes, where N is a set of Lebesgue measure 0. Some extensions of this result were considered by Koplowitz [66], who showed various properties of sets A which allowed for such a decision and gave some characterizations of the set N.

In this chapter, we extend the result of [28] by allowing the set A to be uncountable, not necessarily of measure 0, such that it satisfies the following structural assumption:

Assumption There exists a monotone sequence of sets A_m increasing to A and an appropriate positive sequence $\epsilon(m) \rightarrow_{m \rightarrow \infty} 0$ such that, for each m, the open blow up $B_m = A_m^{(\sqrt{2\epsilon(m)})} \hat{=} \{x : d(x, A_m) < \sqrt{2\epsilon(m)}\}$ is such that the Lebesgue measure

of $B_m \backslash A_m$ is smaller than $1/m^2$. (We will use the fact that the open blow ups B_m satisfy $(d(A_m, B_m^c))^2 \geq 2\epsilon(m) > 0$.)

We note that this Assumption implies that if A has Lebesgue measure zero then it is a countable union of nowhere dense sets (i.e., is of the first category). The Assumption is satisfied by a class of interesting uncountable sets A, e.g. the Cantor set. Obviously, for countable sets, the Assumption is satisfied. For more along these lines, c.f. Lemma 8.1 and the remarks which follow Theorem 8.1.

In Section 8.2, we describe a decision algorithm which changes its decisions after increasingly longer and longer intervals. Those intervals are chosen using entropy bounds. We prove that this algorithm shares the properties of Cover's decision rule, i.e. it makes a finite number of mistakes a.s. on the set A and on $A^c \backslash N$ for an appropriate set N of Lebesgue measure 0. (A characterization of N follows from our proof and is related to the one given in [66]). In Section 8.3, the results are extended to allow a (countable) sub-decision inside the set A.

8.2 Classifying the Mean in A Versus A^c

We begin by first describing the proposed decision rule. Let $\beta(m)$ be a given sequence of increasing positive integers, to be defined below. For any input sequence $x_1, x_2, \ldots$, parse $x_1, x_2, \ldots$ into the subsequences

$$X^m \hat{=} (x_{\beta(m-1)}, \cdots, x_{\beta(m)-1}) \qquad 1 \leq m < \infty$$

Let $\overline{\mu}_X^m$ denote the empirical mean of the sequence X^m. At the end of each parsing, make a decision whether $\overline{\mu}_T \in A$ according to whether $\overline{\mu}_X^m \in B_m$ or not. Between parsings, don't change the decision. For the sequence $\beta(m)$ defined below in equation (8.7), we claim:

Theorem 8.1 *a) For any measure P_T with $\overline{\mu}_T \in A$, the decision rule will make (a.s.) only a finite number of mistakes, i.e. for a.e. ω there exists an $n(\omega)$ such that the decision is A for all $n > n(\omega)$.*

b) For any measure P_T with $\overline{\mu}_T \in A^c \backslash N$, where N is a set of Lebesgue measure 0, the decision rule will make (a.s.) only a finite number of mistakes, i.e. for a.e. ω there exists an $n(\omega)$ such that the decision is A^c for all $n > n(\omega)$.

Before proving the theorem, we introduce some notation and define the sequence $\beta(m)$. For a set $E \subset [0,1]$, E^c denotes the complement of E and $\overline{E}$ denotes the closure of E, whereas E^o denotes the interior of E. Let μ be a probability measure with support in $[0,1]$. The mean of μ is denoted $\overline{\mu}$. Let $M_\mu(\lambda) \hat{=} E_\mu(\exp(\lambda x))$ denote the moment generating function of μ and let $\Lambda(\lambda) \hat{=} \log(M(\lambda))$. Let $I_\mu(x) = \sup_\lambda(\lambda x - \Lambda(\lambda))$ be the Legendre transform of $\Lambda(\lambda)$, and let $H(\nu|\mu)$ denote the relative entropy of ν with respect to μ, i.e. $H(\nu|\mu) = \int_0^1 d\nu(x) \log(\frac{d\nu}{d\mu})$ if $\frac{d\nu}{d\mu}$ exists and ∞ otherwise. It is known that both $I(x)$ and $H(\nu|\mu)$ are convex, lower semicontinuous functions (e.g, see [32]). Further, it is well known that for any open (closed) set C in $[0,1]$,

$$\inf_{x \in C} I_\mu(x) = \inf_{\{\nu : \int_0^1 x d\nu(x) \in C\}} H(\nu|\mu) \tag{8.1}$$

Next, let $\overline{\mu}_n \hat{=} \frac{1}{n} \sum_{i=1}^n x_i$ denote the empirical mean of the sequence $x_1, x_2, \cdots, x_n$, and let $L_n \hat{=} \frac{1}{n} \sum_{i=1}^n \delta_{x_i}$ denote the empirical measure of the sequence $x_1, x_2, \cdots, x_n$. By the classical Cramer theorem, one has that, for any closed set C, and any probability measure μ with support in $[0,1]$, (c.f., [32], proof of Lemma 1.2.5),

$$P_\mu(\overline{\mu}_n \in C) \leq 2 \exp(-n \inf_{x \in C} I_\mu(x)). \tag{8.2}$$

We next define the sequence $\beta(m)$: for any m, let B_m be the open cover of the set A_m described in the Assumption above. For any m, compute

$$I_m \hat{=} \inf_{\{\mu : \overline{\mu} \in A_m\}} \inf_{x \in B_m^c} I_\mu(x). \tag{8.3}$$

Note that by (8.1), one also has that

$$I_m = \inf_{\{\mu : \overline{\mu} \in A_m\}} \inf_{\{\nu : m_\nu \in B_m^c\}} H(\nu|\mu). \tag{8.4}$$

Since $d(A_m, B_m^c)^2 \geq 2\epsilon(m)$, one has that $I_m \geq \epsilon(m)$. Indeed, by [32], Exercise 3.2.24, $2H(\nu|\mu) \geq ||\nu - \mu||_{var}^2 \geq (d(A_m, B_m^c))^2$, where the last inequality holds for $\{\nu : m_\nu \in B_m^c\}$ and $\{\mu : \overline{\mu} \in A_m\}$. Next, let

$$\alpha(m) \hat{=} \frac{\log 2 + 2 \log m}{I_m} \tag{8.5}$$

Note that, by (8.2), for any μ such that $\overline{\mu} \in A_m$,

$$P_\mu(\overline{\mu}_{\alpha(m)} \in B_m^c) \leq \frac{1}{m^2} \tag{8.6}$$

Finally, let

$$\beta(m) = \sum_{i=1}^{m} \alpha(i), \qquad \beta(0) = 0. \tag{8.7}$$

Proof of Theorem 8.1:

a) Assume $\overline{\mu}_T \in A$. Then there exists an m such that $\overline{\mu}_T \in A_m$. Note however that the event of making an error infinitely often is equivalent to the event of making an error **at the parsing intervals** infinitely often. However,

$$\sum_{m=1}^{\infty} \text{ Prob error in m-th parsing} \leq \sum_{m=1}^{\infty} \frac{1}{m^2} < \infty$$

where we have used (8.6) above. Therefore, part a) of the theorem follows by the Borel-Cantelli lemma.

b) Let C_m denote the $2\sqrt{2\epsilon(m)}$ blow up of B_m. Let

$$N = \bigcap_{n=1}^{\infty} \bigcup_{m=n}^{\infty} C_m \backslash A$$

Clearly, the Lebesgue measure of N is zero. Now we may repeat the arguments of part a) in the following way: let $\overline{\mu}_T \in B \backslash N$. For an m_0 large enough, $\overline{\mu}_T \in C_m^c$ for all $m > m_0$. On the other hand, $d(\overline{\mu}_T, B_m)^2 \geq 2\epsilon(m)$ by our construction. Noting that the rate function $\inf_{x \in B_m} I_{P_T}(x) > \epsilon(m)$, the proof follows identically as in part a).

□

Remarks

1) The theorem could have been proved by obtaining (8.6) using more traditional bounds but with a slower decision procedure (i.e., larger $\alpha(m)$).

2) It is interesting to note that the Cantor set satisfies the Assumption. Indeed, the covering sets B_m are just the intervals associated with the Cantor partition.

3) By modifying the structure of the decision rule, one may also make a hypothesis test inside A. This is pursued in Section 8.3.

We conclude this section by a (partial) characterization of the sets A of measure 0 which satisfy the Assumption:

Lemma 8.1 *A set A which is of measure 0 and which satisfies the Assumption is of the first category (i.e., A is a countable union of nowhere dense sets). Conversely, a* **closed** *set A of Lebesgue measure zero satisfies the Assumption if A is of the first category.*

Proof:

($\Longrightarrow$) From the Assumption, $A = \cup_m A_m$. We need only show that each A_m is nowhere dense. But this follows immediately from the existence of a sequence of open blow ups of A_m with arbitrarily small Lebesgue measure (namely, B_k for $k \geq m$).

($\Longleftarrow$) If A is of the first category then $A = \cup_i S_i$ where each S_i is nowhere dense. Let $A_m = \cup_{i=1}^m S_i$. Clearly, the A_m monotonically increase to A. Also, since A_m is nowhere dense, and A is closed, $|A_m^{(\delta)}| \to 0$ as $\delta \to 0$ where $|\cdot|$ denotes Lebesgue measure and $A_m^{(\delta)} = \{x : d(x, A_m) < \delta\}$ is the (open) δ-neighborhood of A_m. For each m, choose any $\delta_m > 0$ such that $|A_m^{(\delta_m)}| < 1/m^2$. Then the Assumption is satisfied with $B_m = A_m^{(\delta_m)}$ and $\epsilon(m) = \delta_m^2/2$.

□

8.3 Countable Hypothesis Testing

In this section, we refine the decision rule to allow for deciding among a countable set of hypotheses. In addition to deciding whether or not $\overline{\mu}_T \in A$, we also make a hypothesis test inside A. Suppose that A is written as $A = \cup_{i=1}^\infty S_i$ where the S_i are disjoint. We are interested not only in whether $\overline{\mu}_T \in A$, but if so to which of the S_i does $\overline{\mu}_T$ belong. Specifically, we wish to decide among the following countable set of hypotheses:

$$H_i : \overline{\mu}_T \in S_i, \quad i = 1, 2, \ldots$$

$$H_0 : \overline{\mu}_T \notin A$$

For the theorem below, restrictions must be placed on the decomposition of A. Namely, we assume that the S_i are pairwise positively separated meaning that $d(S_i, S_j) > 0$

for every $i \neq j$. (Note that, as before, A is required to satisfy the structural Assumption of the introduction.)

We modify our previous decision rule as follows. At the end of each parsing (defined by the sequence $\beta(m)$), find the least index k (if one exists) such that $\overline{\mu}_{\alpha(m)}$ is contained in the $\sqrt{2\epsilon(m)}$ open blow up of $S_k \cap A_m$. If such a k exists, then decide that $\overline{\mu}_T \in S_k$. Otherwise (if $m_{\alpha(m)} \notin (S_i \cap A_m)^{(\sqrt{2\epsilon(m)})}$ for all i) decide that $\overline{\mu}_T \notin A$. Alternatively, we can think of this decision procedure as first deciding whether or not $\overline{\mu}_T \in A$ as before. Then, if the decision is that $\overline{\mu}_T \in A$, make a refinement by deciding that $\overline{\mu}_T \in S_k$ where k is the least index such that $m_{\alpha(m)} \in (S_i \cap A_m)^{(\sqrt{2\epsilon(m)})}$.

Theorem 8.2 *If $A = \cup_{i=1}^{\infty} S_i$ satisfies the Assumption and the S_i are pairwise positively separated then*

a) For any measure P_T with $\overline{\mu}_T \in S_i$ for some i, the decision rule will make (a.s.) only a finite number of mistakes, i.e. for a.e. ω there exists an $n(\omega)$ such that the decision is S_i for all $n > n(\omega)$.

b) For any measure P_T with $\overline{\mu}_T \in A^c \backslash N$, where N is a set of Lebesgue measure 0, the decision rule will make (a.s.) only a finite number of mistakes, i.e. for a.e. ω there exists an $n(\omega)$ such that the decision is A^c for all $n > n(\omega)$.

Proof:

a) Suppose that $\overline{\mu}_T \in S_i$. By the same considerations that led to (8.6), for any μ such that $\overline{\mu} \in S_i \cap A_m$ we have

$$P_\mu(\overline{\mu}_{\alpha(m)} \notin (S_i \cap A_m)^{(\sqrt{2\epsilon(m)})}) \leq \frac{1}{m^2} \tag{8.8}$$

Since $\overline{\mu}_T \in S_i \subseteq A$, for sufficiently large m, $\overline{\mu}_T \in A_m$. Also, since the S_j are pairwise positively separated and i is finite, for large enough m the sets $(S_j \cap A_m)^{(\sqrt{2\epsilon(m)})}$ and $(S_i \cap A_m)^{(\sqrt{2\epsilon(m)})}$ are disjoint for all $j < i$. That is, for sufficiently large m, denoted $m_0(i)$, as long as $\overline{\mu}_{\alpha(m)} \in (S_i \cap A_m)^{(\sqrt{2\epsilon(m)})}$ we have $\overline{\mu}_{\alpha(m)} \notin (S_j \cap A_m)^{(\sqrt{2\epsilon(m)})}$ for all $j < i$. Hence, for all $m > m_0(i)$, i is the least index satisfying the requirements of the decision procedure (so that a correct decision is made) iff $\overline{\mu}_{\alpha(m)} \in (S_i \cap A_m)^{(\sqrt{2\epsilon(m)})}$. Therefore,

$$\sum_{m=1}^{\infty} \text{Prob error in m-th parsing} \leq m_0(i) + \sum_{m=m_0(i)+1}^{\infty} P(\overline{\mu}_{\alpha(m)} \notin (S_i \cap A_m)^{(\sqrt{2\epsilon(m)})})$$

$$\leq m_0(i) + \sum_{m=1}^{\infty} \frac{1}{m^2} < \infty$$

so that part a) follows by the Borel-Cantelli Lemma.

b) This part is identical to part b) of Theorem 8.1.

□

Remarks

1) Cover's result on countable hypothesis testing is a special case of this result since every countable set A clearly satisfies the Assumption and can be written as the union of pairwise positively separated sets.

2) If one is willing to allow the test to fail for some points in A, then the requirement that the S_i be pairwise positively separated can be dropped. The set $N_2 \subset A$ on which the test fails in the general case can be characterized, and presumably conditions on the S_i for which N_2 is a null set could be obtained.

8.4 A Symmetric Decision Criterion and the Lebesgue Density Theorem

In the previous sections as well as in [28, 66] mistakes were allowed on a set of measure zero, but only in A^c. In this sense, the criterion was asymmetric in the roles played by A and A^c. Suppose instead we allow mistakes on *any* set of measure zero (i.e., in either A and/or A^c). In this section, we consider such a symmetric decision criterion and show that in this case the test can be accomplished for any (measurable) A. The essential approach is based on the Lebesgue density theorem, but much of the details of the parsing and large deviations bounds are the same as those in Section 8.2.

Let A be a measurable set. The Lebesgue density of A at a point x is defined as

$$D_A(x) \doteq \lim_{r \to 0} \frac{m(A \cap B_r(x))}{m(B_r(x))}$$

when the limit exists. Below we state the Lebesgue density theorem which is a classical result from measure theory.

Theorem 8.3 (Lebsgue Density Theorem) *For any measurable set A, $D_A(x)$ exists and equals 1 if $x \in A$ and 0 if $x \notin A$ except for a set of Lebesgue measure 0.*

Proof: See [41], p.14.

□

We will construct a decision procedure for the symmetric criterion based on this theorem. The essential idea is that, based on convergence of the empirical mean obtained from the observation sequence, we can compute an approximation to the Lebesgue density of A at the true mean. As long as our approximation to the Lebesgue density of A at $\bar{\mu}$ converges, by the Lebesgue density theorem we can make a correct decision as to whether or not $\bar{\mu} \in A$ except on a set of measure zero.

Specifically, the decision procedure is as follows. Let δ_n be any positive sequence with $\delta_n \to 0$. As before parse the input sequence $x_1, x_2, \ldots$ into the subsequences

$$X^n \hat{=} (x_{\beta(n-1)}, \ldots, x_{\beta(n)-1})$$

where the parsing sequence $\beta(n)$ will be defined below (and depends on the choice of δ_n). At the end of each parsing compute the 'relative measure' of A in an interval $(\bar{\mu}_n - \delta_n, \bar{\mu}_n + \delta_n)$ around the empirical mean $\bar{\mu}_n$ of the n-th subinterval. I.e., compute

$$d_n \hat{=} \frac{m(A \cap B_{\delta_n}(\bar{\mu}_n))}{2\delta_n}$$

where $B_r(x)$ is the open ball of radius r centered at x, $\bar{\mu}_n$ is the empirical mean of the sequence X^n, and $m(\cdot)$ denotes Lebesgue measure. Decide $\bar{\mu} \in A$ if $d_n > 1/2$ and decide $\bar{\mu} \notin A$ otherwise.

Since $\delta_n \to 0$, we have that $D_A(\bar{\mu}) = \lim_{n\to\infty} m(A \cap B_{\delta_n}(\bar{\mu}))/m(B_{\delta_n}(\bar{\mu}))$. Since we decide $\bar{\mu} \in A$ iff $d_n > 1/2$, we will eventually stop making incorrect decisions as long as there exists N such that for $n > N$ $|d_n - m(A \cap B_{\delta_n}(\bar{\mu}))/m(B_{\delta_n}(\bar{\mu}))| < 1/4$. But, we have $|m(A \cap B_{\delta_n}(\bar{\mu}_n)) - m(A \cap B_{\delta_n}(\bar{\mu}))| \leq |\bar{\mu}_n - \bar{\mu}|$, so that if $|\bar{\mu}_n - \bar{\mu}| < \delta_n/2$ then we will make a correct decision.

Now, we can find an appropriate parsing sequence $\beta(n)$ using large deviations techniques as before. Specifically, we can find $\beta(n)$ such that

$$P(|\bar{\mu}_n - \bar{\mu}| > \delta_n/2) < 1/n^2$$

For example, let

$$I_n(\bar{\mu}) = \inf_{\{\eta: E\eta = \bar{\mu}\}} \inf_{\{\nu: E\nu \in B_{\frac{\delta_n}{2}}(\bar{\mu})^c\}} H(\nu|\eta)$$

and let $I_n = \inf_{\bar{\mu}} I_n(\bar{\mu})$. For η and ν such that $E\eta = \bar{\mu}$ and $|E\nu - \bar{\mu}| \geq \delta_n/2$ we have $H(\eta|\nu) \geq \frac{1}{2}||\eta - \nu||_{var} \geq d(\bar{\mu}, B_{\frac{\delta_n}{2}}(\bar{\mu})^c)^2 = \delta_n^2/4$, and so $I_n \geq \delta_n^2/4$. As before let

$$\alpha(n) = \frac{\log 2 + 2\log n}{I_n}$$

Finally, let (again as before)

$$\beta(n) = \sum_{i=1}^{n} \alpha(i), \qquad \beta(0) = 0$$

We can show the following result.

Theorem 8.4 *For any measurable A, there is a set N_0 of Lebesgue measure zero such that for any measure μ with $\bar{\mu} \notin N_0$ the decision procedure a.s. makes only a finite number of mistakes in deciding whether or not $\bar{\mu} \in A$.*

Proof: Let N_0 be the set on which $D_A(x)$ is not equal to the characteristic function of A at x. By the Lebesgue density theorem, N_0 has measure zero. Now for $\bar{\mu} \notin N_0$,

$$\begin{aligned}
\sum_{n=1}^{\infty} \text{Prob. error in n-th parsing} &\leq N + \sum_{n=N+1}^{\infty} \text{Prob. error in n-th parsing} \\
&\leq N + \sum_{n=N+1}^{\infty} P(|\bar{\mu}_n - \bar{\mu}| > \frac{\delta_n}{2}) \\
&< N + \sum_{n=N+1}^{\infty} \frac{1}{n^2} < \infty
\end{aligned}$$

Hence, the theorem follows by the Borel-Cantelli lemma and the fact that making an error infinitely often is equivalent to making an error at the parsing intervals infinitely often.

□

Remarks:

1) As before this should be extendible to a countable hypothesis test. In this case, we expect that the test could be accomplished for an arbitrary (measurable) partition except on some set of measure zero.

2) Perhaps the results of Sections 8.2 and 8.3 on making no errors inside A can be obtained by computing the Lebesgue density of blow ups of the A_n's that were used in those sections.

3) It may be possible to extend the results by requiring the set on which mistakes are made to be of measure zero with respect some measure other than Lebesgue measure. Essentially the same argument should go through as long as a density theorem analogous to the Lebesgue density theorem holds.

8.5 Discussion and Open Problems

We provided a sufficient condition under which the mean of an unknown random variable can be determined a.s. to belong to a set A or its complement. Our criterion extends previous results [28, 66] on this problem. It would be interesting to obtain a better understanding of the structural assumption required for our decision procedure. We provided a result which under certain conditions relates the structural assumption to the notion of first category, but other equivalent characterizations of sets which satisfy the assumption are lacking.

There are several directions that may be worthwhile to pursue further. First, a natural generalization is to consider splitting the set of probability measures on $[0,1]$ into two classes (not necessarily according to the means of the distributions), and attempting to decide which class the unknown measure belongs. This is pursued in Chapter 9. Another interesting direction would be to obtain *necessary* conditions under which a test satisfying the same criteria could be performed. Dembo and Peres [31] have some recent results along these lines. A perhaps difficult direction would be to study optimal rates for the decision procedure, e.g., in terms of the decay rate of the probability of an incorrect decision. A generalization which should not be too difficult would be to require the set N on which failure is allowed to be of measure zero with respect to some different measure. That is, we considered the case in which N is required to be of Lebesgue measure zero. Instead, one could consider an arbitrary probability measure on the interval $[0,1]$ and require N to have probability zero. We feel that there are connections between the "learning model" presented in this chapter and other more standard learning models, such as the PAC model discussed in previous chapters and identification in the limit [47]. A somewhat vague, but perhaps extremely interesting, direction would be to develop such connections.

Chapter 9

A General Classification Rule for Probability Measures

9.1 Introduction

Let $x_1, \ldots, x_n$ be i.i.d. samples drawn from some distribution μ. We assume x_i takes values in some compact Polish space Σ, which for concreteness should be thought of as $[0,1]^d \subset \mathbf{R}^d$. Let $\mathcal{M}_1(\Sigma)$ denote the space of probability measures on Σ. We put on $\mathcal{M}_1(\Sigma)$ the Prohorov metric, whose topology is equivalent to the weak topology.

We consider here the following problems:

P-1) Based on the sequence of observations $(x_1, x_2, \ldots)$, decide whether $\mu \in A$ or $\mu \in A^c$, where A is some given set satisfying certain structural properties (c.f. A-1 below).

P-2) Based on the sequence of observations $(x_1, x_2, \ldots)$, decide whether $\mu \in A_i$ where all $A_i \subset \mathcal{M}_1(\Sigma)$, $i = 1, 2, \ldots$ are sets satisfying structural properties (c.f. A-1 below).

Relaxations of the basic assumption concerning the i.i.d. structure of the observations $x_1, x_2, \ldots$ are presented in Sections 9.3 and 9.4.

Since $\mathcal{M}_1(\Sigma)$ is a Polish space, there exist on $\mathcal{M}_1(\Sigma)$ many finite measures which we may assume to be normalized to have a total mass 1. Suppose one is given a particular measure, denoted G, on $\mathcal{M}_1(\Sigma)$. We assume that G charges all open sets in $\mathcal{M}_1(\Sigma)$, i.e. for any open set Φ, $G(\Phi) > 0$. G will play the role of the Lebesgue measure in the following structural condition, which is reminiscent of the assumption in [71]:

A-1) There exists a sequence of open sets C_m and closed sets B_m, and a sequence of positive constants $\epsilon(m)$ such that:

1) $\forall \mu \in A \; \exists m_0(\mu) < \infty \; s.t. \; \forall m > m_0(\mu), \mu \in B_m$.

2) $d(B_m, C_m^c) = \sqrt{2\epsilon(m)} > 0$.

3) $G(\bigcap_{n=1}^{\infty} \bigcup_{m=n}^{\infty} (C_m \backslash A)) = 0$.

The underlying idea here is the same as that in Chapter 8. A-1) is an embellishment of the structural assumption in Chapter 8, which corresponds to the case where B_m is a monotone sequence and C_m are taken as the $2\sqrt{2\epsilon(m)}$ blow up of B_m. This form of A-1) was proposed to us by A. Dembo and Y. Peres. We note that as in Chapter 8, the assumption is immediately satisfied for countable sets A by taking as B_m the union of the first m components of A and noting that, for a finite measure on a metric space, $G(B(x,\delta)\backslash\{x\}) \rightarrow_{\delta \rightarrow 0} 0$ where $B(x,\delta)$ denotes the open ball of radius δ around x. More generally, A-1) is satisfied for any closed set by taking $B_m = A$ and using for C_m a sequence of open sets which include A whose measure converges to the outer measure of A. Since C_m is open and Σ is compact, it follows that $d(A, C_m^c) > 0$, and A-1) is satisfied. By the same considerations, it also follows that A-1) is satisfied for any countable union of closed sets.

9.2 Classification in A versus A^c

The definition of success of the decision rule will be similar to the one used in [71]. Namely, a test which makes at each instant n a decision whether $\mu \in A$ or $\mu \in A^c$ based on $x_1, x_2, \ldots$ will be called successful if:

(S.1) $\forall \mu \in A$, a.s. ω, $\exists T(\omega)$ s.t. $\forall n > T(\omega)$, the decision is 'A'.

(S.2) $\exists N \subset \mathcal{M}_1(\Sigma)$ s.t.

(S.2.1) $G(N) = 0$

(S.2.2) $\forall \mu \in A^c \backslash N$, a.s. ω, $\exists T(\omega)$ s.t. $\forall n > T(\omega)$, the decision is 'A^c'.

Note that the outcome is unspecified on N. Note also that the definition is asymmetric in the roles played by A, A^c in the sense that errors in A are not allowed at all.

Let $\mu_n = \frac{1}{n} \sum_{i=1}^{n} \delta_{x_i}$ where δ_{x_i} is the probability measure concentrated at x_i. We

recall that μ_n satisfies a large deviation principle, i.e.

$$\begin{aligned} -\inf_{\theta\in\bar{A}} H(\theta|\mu) &\leq \liminf_{n\to\infty}\frac{1}{n}\log P(\mu_n\in A) \\ &\leq \limsup_{n\to\infty}\frac{1}{n}\log P(\mu_n\in A) \\ &\leq -\inf_{\theta\in A^0} H(\theta|\mu) \end{aligned} \tag{9.1}$$

where $\bar{A}(A^o)$ denotes the closure (interior) of a set $A\subset\mathcal{M}_1(\Sigma)$ in the weak topology, respectively, and

$$H(\theta|\mu)=\begin{cases}\int_\Sigma d\theta\log\dfrac{d\theta}{d\mu} & \text{if } d\theta << d\mu \\ \infty & \text{otherwise}\end{cases} \tag{9.2}$$

Before turning to a description of the proposed decision rule, we will need a strengthened version of the upper bound in Sanov's theorem (9.1). To do that, we use the notion of metric entropy. Although metric entropy was defined in Chapter 5, for convenience we repeat here the definition for the specific case at hand.

Definition Let $\epsilon>0$ be given. The metric entropy of a set $B\subset\mathcal{M}_1(\Sigma)$, denoted $N(\epsilon,B)$, is defined by

$$N(\epsilon,B)\stackrel{\triangle}{=}\inf\{n|\exists y_1,\ldots,y_n\in\mathcal{M}_1(\Sigma)\quad\text{s.t.}\quad B\subset\cup_1^n B(y_i,\epsilon)\} \tag{9.3}$$

where $B(y,\epsilon)$ denotes a ball of radius ϵ (in the Prohorov metric) around y.

Similarly, for any given ϵ, denote by $N^\Sigma(\epsilon)$ the metric entropy of Σ, i.e.

$$N^\Sigma(\epsilon)\stackrel{\triangle}{=}\inf\{n|\exists\tilde{y}_1,\ldots,\tilde{y}_n\in\Sigma\quad\text{s.t.}\quad\Sigma\subset\cup_1^n B(\tilde{y}_i,\epsilon)\}. \tag{9.4}$$

where $B(\tilde{y}_i,\epsilon)$ are taken in the metric corresponding to Σ.

We claim now:

Lemma 9.1

$$N(\epsilon,\mathcal{M}_1(\Sigma))\leq 2\left(\frac{e}{\epsilon}\right)^{(N^\Sigma(\epsilon))}\stackrel{\triangle}{=}\bar{N}(\epsilon,\mathcal{M}_1(\Sigma)) \tag{9.5}$$

Proof: In order to prove the lemma, we will explicitly construct an ϵ-cover of $\mathcal{M}_1(\Sigma)$ with $\bar{N}(\epsilon,\mathcal{M}_1(\Sigma))$ elements.

Let $\tilde{y}_1,\ldots,\tilde{y}_{N^\Sigma(\epsilon)}$ be the centers of a set of ϵ balls in Σ which create the cover

$N^{\Sigma}(\epsilon)$ in (9.4). Let $\delta_i \stackrel{\Delta}{=} \delta_{\tilde{y}_i}$, i.e. the distribution concentrated at $\tilde{y}_i$, and let

$$\mu_i^j \stackrel{\Delta}{=} j \cdot \left(\frac{\epsilon}{N^{\Sigma}(\epsilon)}\right) \cdot \delta_i \quad , \quad j = 0, 1, \ldots, \frac{N^{\Sigma}(\epsilon)}{\epsilon}$$

Define $Y \stackrel{\Delta}{=} \{y \in \mathcal{M}_1(\Sigma) : \exists\, (i_1, j_1) \cdots (i_k, j_k) \text{ s.t. } y = \sum_{\alpha=1}^{k} \mu_{i_\alpha}^{j_\alpha}\}$. Note that Y is a finite set, for it includes at most $\left(\frac{N^{\Sigma}(\epsilon)}{\epsilon} + 1\right)^{N^{\Sigma}(\epsilon)}$ members. Also, note that Y is an ϵ-cover of $\mathcal{M}_1(\Sigma)$, i.e. for any $\mu \in \mathcal{M}_1(\Sigma)$ there exists a $y \in Y$ such that for any open set $C \subset \Sigma$, $\mu(C) \leq y(C^{\epsilon}) + \epsilon$. To see that, choose as y the following approximation to μ:

Let $i_\alpha = \alpha$, $\alpha = 1, \ldots, N^{\Sigma}(\epsilon) - 1$, and choose $j_\alpha = \lfloor \mu(B(\tilde{y}_\alpha, \epsilon) \backslash \left(\cup_{k=1}^{\alpha-1} B(\tilde{y}_k, \epsilon)\right) \rfloor$, where by $\lfloor \times \rfloor$ we mean the closest approximation to $\times$ on the $\frac{N^{\Sigma}(\epsilon)}{\epsilon}$ j-net from below. Finally, let $j_{N^{\Sigma}(\epsilon)} \stackrel{\Delta}{=} 1 - \sum_{\alpha=1}^{N^{\Sigma}(\epsilon)-1} j_\alpha$. Take now $y = \sum_{\alpha=1}^{N^{\Sigma}(\epsilon)} \mu_{i_\alpha}^{j_\alpha}$. It follows that y is a probability measure based on a finite number of atoms and, furthermore, $d(y, \mu) < \epsilon$. We need therefore only to estimate the cardinality of the set Y, denoted $|Y|$. Note that $|Y|$ is just the number of vectors $(j_1, \ldots, j_{N^{\Sigma}(\epsilon)})$ such that $\sum_{i=1}^{N^{\Sigma}(\epsilon)} j_i = 1$ and $j_i \in \{0, \frac{\epsilon}{N(\epsilon)}, \frac{\alpha\epsilon}{N(\epsilon)}, \ldots, 1\}$.

It follows that

$$\begin{aligned} |Y| &\leq \left(\tfrac{N^{\Sigma}(\epsilon)}{\epsilon} + 1\right)^{N^{\Sigma}(\epsilon)} \int_0^1 \cdots \int_0^{x_3} \int_0^{x_2} dx_1 \ldots dx_{N^{\Sigma}(\epsilon)} \\ &= \left(\tfrac{N^{\Sigma}(\epsilon)}{\epsilon} + 1\right)^{N^{\Sigma}(\epsilon)} \cdot \tfrac{1}{N^{\Sigma}(\epsilon)!} \end{aligned} \tag{9.6}$$

However, by Stirling's formula

$$\log\left(N^{\Sigma}(\epsilon)!\right) \geq N^{\Sigma}(\epsilon) \log\, N^{\Sigma}(\epsilon) - N^{\Sigma}(\epsilon) \tag{9.7}$$

Substituting (9.7) into (9.6), one has

$$|Y| \leq \left(\frac{N^{\Sigma}(\epsilon)}{\epsilon} + 1\right)^{N^{\Sigma}(\epsilon)} \epsilon^{N^{\Sigma}(\epsilon)} \cdot \frac{1}{\left(N^{\Sigma}(\epsilon)\right)^{N^{\Sigma}(\epsilon)}} \tag{9.8}$$

which implies that

$$\begin{aligned} N(\epsilon, \mathcal{M}_1(\Sigma)) &\leq \left(\tfrac{1}{\epsilon}\left(1+\tfrac{\epsilon}{N^\Sigma(\epsilon)}\right)\right)^{N^\Sigma(\epsilon)} e^{N^\Sigma(\epsilon)} \\ &= \left(\tfrac{e}{\epsilon}\right)^{N^\Sigma(\epsilon)} \left(1+\tfrac{\epsilon}{N^\Sigma(\epsilon)}\right)^{N^\Sigma(\epsilon)} \leq 2\left(\tfrac{e}{\epsilon}\right)^{N^\Sigma(\epsilon)} = \bar{N}(\epsilon, \mathcal{M}_1(\Sigma)) \end{aligned}$$

□

For completeness, we now prove a lower bound for the metric entropy of $\mathcal{M}_1(\Sigma)$ with respect to the Prohorov metric. This lower bound exhibits a behavior similar to the upper bound of (9.5), so that these bounds cannot be much improved. In the proof below, $M(\epsilon, Y, d)$ denotes the ϵ-capacity (or packing number) of the space Y with respect to the metric d. That is, $M(\epsilon, Y, d)$ represents the maximum number of non-overlapping balls of *diameter* ϵ with respect to the metric d that can be packed in Y. The well known relationship

$$N(2\epsilon, Y, d) \leq M(2\epsilon, Y, d) \leq N(\epsilon, Y, d)$$

between covering numbers and packing numbers is easy to show and is used in the proof below.

Lemma 9.2 *Let Σ be compact Polish space with metric d, and let $\mathcal{M}^1(\Sigma)$ denote the set of probability measures on X with the Prohorov metric ρ. Then*

$$N(\epsilon, \mathcal{M}^1(\Sigma), \rho) \geq (\frac{1}{8\epsilon})^{N(2\epsilon, \Sigma, d)}$$

Proof: First, we can find $N = N(\epsilon, \Sigma, d)$ points $x_1, \ldots, x_N$ which are pairwise greater than or equal to ϵ apart. Each measure supported on these N points corresponds to a point in $\mathbf{R}^N$ in the natural way. Then, the set of all probability measures supported on $x_1, \ldots, x_N$ corresponds to the simplex S^N in $\mathbf{R}^N$.

Now, let p, q be points on the simplex S^N and suppose that $d_{\ell^1}(p, q) \geq 2\epsilon$. Then on some subset $G \subset \{1, \ldots, N\}$ of coordinates either $\sum_{i \in G} p_i \leq \sum_{i \in G} q_i + \epsilon$ or $\sum_{i \in G} p_i \leq \sum_{i \in G} q_i + \epsilon$. Then, considered as probability measures on Σ, $\rho(p, q) \geq \epsilon$ since there is a closed set $F \subset \Sigma$, namely $F = \{x_i \mid i \in G\}$, for which either $p(F) \geq q(F^\epsilon) + \epsilon$ or $q(F) \geq p(F^\epsilon) + \epsilon$. Hence,

$$N(\epsilon/2, \mathcal{M}^1(\Sigma), \rho) \geq M(\epsilon, \mathcal{M}^1(\Sigma), \rho) \geq M(2\epsilon, S^N, d_{\ell^1}) \geq N(2\epsilon, S^N, d_{\ell^1})$$

Finally, to get a lower bound on $N(2\epsilon, S^N, d_{\ell^1})$, we note that the volume of the

simplex S^N is $1/N!$ while the volume of an ℓ^1 ball of radius 2ϵ is $(4\epsilon)^N/N!$ which implies $N(2\epsilon, S^N, d_{\ell^1}) \geq (1/4\epsilon)^N$. Thus, $N(\epsilon/2, \mathcal{M}^1(\Sigma), \rho) \geq (1/4\epsilon)^{N(\epsilon,\Sigma,d)}$ or equivalently $N(\epsilon, \mathcal{M}^1(\Sigma), \rho) \geq (1/8\epsilon)^{N(2\epsilon,\Sigma,d)}$.

□

The existence of the bound $\bar{N}$ permits us to mimic the computation in [71] for the case in hand. Indeed, a crucial step needed in [71] was bounding the probability of complements of balls, for all n, uniformly over all measures, as follows:

Theorem 9.1

$$P(\mu_n \in B(\mu, \delta)^c) \leq \bar{N}\left(\frac{\delta}{2}, \mathcal{M}_1(\Sigma)\right) e^{-n\left(\frac{\delta}{2}\right)^2}$$

Proof: The proof follows the standard Chebycheff bound technique, without taking n limits as in the large deviation framework. Indeed,

$$P(\mu_n \in B(\mu, \delta)^c) \leq \bar{N}\left(\frac{\delta}{2}, \mathcal{M}_1(\Sigma)\right) \cdot \sup_{y \in \mathcal{M}_1(\Sigma), d(y,\mu) \geq \delta} P(\mu_n \in B(y, \frac{\delta}{2}))$$

Let P_n denote the law of the random variable μ_n. The second term can be bounded as follows. By the Chebycheff bound, it follows that for any $\theta \in C_b(\Sigma)$,

$$\begin{aligned}
\sup_{y \in \mathcal{M}_1(\Sigma), d(y,\mu) \geq \delta} & P(\mu_n \in B(y, \frac{\delta}{2})) \\
&\leq \sup_{y \in \mathcal{M}_1(\Sigma), d(y,\mu) \geq \delta} \int_{B(y, \frac{\delta}{2})} e^{n<\theta,\nu>} e^{-n<\theta,\nu>} dP_n(\nu) \\
&\leq \exp\left(-n \sup_{\theta \in C_b(\Sigma)} \inf_{\nu \in B(y, \frac{\delta}{2}), d(y,\mu) \geq \delta} (<\theta, \nu> - \frac{1}{n} \log E_{P_n}(e^{n<\theta,\nu>})\right) \\
&= \exp\left(-n \inf_{\nu \in B(y, \frac{\delta}{2}), d(y,\mu) \geq \delta} \sup_{\theta \in C_b(\Sigma)} (<\theta, \nu> - \frac{1}{n} \log E_{P_n}(e^{n<\theta,\nu>})\right) \\
&= \exp\left(-n \inf_{\nu \in B(y, \frac{\delta}{2}), d(y,\mu) \geq \delta} H(\nu|\mu)\right) \\
&\leq \exp\left(-n \inf_{\nu \in B(\mu, \frac{\delta}{2})^c} H(\nu|\mu)\right) \\
&\leq e^{-n\left(\frac{\delta}{4}\right)^2} \qquad (9.9)
\end{aligned}$$

where the first equality in (9.9) follows from the min-max theorem for convex compact sets (c.f. [39]), the second equality follows by ([32], Lemma 3.2.13), and the last

inequality from the fact that ([32], Exercise 3.2.24) for any $\theta \in B(\mu, \delta/2)^c$,

$$\frac{\delta}{2} \leq d(\theta, \mu) \leq \|\theta - \mu\|_{var} \leq 2H^{1/2}(\theta|\mu)$$

□

Corollary 9.1 *Let $B_m \subset \mathcal{M}_1(\Sigma)$ be a measurable set such that $\mu \in B_m$. Let B_m^δ denote an open set such that $d(B_m, (B_m^\delta)^c) \geq \delta$. Then*

$$P_\mu\left(\mu_n \in (B_m^\delta)^c\right) \leq \bar{N}\left(\frac{\delta}{2}, \mathcal{M}_1(\Sigma)\right) e^{-n\left(\frac{\epsilon}{4}\right)^2} \tag{9.10}$$

We can turn now to the proposed classification algorithm, much as in [71]. Define

$$\alpha(m) = \frac{4}{\epsilon(m)}\left[2\log m + \log 2 + N^\Sigma\left(\sqrt{\frac{\epsilon_m}{2}}\right)\left(1 - \log\sqrt{\frac{\epsilon_m}{2}}\right)\right] \tag{9.11}$$

and note that, for all $m > m_0(\mu)$,

$$P_\mu(\mu_{\alpha(m)} \in C_m^c) \leq \frac{1}{m^2} \tag{9.12}$$

Let

$$\beta(m) = \sum_{i=1}^{m} \alpha(i), \quad \beta(0) = 0. \tag{9.13}$$

For any input sequence $x_1, x_2, \ldots$, form the subsequences

$$X^m \hat{=} (x_{\beta(m-1)}, \cdots, x_{\beta(m)-1}).$$

The endpoints of these subsequences X^m form a parsing of the original sequence $x_1, x_2, \ldots$. Let μ_{X^m} denote the empirical measure of the sequence X^m. At the end of each parsing, make a decision of whether $\mu_T \in A$ according to whether $\mu_{X^m} \in C_m$ or not. Between parsings, don't change the decision.

We now claim:

Theorem 9.2 *The decision rule defined by the parsing $\beta(m)$ as above is successful.*

Proof: The proof is identical to the proof of Theorem 8.1 in Chapter 8 (or Theorem 1 in [71]).

□

9.3 Classification Among a Countable Number of Sets

In this section, we refine the decision rule to allow for classification among a countable number of sets. Specifically, if $A_1, A_2, \ldots$ are a countable number of subsets of $\mathcal{M}^1(\Sigma)$ we are interested in deciding to which of the A_i the unknown measure μ belongs. The only assumption we make on the A_i is that each A_i satisfies the structural assumption (A-1). The A_i are not required to be either disjoint or nested, although these special cases are most commonly of interest in applications. In general, after a finite number of observations one cannot expect to determine the membership status of μ in all of the A_i. However, we will show that for all μ except in a set of G-measure zero in $\mathcal{M}^1(\Sigma)$ there is a decision procedure that a.s. will eventually determine the membership of μ in any finite subset of the A_i. In the special cases of disjoint or nested A_i, the membership status of μ in any of the countable A_i is completely determined by membership in some finite subset. Hence, in these cases, except for μ in a set of G-measure zero the membership of μ in all the A_i will a.s. be eventually determined.

We modify our previous decision rule as follows. The observations $x_1, x_2, \ldots$ will still be parsed into increasingly larger blocks in a manner to be defined below. However, now, at the end of the m-th block, we will make a decision as to the membership of μ in the first m of the A_i. The decisions of whether μ belongs to $A_1, \ldots, A_m$ are made separately for each A_i using a procedure similar to that of the previous section.

Specifically, for each A_i let $B_{i,m}$ be a sequence of closed sets, $C_{i,m}$ a sequence of open sets and $\epsilon_i(m) \to_{m\to\infty} 0$ a positive sequence satisfying the requirements of the structural assumption (A-1). From the same considerations that led to (9.12), for

$$\alpha_i(m) = \frac{4}{\epsilon_i(m)} \left[2 \log m + \log 2 + N^{\Sigma} \left(\sqrt{\epsilon_i(m)/2} \right) \left(1 - \log \sqrt{\epsilon_i(m)/2} \right) \right] \tag{9.14}$$

we have

$$P_\mu(\mu_{\alpha_i(m)} \in C_{i,m}^c) \leq \frac{1}{m^2} \tag{9.15}$$

As before, the observation sequence $x_1, x_2, \ldots$ will be parsed into non-overlapping blocks

$$X^m = (x_{\beta(m-1)+1}, \ldots, x_{\beta(m)}) \tag{9.16}$$

where the $\beta(m)$ are defined below. At the end of the m-th block, a decision will be made about the membership of μ in $A_1, \ldots, A_m$. This decision will be made

separately for each $i = 1, \ldots, m$ using the observation sequence X^m exactly as before. That is, at the end of the parsing sequence X^m, for $i = 1, \ldots, m$ decide that $\mu \in A_i$ according to whether or not $\mu_{X^m} \in C_{i,m}$, and don't change the decision except at the end of a parsing sequence. We define the parsing sequence $\beta(m)$ by $\beta(0) = 0$ and $\beta(m) - \beta(m-1) = \sup_{1 \le i \le m} \alpha_i(m)$ or equivalently

$$\beta(m) = \sum_{k=1}^{m} \sup_{1 \le i \le k} \alpha_i(k), \qquad \beta(0) = 0 \tag{9.17}$$

For this decision rule we have the following theorem.

Theorem 9.3 *Let $A_i \subset \mathcal{M}^1(\Sigma)$ for $i = 1, 2, \ldots$ satisfy the structural assumption (A-1). There is a set $N \subset M_1(\Sigma)$ of G-measure zero such that for every $\mu \in \mathcal{M}^1(\Sigma) \setminus N$ and every $k < \infty$ the decision rule will make (a.s.) only a finite number of mistakes in deciding the membership of μ in $A_1, \ldots, A_k$. That is, given any $\mu \in \mathcal{M}^1(\Sigma) \setminus N$, for a.e. ω there exists $m(\omega) = m(\omega, \mu, k)$ such that for all $m > m(\omega)$ the algorithm makes a correct decision as to whether $\mu \in A_i$ or $\mu \in A_i^c$ for $i = 1, \ldots, k$.*

Proof: Let

$$N_i = \bigcap_{n=1}^{\infty} \bigcup_{m=n}^{\infty} C_{i,m} \setminus A_i \tag{9.18}$$

and let

$$N = \bigcup_{i=1}^{\infty} N_i \tag{9.19}$$

Then from the assumption (A-1) it follows that the G-measure of each N_i is zero, and so the G-measure of N is also zero.

Now, let $\mu \in \mathcal{M}^1(\Sigma) \setminus N$, and let $k < \infty$. For each $i = 1, \ldots, k$, there exists $m_i < \infty$ such that if $\mu \in A_i$ then $\mu \in S_{i,m}$ for all $m > m_i$, while if $\mu \in A_i^c$ then $\mu \in C_{i,m}^c$ for all $m > m_i$ (since $\mu \notin N_i$). Recall that at the end of the parsing sequence X^m, the algorithm decides $\mu \in A_i$ iff $\mu_{X^m} \in C_{i,m}$, so that if $\mu \in A_i$ then an error is made about membership in A_i iff $\mu_{X^m} \notin C_{i,m}$ while if $\mu \notin A_i$ an error is made iff $\mu_{X^m} \in C_{i,m}$. If $\mu \in A_i$ then using Corollary 9.1 and the fact that $d(B_{i,m}, C_{i,m}^c)^2 \ge 2\epsilon_i(m)$, we have that the probability of making an incorrect decision is less than $1/m^2$ for $m > m_0^i(\mu)$. On the other hand, if $\mu \in A_i^c$ then since $B_{i,m} \subset ((C_{i,m}^c)^{\sqrt{2\epsilon(m)}})^c$ we also have probability of error less than $1/m^2$ for $m > m_0^i$ (again using Corollary 9.1 and the expression for $\alpha(m)$). Hence, for $m > m_0 = \max(m_0^i, \ldots, m_0^k)$ the probability of making an error about the membership of μ in *any* of $A_i, \ldots, A_k$ is less than k/m^2.

Then

$$\sum_{m=1}^{\infty} \text{Prob}\{\text{error in any } A_i \text{ on } m\text{-th parsing}\} \leq m_0 + k \sum_{m=m_0+1}^{\infty} \frac{1}{m^2} < \infty$$

so that the theorem follows by the Borel-Cantelli Lemma.

□

Note that if one also wants to make a correct decision after some finite time whether or not μ is in *any* of the A_i for $i = 1, 2, \ldots$ then the decision procedure can be easily modified to handle this. Specifically, it is easy to show that sets satisfying the structural assumption are closed under countable union. Hence, one could include in the hypothesis testing the set $A_0 = \cup_{i=1}^{\infty} A_i$, so that after some finite time a correct decision would be made about the membership of $\mu \in A_0$.

Also, it is worthwhile to note that if the A_i have more structure then some improvements can be made. For example, if the membership status of μ in A_i for $i = 1, 2, \ldots$ is determined by its membership status in some finite number of the A_i then a correct decision regarding the membership of μ in all of the A_i can be guaranteed (a.s.) after some finite time (depending on μ). This is the case for disjoint or nested A_i, which may be of particular interest in some applications. For these cases, by letting $A_0 = \cup_{i=1}^{\infty} A_i$ and running the decision rule on $A_0, A_1, A_2, \ldots$ as mentioned above, we have the following corollary of Theorem 9.3.

Corollary 9.2 *Let $A_i \subset \mathcal{M}^1(\Sigma)$ for $i = 1, 2, \ldots$ satisfy the structural assumption (A-1) and suppose the A_i are either disjoint or nested. There is a set $N \subset M_1(\Sigma)$ of G-measure zero such that for every $\mu \in \mathcal{M}^1(\Sigma) \setminus N$ the decision rule will make (a.s.) only a finite number of mistakes in deciding the membership of μ in all of the A_i. That is, given any $\mu \in \mathcal{M}^1(\Sigma) \setminus N$, for a.e. ω there exists $m(\omega) = m(\omega, \mu)$ such that for all $m > m(\omega)$ the algorithm makes a correct decision as to whether $\mu \in A_i$ for all $i = 1, 2, \ldots$.*

It is worthwhile to note that the results of this section may be used also in the case that Σ is locally compact but not compact. In that case, one may first intersect the A_i with compact sets K_m which sequentially approximate Σ and then use $m(n) \to \infty$. We do not consider this issue here.

We conclude this section with an example from the area of density estimation. Let $\Sigma = [0, 1]$ and assume that $x_1, \ldots, x_n$ are i.i.d. and drawn from a distribution with

law μ_θ, $\theta \in \Theta$. When some structure is given on the set $\mathcal{F} = \bigcup_{\theta\in\Theta} \mu_\theta$, there exists a large body of literature which enables one to obtain estimates of the error after n observations (e.g, see [33, 34, 50] and references contained therein). All these results assume an a-priori structure, e.g. a bound on the L^2 norm of the density $f_\theta = \frac{d\mu_\theta}{dx}$. If such information is not given a-priori, it may be helpful to design a test to check for this information and thus to be able to estimate eventually whether the distribution belongs to a nice set and if so to apply the error estimates alluded to above.

Thus, let

$$A_i = \{\mu \in \mathcal{M}_1(\Sigma) : \int_0^1 (\frac{d\mu(x)}{dx})^2 \leq i\}.$$

Note that the sets A_i are closed w.r.t. the Prohorov metric and therefore they satisfy the structural assumption A-1). Moreover, they are nested and thus Corollary 9.2 may be applied to yield a decision rule which will asymptotically decide correctly on the appropriate class of densities. This idea is in the spirit of a suggestion by Cover [27], and the case where the A_i consist of single points (i.e., each A_i contains a single probability measure) is related to a model considered by Barron and Cover [12, 13]. It would be interesting to make a more formal comparison between the two models.

9.4 Applications to Order Determination of Markov Processes

In this section, we extend the model of the observations to allow for a Markov dependence in the observation. Specifically, let Σ be a compact Polish space as before, but assume that the observations $x_1, \ldots, x_n$ are the outcome of a Markov chain of order j, i.e.

$$\text{Prob}(x_k \in A | x_{k-1}, x_{k-2}, \ldots, x_1) = \pi^j(x_k \in A | x_{k-1}, x_{k-2}, \ldots, x_{k-j})$$

where A is a Borel measurable subset of Σ and $k > j$. We assume that j is unknown, and our task is to decide (correctly) on the order j. In order to avoid technicalities, we assume that all Markov chains involved are ergodic and satisfy a uniformity condition w.r.t. the initial conditions (c.f below), and therefore there exists a stationary measure $P_{\pi^j} \in \mathcal{M}_1(\Sigma^j)$ such that for any measurable set A in Σ^j,

$$P_{\pi^j}(A) = \int_{\{x_{2j},x_{2j-1},\ldots,x_{j+1}\}\in A} d\pi^j(x_{2j}|x_{2j-1},\ldots,x_j) \cdots d\pi^j(x_{j+1}|x_j,\ldots,x_1)dP_{\pi^j}(x_j,\ldots,x_1) \quad (9.20)$$

This problem has already been considered in the literature. For a discussion of available results we refer the reader to [83],[52]. Most of the results in the literature are either for the discrete alpha-bet setup or for the case of a linear model $x_{n+k} = \sum_{i=1}^{k} a_i x_{n+i} + \epsilon_{n+k}$, where ϵ_n is a white sequence and a_i are deterministic, unknown constants. Here, we consider the general setup and show how a strongly consistent decision rule may be constructed based on the general paradigm of this paper. Towards this end, we need to extend the basic estimates of Section 2 to the Markov case, as follows:

Let $\Omega = \Sigma^Z$, define x_i to be the coordinate map $x_i(\omega) = \omega_i$, and let the shift operator be defined by $x_i(T\omega) = x_{i+1}(\omega)$. Define the k-th order empirical measure on $\mathcal{M}_1(\Sigma^k)$ by

$$\mu_n^k = \frac{1}{n}\sum_{i=1}^{n} \delta_{x_1(T^i\omega),x_2(T^i\omega),\ldots,x_k(T^i\omega)}$$

As before, we endow $\mathcal{M}_1(\Sigma^k)$ with the Prohorov topology, and recall that under the structural assumptions on the Markov chain described above, a large deviations principle holds for the empirical measure μ_n^{j+1}, viz. for any set $A \subset \mathcal{M}_1(\Sigma^{j+1})$ a large deviations statement of the form (9.1) holds, with the relative entropy $H(\nu|\mu)$ being replaced by

$$H_j(\theta|\mu) = \begin{cases} \int_{\Sigma^j} d\theta(y_1,\ldots,y_{j+1}) \log \dfrac{d\theta(y_{j+1}|y_j,\ldots,y_1)}{d\mu(y_{j+1})|y_j,\ldots,y_1)} \\ \qquad \text{if } d\theta(\cdot|y_j,\ldots,y_1) << d\mu(\cdot|y_j,\ldots,y_1) \\ \infty \qquad\qquad \text{otherwise} \end{cases} \quad (9.21)$$

For any measure $\mu(x_1,\ldots,x_k) \in \mathcal{M}_1(\Sigma^k)$, denote by μ_i the marginal defined by

$$\mu_i(\{x_1,\ldots,x_i\} \in A) \stackrel{\triangle}{=} \mu(\{x_1,\ldots,x_i\} \in A, \{x_{i+1},\ldots,x_k\} \in \Sigma^{k-i})$$

and by $\mu_{i|i-1,\ldots,i-t}$ the regular conditional probability $\mu(x_i|x_{i-1},\ldots,x_{i-t})$. Define the measure $\mu_{i-k} \otimes \mu_{i-(k-1)|i-k,\ldots,1} \otimes \cdots \otimes \mu_{i|i-1,\ldots,2} \in \mathcal{M}_1(\Sigma^i)$ as the measure which, for

any measurable set $A \subset \Sigma^i$,

$$\mu_{i-k} \otimes \mu_{i-(k-1)|i-k,\ldots,1} \otimes \cdots \otimes \mu_{i|i-1,\ldots,2}(A) = \int_A d\mu_{i-k}(x_1,\ldots,x_{i-k}) d\mu_{i-(k-1)|i-k,\ldots,1}(x_{i-(k-1)}|x_{i-k},\ldots,x_1) \cdots d\mu_{i|i-1,\ldots,2}(x_i|x_{i-1},\ldots,x_2) \quad (9.22)$$

Let π^j be a given j-th order Markov kernel, P_{π^j} its corresponding stationary measure, and denote by Pr^{π^j} the stationary measure on Ω generated by this kernel. Assume that the empirical measures μ_n^{j+k}, $k = 2, 3, \ldots$ are formed from a Markov sequence generated by this kernel. In order to compute explicitly the sequence of decision rules as in the i.i.d. case, we need to derive the analog of Theorem 9.1 given below.

Theorem 9.4

$$\begin{aligned} Pr^{\pi^j}\left[\mu_n^{j+k} \notin B\left((\mu_n^{j+2})_j \otimes \pi^j \otimes \cdots \otimes \pi^j, \delta\right)\right] &\leq \bar{N}\left(\frac{\delta}{2}, \mathcal{M}_1(\Sigma^{j+k})\right) e^{-n(\frac{\delta}{16})^2} + \cdots + \bar{N}\left(\frac{\delta}{2^k}, \mathcal{M}_1(\Sigma^{j+1})\right) e^{-n(\frac{\delta}{2^{k+3}})^2} \\ &\leq k\bar{N}\left(\frac{\delta}{2^k}, \mathcal{M}_1(\Sigma^{j+k})\right) e^{-n(\frac{\delta}{2^{k+3}})^2} \end{aligned}$$

Proof: We prove the theorem first for the case $k = 2$. The general case follows by induction.

$$\begin{aligned} Pr^{\pi^j}\left[\mu_n^{j+2} \notin B\left((\mu_n^{j+2})_j \otimes \pi^j \otimes \pi^j, \delta\right)\right] &\leq Pr^{\pi^j}\left[d(\mu_n^{j+2}, (\mu_n^{j+2})_j \otimes \pi^j \otimes \pi^j) \geq \delta, d((\mu_n^{j+2})_{j+1}, (\mu_n^{j+2})_j \otimes \pi^j) < \delta/4\right] \\ &\quad + Pr^{\pi^j}\left[d((\mu_n^{j+2})_{j+1}, (\mu_n^{j+2})_j \otimes \pi^j) \geq \delta/4\right] \\ &\triangleq P_1 + P_2 \end{aligned} \quad (9.23)$$

By repeating the argument in (9.9),

$$\frac{P_1}{\bar{N}\left(\frac{\delta}{2}, \mathcal{M}_1(\Sigma^{j+2})\right)} \leq \sup_{y \in \mathcal{M}_1(\Sigma^{j+2})} P[\mu_n^{j+2} \in B(y, \frac{\delta}{2}), (\mu_n^{j+2})_{j+1} \in B(P_{\pi^j} \otimes \pi^j, \frac{\delta}{4}), d((\mu_n^{j+2})_j \otimes \pi^j \otimes \pi^j, y) \geq \delta] \quad (9.24)$$

Therefore, by the Chebycheff bound, denoting by P_n the law of the random variable

μ_n^{j+2}, it follows that for any $\theta \in C_b(\Sigma^{j+2})$,

$$\frac{P_1}{\bar{N}\left(\frac{\delta}{2}, \mathcal{M}_1(\Sigma^{j+2})\right)} \tag{9.25}$$

$$\leq \sup_{y \in \mathcal{M}_1(\Sigma^{j+2})} \int_{B(y,\frac{\delta}{2})} e^{n<\theta,\nu>} e^{-n<\theta,\nu>} \cdot$$

$$1_{(\mu_n^{j+2})_{j+1} \in B((\mu_n^{j+2})_j \otimes \pi^j, \delta/4), y \in B((\mu_n^{j+2})_j \otimes \pi^j \otimes \pi^j, \delta))} dP_n(\nu)$$

$$\leq \exp\left[-n \sup_{\theta \in C_b(\Sigma^{j+2})} \inf_{\substack{\nu \in B(y,\frac{\delta}{2}) \bigcap B(\nu_j \otimes \pi^j, \frac{\delta}{4}) \\ d(y, \nu_j \otimes \pi^j \otimes \pi^j) \geq \delta}} \left(< \theta, \nu > - \frac{1}{n} \log E_{P_n}(e^{n<\theta,\nu>})\right)\right]$$

$$\leq \exp\left[-n \inf_{\substack{\nu \in B(y,\frac{\delta}{2}) \bigcap B(\nu_j \otimes \pi^j, \frac{\delta}{4}) \\ d(y, \nu_j \otimes \pi^j \otimes \pi^j) \geq \delta}} \sup_{\theta \in C_b(\Sigma^{j+2})} \left(< \theta, \nu > - \frac{1}{n} \log E_{P_n}(e^{n<\theta,\nu>})\right)\right]$$

However,

$$\sup_{\theta \in C_b(\Sigma^{j+2})} \left(< \theta, \nu > - \frac{1}{n} \log E_{P_n}(e^{n<\theta,\nu>})\right)$$

$$= \sup_{\theta \in C_b(\Sigma^{j+2})} \left(< \theta, \nu > - \frac{1}{n} \log E(e^{\theta(x_1,\ldots,x_{j+2}) + \cdots + \theta(x_{n-j-2},\ldots,x_n)})\right) \tag{9.26}$$

$$= \sup_{\theta \in \mathcal{B}(\Sigma^{j+2})} \left(< \theta, \nu > - \frac{1}{n} \log E(e^{\theta(x_1,\ldots,x_{j+2}) + \cdots + \theta(x_{n-j-2},\ldots,x_n)})\right)$$

where $\mathcal{B}(\Sigma^{j+2})$ denotes the space of bounded measurable functions on Σ^{j+2} and the last equality follows from dominated convergence. We assume now that ν is absolutely continuous w.r.t. $\nu_{j+1} \otimes \pi^j$ and that the resulting Radon-Nikodym derivative is uniformly bounded from above and below (these assumptions may be relaxed exactly as in [32], pg. 69). In this case, we may take $\log \theta$ in (9.26) as this Radon-Nikodym derivative, i.e. $\theta(x_1, \ldots, x_{j+2}) = \log \frac{d\nu}{d\nu_{j+1} \otimes \pi^j}$ to obtain:

$$\sup_{\theta \in \mathcal{B}(\Sigma^{j+2})} \left(< \theta, \nu > - \frac{1}{n} \log E(e^{\theta(x_1,\ldots,x_{j+2}) + \cdots + \theta(x_{n-j-1},\ldots,x_n)})\right)$$

$$\geq H(\nu | \nu_{j+1} \otimes \pi^j) - \frac{1}{n} \log \int_{\Sigma^{j+2}} d\nu(x_n | x_{n-1}, \ldots, x_{n-j-1})$$

$$\cdots d\nu(x_{j+2} | x_{j+1}, \ldots, x_1) d\nu_{j+1}(x_{j+1}, \ldots, x_1)$$

$$= H(\nu | \nu_{j+1} \otimes \pi^j) \tag{9.27}$$

Substituting (9.27) in (9.25) and recalling the inequality

$$2H^{1/2}(\theta|\mu) \geq d(\theta,\mu)$$

one obtains

$$\begin{aligned}
&\frac{P_1}{\bar{N}\left(\frac{\delta}{2},\mathcal{M}_1(\Sigma^{j+2})\right)} \\
&\leq \exp\left(-n \inf_{\substack{\nu\in B(y,\frac{\delta}{2})\bigcap B((\mu_n^{j+2})_j\otimes\pi^j,\frac{\delta}{4}) \\ d(y,(\mu_n^{j+2})_j\otimes\pi^j\otimes\pi^j)\geq\delta}} H(\nu|\nu_{j+1}\otimes\pi^j)\right) \\
&\leq \exp\left(-n \inf_{\substack{\nu\in B(y,\frac{\delta}{2})\bigcap B((\mu_n^{j+2})_j\otimes\pi^j,\frac{\delta}{4}) \\ d(y,(\mu_n^{j+2})_j\otimes\pi^j\otimes\pi^j)\geq\delta}} d(\nu,\nu_{j+1}\otimes\pi^j)^2/4\right) \\
&\leq \exp[-n \inf_{\substack{\nu\in B(y,\frac{\delta}{2})\bigcap B((\mu_n^{j+2})_j\otimes\pi^j,\frac{\delta}{4}) \\ d(y,(\mu_n^{j+2})_j\otimes\pi^j\otimes\pi^j)\geq\delta}} \\
&\left([d(\nu,(\mu_n^{j+2})_j\otimes\pi^j\otimes\pi^j) - d(\nu_{j+1}\otimes\pi^j,(\mu_n^{j+2})_j\otimes\pi^j\otimes\pi^j)]\vee 0\right)^2/4] \\
&\leq \exp\left(-n\,(\delta/16)^2\right) \qquad (9.28)
\end{aligned}$$

Similarly,

$$\frac{P_2}{\bar{N}\left(\frac{\delta}{4},\mathcal{M}_1(\Sigma^{j+1})\right)} \leq \exp\left(-n\,(\delta/32)^2\right) \qquad (9.29)$$

Substituting (9.28), and (9.29) in (9.23) yields the theorem for $k = 2$. The general case is similar and follows by induction.

□

We are now ready to return to the order determination problem described in the beginning of this section. Since the set up here differs slightly from the one described in the previous section, we repeat here the main definitions.

Let $A_i \subset \mathcal{M}_1(\Omega)$, $i = 0, 1, \ldots$, be the set of stationary measures generated by Markov chains of order i (with $i = 0$ denoting the i.i.d. case), i.e. for $i = 0, 1, 2, \ldots$,

$\mu \in A_i \iff (\mu)_{i+k} = (\mu)_i \otimes \pi^i \otimes \cdots \otimes \pi^i$ for some Markov kernel π^i and for all $k = 1, 2, \ldots$.

Note that the sets A_i are closed.

A natural candidate for covering sets $C_{i,m}$ are δ_m blow ups of the A_i. That is, define

$$C_{i,m} = \{\nu \in \mathcal{M}_1(\Sigma^{i+m}) : \ d(\nu, (\nu)_i \otimes \pi^i \otimes \cdots \otimes \pi^i) < \delta_m \text{ for some Markov kernel } \pi^i\}$$

It is clear that $C_{i,m}$ is open, and also that

$$\bigcap_{n=1}^{\infty} \bigcup_{m=n}^{\infty} (C_{i,m} \backslash A_i) = \emptyset.$$

Therefore, by using Theorem 9.4 and the procedure described in Theorem 9.3, the sets $C_{i,m}$ are candidates for building a decision rule which, a.s., decides correctly in finite time whether the given observation sequence was generated by a Markov chain of order i. In order to be able to do so, we need only to check that the complements of the sets $C_{i,m}$, which are closed, have the property that they may be covered by small enough spheres (say, $\delta_m/4$ spheres), such that the union of those spheres belongs to the complement of some $C_{i,m}$. This can be seen by using the following lemma:

Lemma 9.3 *Let $\nu, \nu' \in \mathcal{M}_1(\Sigma^k)$. Assume that for some π^i,*

$$d(\nu, (\nu)_i \otimes \pi^i \otimes \cdots \otimes \pi^i) > \delta_m.$$

Further assume that $d(\nu, \nu') < \delta_m/4$. Then

$$d(\nu', (\nu')_i \otimes \pi^i \otimes \cdots \otimes \pi^i) > \delta_m/2.$$

Proof: Note that $d(\nu, \nu') < \delta_m/4$ implies that $d((\nu)_i, (\nu')_i) < \delta_m/4$ and that $d((\nu)_{i+1}, (\nu')_{i+1}) < \delta_m/4$. On the other hand, since π^i is a Markov kernel, it also follows that $d((\nu)_i \otimes \pi^i, (\nu')_i \otimes \pi^i) < \delta_m/4$ and therefore also that $d((\nu)_i \otimes \pi^i \otimes \cdots \otimes \pi^i, (\nu')_i \otimes \pi^i \otimes \cdots \otimes \pi^i) < \delta_m/4$. Hence,

$$\begin{aligned} d(\nu', (\nu')_i \otimes \pi^i \otimes \cdots \otimes \pi^i) \ &\geq \ d(\nu, (\nu)_i \otimes \pi^i \otimes \cdots \otimes \pi^i) - d(\nu, \nu') \\ &\quad - d((\nu)_i \otimes \pi^i \otimes \cdots \otimes \pi^i, (\nu')_i \otimes \pi^i \otimes \cdots \otimes \pi^i) \\ &> \ \delta_m/2 \end{aligned}$$

□

We have now completed all the preparatory steps required for the definition of the

proposed decision rule. Indeed, let $\epsilon_i(m)$ be a sequence of positive numbers, define

$$\alpha_i(m) = \frac{2^{2(m+3)}}{\epsilon_i(m)} \left[2 \log m + \log k + N^{\Sigma^{i+m}} \left(\sqrt{\epsilon_i(m)/2^{k+1}} \right) \left(1 - \log \sqrt{\epsilon_i(m)/2^{k+1}} \right) \right] \tag{9.30}$$

we have, by Lemma 9.3 and Theorem 9.4, that for any Markov measure $\mu \in A_i$,

$$P_\mu(\mu^{i+m}_{\alpha_i(m)} \in C^c_{i,m}) \leq \frac{1}{m^2} \tag{9.31}$$

The construction of the decision rule is then identical to the one described in Theorem 9.3, i.e. one forms the parsing of the observation sequence into the nonoverlapping blocks X^m described in equation (9.16) with $\beta(m)$ chosen as in (9.17). At each step, one forms, based on the block X^m, the empirical measures of order $m, m+1, \ldots, 2m$. The order estimate at the m-th step is now the smallest i such that $\mu^{i+m}_{\alpha_i(m)} \in C_{i,m}$. By the results of section 3, this decision rule achieves a.s. only a finite number of errors, regardless of the true order.

9.5 Discussion and Open Problems

In this chapter, we considered the problem of deciding whether an unknown probability measure belongs to one of several sets based on a set of random samples from the unknown measure. These results represent a substantial generalization of the results of Chapter 8 and those of [28, 66]. We briefly discussed an application of our results to the problem of density estimation. It seems that much more work could be done along these lines. In Section 9.4, we discussed an application/extension of our ideas to the problem of order determination of a Markov chain. This application consisted of partitioning the class of Markov measures into a countable number of sets according to the order of each Markov measure. One could consider the problem of partitioning the set of Markov measures in a more general manner and trying to decide to which set of the partition an unknown measure belongs. This would be a direct generalization of the results of Sections 9.2 and 9.3 to the Markov case.

In Chapter 8, we considered a symmetric decision criteria for the problem of deciding the mean. One could consider a similar criteria for the more general formulation of the present chapter. For a proof analogous to that presented in Section 8.4 to go through, one would need a result like the Lebesgue density theorem holding in spaces much more general than $\mathbf{R}^n$ (specifically, $\mathcal{M}_1(\Sigma)$). In Section 8.5, we discussed sev-

eral possible new directions to pursue for the problem of deciding the mean. Most of these directions suggest analogous problem to pursue for the much more general framework of the present chapter. In particular, it may be interesting to consider conditions necessary to perform the classification (see [31]), rates of convergence, and connections to other more standard learning models.

Chapter 10

Summary and Other Directions

In this chapter, we first summarize our results, and then discuss several other general directions that may be interesting to pursue. Each section deals with a different topic, although there are certainly some interrelations between the various topics. Furthermore, the topics differ substantially in the degree to which they are well defined, their scope, their expected difficulty, and their fundamental importance.

10.1 Summary

We studied a variety of problems in the areas of machine vision and machine learning. The first part of the thesis dealt with computational problems in machine vision. The second part of the thesis dealt with extensions of learning models with a view towards extending the domain of applicability of learning results to areas such as signal processing and machine vision.

We began by considering relationships between variational methods and discrete Markov random field formulations for the problem of image restoration and segmentation. Previous discrete versions of the segmentation problem fail to properly approximate the continuous formulation. We studied a number of properties of Minkowski content and used these to obtain a discrete formulation which correctly approximates the continuous segmentation problem. We then considered two other discrete versions for which the convergence proofs were considerably easier and which may result in more efficient implementations.

The results for the segmentation problem led us to consider a question concerning the computation of the length of a digitized contour. Specifically, we considered a parallel computation with a processor at each pixel of the digitized image. Notions

of local and non-local computations were considered, based on definitions of Minsky and Papert in their study of perceptrons. It was shown that for the usual rectangular digitization, length cannot be computed locally. On the other hand, we showed that for a random tesselation and an appropriate deterministic one, the length of straight line segments can be computed locally.

Another problem in vision that we studied concerns the complexity of model based recognition. Recently, model based approaches to object recognition have been a subject of considerable interest. We showed that certain formulations of model based recognition are NP-complete, so that efficient algorithms for these formulations are not likely to exist. The results are helpful in suggesting approaches that may lead to formulations with efficient algorithms.

Next, we considered a variety of extensions to the PAC model in machine learning. We began by showing some new relationships between metric entropy and the Vapnik-Chervonenkis (VC) dimension. We then considered the problem of learning over a class of distributions, and for certain special classes, we characterized when such learning can take place. The results provide some information on when prior knowledge regarding the distribution increases learnability.

Another extension of the PAC model that we considered involves active learning using arbitrary binary valued queries. A number of researchers have considered the use of various oracles and their effect on learnability. Our results provide bounds on the maximum gain that can be expected from using finite valued oracles. Surprisingly, asking arbitrary yes/no questions does not increase the set of learnable concept classes in either the fixed distribution or distribution-free settings, but as expected it can reduce the sample complexity.

A third extension of the PAC model that we considered involves learning from generalized samples. In the usual PAC model, the data received by the learner consists of values of the unknown concept at random points. We considered a model in which the information received by the learner can consist of general operators applied to the unknown concept. It appears that this model can be applied to a number of problems in geometric reconstruction, stereology, and signal processing to provide sample size bounds under a PAC criterion. We studied a particular application of the model to a problem of reconstructing a curve by counting intersections with straight lines. Our results our closely related to and, in a sense, refine a classical result from stochastic geometry.

Finally, we considered a problem concerning the classification of an unknown

probability measure from empirical data. Suppose we observe i.i.d. samples from an unknown distribution and wish to decide to which of a countable number of classes the unknown distribution belongs. The criteria for success is a type of almost sure class identification in the limit. Using large deviations techniques, we simplified and extended previous results in the case of classifying the mean. We also studied the much more general case of classifying the measure itself, and considered applications to density estimation and the problem of order determination of a Markov chain.

10.2 Nonuniform Learning, Misfit Versus Complexity Tradeoff, and Universal Coding

The standard PAC model requires the number of samples needed for learning to be uniformly bounded over all probability measures and all concepts. Various types of nonuniform learning can and have been considered [15, 18, 25, 77]. In particular, nonuniform learning with respect to the concepts in the concept class has received the most attention [18, 25, 77]. Conditions and algorithms for this problem have been obtained, but the fundamental limitations and optimality results are lacking. In fact, in contrast with uniform learning where the uniform bound on the sample size provides a measure of performance, a good measure of performance in the case of nonuniform learning is not obvious.

For a fixed distribution, a concept class is uniformly learnable iff it has finite metric entropy (see Chapter 5), i.e. iff a finite ϵ-approximation can be found. If X is a finite dimensional Euclidean space and P is any probability measure on X, then it turns out that any concept class C over X has a countable ϵ-approximation for each $\epsilon > 0$. This is sufficient to guarantee nonuniform learnability (with respect to concepts). One can also consider nonuniform active learnability, where, as in Chapter 6, the learner is allowed to ask arbitrary binary valued questions. Then for a countable ϵ-approximation, active learning is essentially equivalent to coding the integers. In this case, some fundamental limitations of active learnability are related to ideas on universal coding and priors for the integers [40, 98, 99].

For distribution-free nonuniform learning, one typically takes $C = \cup_{1=i}^{\infty} C_i$ with $C_1 \subset C_2 \subset \cdots$ and the VC dimension of each C_i finite but growing unboundedly with i. The 'complexity' or 'order' of a concept c is the smallest k for which $c \in C_k$. Typical learning algorithms for such problems either limit the complexity of the hy-

pothesis as a function of the number of samples, or first estimate the complexity of the target concept and then produce a hypothesis with complexity less than or equal to the estimate. O. Zeitouni has suggested a possible way to measure optimal performance for such a two step learning learning procedure, based on some work that has been done in estimating the order of Markov chains. Specifically, requiring the probability of overestimating the order to decrease exponentially at a given rate, an optimal learning procedure is defined to have the fastest exponential rate of decrease in underestimating the order. Such a criteria would allow comparisons to be made between different learning algorithms and the best achievable, which at present time is not possible. Finally, for both fixed distribution and distribution-free nonuniform learning it may be interesting to formulate learning algorithms in terms of the classic misfit vs. complexity tradeoff characteristic of minimum description length approaches [98, 99].

10.3 Prior Information and Mean Sample Size Bounds in Learning

A potentially powerful way to obtain stronger learnability results is to provide the learner with prior information. Fixing the concept class (and the distribution in the case of learnability under a fixed distribution) is in fact providing prior information. A more general form of prior information might be to provide a prior distribution on the concept class and/or over the class of distributions. D. Tse [116] has considered a form of this in which the concept class is partitioned into disjoint subsets and a probability is assigned to each subset. In general one may be able to put an arbitrary measure on the concept class. One approach might be to assign discrete measures to ϵ-approximations. Assigning prior distributions would allow studying mean sample size bounds as opposed to the usual worst-case bounds. Interestingly, active learning with a prior distribution would essentially be formally equivalent to problems in information theory.

10.4 Learning Under General Metric Uncertainty

In the standard PAC model, the probability distribution is used both to provide random samples as well as to measure the performance of learning. One can imagine situations where the goal is to learn according to some criteria not necessarily directly tied to the means of gathering data. This suggests formulating a learning problem in a general metric space setting where the metric need not be induced by the distribution used to generate samples. One natural question that immediately comes to mind is what compatibility conditions must be satisfied between the metric, the (fixed) distribution, and the concept class to allow learning to take place. A special case of this in which the metric is induced by a distribution (but not necessarily the same one used to generate samples) would provide a sort of robustness result with respect to the sampling distribution in the usual PAC model. For example, this would correspond to the case in which noise is added to the samples prior to being labeled, so that the learner is getting correct information, but not according to the correct distribution.

A more general model would allow some metric uncertainty in the sense that the learner may not know the metric exactly but only that it belongs to some class of metrics. This corresponds to the case in which the learner doesn't know exactly the criteria he is trying to optimize. This could be studied with respect to a variety of information gathering mechanisms — i.e. fixed distribution, class of distributions, active learning, learning by distances [16], etc. This model is natural in a sense since it separates uncertainty in the information gathering from uncertainty in the performance criteria. It is likely that learning can take place only if there are close ties between the two, which would still be an interesting result. However, it may be difficult to prove results in the general case since the powerful uniform convergence results may no longer apply.

10.5 Learning with Unions of Hyperspheres and Attribute Noise

An idea suggested by J. Koplowitz is to learn a concept by placing spheres (hyperspheres in general) around each positive example, and letting the radii of the spheres decrease appropriately as the number of samples increases. This is a natural idea

from the point of view of pattern recognition or density estimation, although as far as we know it has not been studied formally in either the pattern recognition or density estimation literature. Apparently some work in learning is currently being pursued along these lines [63]. In general, the number of samples will depend on the target concept, but perhaps distribution independent sample size bounds can be obtained. Although this may result in nonuniform learning, it seems that a very broad class of concepts will be learnable with this single representation for hypotheses, and the hypotheses can be easily generated from the data. Another potential advantage of this representation is that it may be possible to prove results on learning with attribute noise. This type of noise has been studied in discrete domains, but as far as we know there has been no general treatment.

10.6 Tracking Time-Varying Concepts

An interesting problem suggested recently by D. Helmbold is to study the tracking or learnability of mutable concepts. Some work is currently being pursued along these lines [59]. For the dynamics of the target concept, some possibilities are to allow movement unrestricted in direction or smoothness but with a bounded rate, or if the concept class is a differentiable manifold to consider flows on the concept class. The type of result that one might hope or expect to obtain is that learnability can take place if the dynamics of the target are not too large compared with the sampling rate, and one would like the criteria for learnability to be somewhat robust with respect to the specific concept class. Many interesting directions may arise by introducing dynamics into the learning model.

Bibliography

[1] Abelson, H., "Towards a theory of local and global computation," *Theoretical Computer Science*, Vol. 6, pp. 41-67, 1978.

[2] Alexandrov, A.D. and Yu.G. Reshetnyak, *General Theory of Irregular Curves*, Mathematics and Its Applications (Soviet Series) Vol. 29, Kluwer Academic Publishers, 1989.

[3] Ambartzumian, R.V., edt., *Stochastic and Integral Geometry*, Kluwer Academic Publishers, 1987.

[4] Ambrosio, L., "Existence Theory for a New Class of Variational Problems," Center for Intelligent Control Systems Report CICS-P-93, MIT, 1988

[5] Ambrosio, L., "Variational Problems in SBV," Center for Intelligent Control Systems Report CICS-P-86, MIT, 1988.

[6] Amsterdam, J., "Extending the Valiant learning model," *Proc. 5th Int. Conf. on Machine Learning*, pp. 381-394, 1988.

[7] Angluin, D. and P. Laird, "Learning from noisy examples," *Machine Learning*, Vol. 2, pp.343-370, 1988.

[8] Angluin, D., "Types of Queries for Concept Learning," Technical Report YALEU/DCS/TR-479, Yale Univ., Dept. of Computer Science, June 1986.

[9] Angluin, D., "Queries and Concept Learning," *Machine Learning*, Vol. 2, pp.319-342, 1988.

[10] Attouch, H., *Variational Convergence for Functions and Operators*, Pitman Publishing Inc., 1984.

[11] Baddeley, A.J., "Stochastic geometry and image analysis," in *CWI Monographs*, North Holland, 1986.

[12] Barron, A.R., "Logically Smooth Density Estimation," Ph.D. thesis, Dept. of Electrical Engineering, Stanford University, Sept., 1985.

[13] Barron, A.R. and T.M. Cover, "Minimum complexity density estimation," to appear in *IEEE Trans. on Info. Theory.*

[14] Baum, E.B., "Complete representations for learning from examples," From *Complexity in Information Theory*, Y.S. Abu-Mostafa, edt., pp. 77-98, Springer-Verlag, 1988.

[15] Ben-David, S., G.M. Benedek, Y. Manosur, "A parametrization scheme for classifying models of learnability," *Proc. Second Annual Workshop on Computational Learning Theory*, pp. 285-302, 1989.

[16] Ben-David, S., A. Itai, E. Kushilevitz, "Learning by distances," *Proc. of Third Annual Workshop on Computational Learning Theory*, pp. 232-245, 1990.

[17] Benedek, G.M. and A. Itai, "Learnability by fixed distributions," *Proc. of First Workshop on Computational Learning Theory*, pp. 80-90, 1988.

[18] Benedek, G.M. and A. Itai, "Nonuniform learnability," *ICALP*, pp. 82-92, 1988.

[19] Besl, P.J. and R. C. Jain, "Three-dimensional object recognition," *Computing Surveys*, 17(1):75–145, 1985.

[20] Beuscher, K. and P.R. Kumar, "Simultaneous learning and estimation for classes of probabilities," to be presented at COLT '91, Workshop on Computational Learning Theory, Santa Cruz, CA, Aug 1991.

[21] Binford, T.O., "Survey of model-based image analysis systems," *International Journal of Robotics Research*, 1(1):18–64, 1982.

[22] Blake, A. and A. Zisserman, "Invariant surface reconstruction using weak continuity constraints," *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, Miami, pp. 62-67, 1986.

[23] Blake, A. and A. Zisserman, *Visual Reconstruction*, MIT Press, 1987.

[24] Blumer, A., A. Ehrenfeucht, D. Haussler, M. Warmuth, "Classifying learnable geometric concepts with the Vapnik-Chervonenkis dimension," *Proc. 18th ACM Symp. on Theory of Comp.*, Berkeley, CA, pp. 273-282, 1986.

[25] Blumer, A., A. Ehrenfeucht, D. Haussler, M. Warmuth, "Occam's razor," *Info. Proc. Let.*, Vol. 24, pp. 377-380, 1987.

[26] Chin, R.T. and C. R. Dyer, "Model-based recognition in robot vision," *ACM Computing Surveys*, 18(1):67–108, 1986.

[27] Cover, T.M., "A hierarchy of probability density function estimates," in *Frontiers of Pattern Recognition*, Academic Press Inc., 1972.

[28] Cover, T.M., "On determining the irrationality of the mean of a random variable," *The Annals of Statistics*, Vol 1, pp. 862-871, 1973.

[29] De Giorgi, E., "Convergence problems for functionals and operators," Proc. Int. Meeting on Recent Methods in Nonlinear Analysis, Rene 1978, ed. E. De Giorgi, Magenes, Mosco Pitagora, Bologna, pp. 131-188, 1979.

[30] De Giorgi, E., "New problems in Γ-convergence and G-convergence," Proc. Meeting on Free Boundary Problems, Pavia 1979, Istituto Nazionale di Atta Matematica, Roma, Vol. II, pp. 183-194, 1980.

[31] Dembo, A. and Y. Peres, "A topological criterion for hypothesis testing," preprint, to be submitted to *The Annals of Statistics.*

[32] Deuschel, J.D. and D.W. Stroock, *Large deviations*, Academic Press, Boston, 1989.

[33] Devroye, L., *A Course in Density Estimation*, Vol. 14 in the series Progress in Probability and Statistics, Birkhauser, 1987.

[34] Devroye, L. and L. Gyorfi, *Nonparametric Density Estimation: The L^1 View*, John Wiley, 1985.

[35] Dubuc, B., C. Roques-Carmes, C. Tricot, and S.W. Zucker. The variation method: a technique to estimate the fractal dimension of surfaces. *SPIE Vol. 845 Visual Communications and Image Processing II*, pp. 241-248, 1987.

[36] Dudley, R.M., "Central limit theorems for empirical measures," *Ann. Probability*, Vol. 6, No. 6, pp. 899-929, 1978.

[37] Dudley, R.M., "Metric entropy of some classes of sets with differentiable boundaries," *J. Approx. Theory*, Vol. 10, No. 3, pp. 227-236, 1974.

[38] Eisenberg, B. and R.L. Rivest, "On the Sample Complexity of Pac-Learning Using Random and Chosen Examples," *Proc. Third Annual Workshop on Computational Learning Theory*, pp. 154-162, 1990.

[39] Ekeland, I. and R. Temam, *Convex analysis and variational problems*, North-Holland, 1976.

[40] Elias, P., "Universal codeword sets and representations of the integers," *IEEE Trans. on Info. Theory*, Vol. IT-21, No. 2, pp. 194-203, 1975.

[41] Falconer, K.J., *The Geometry of Fractal Sets*, Cambridge University Press, 1985.

[42] Federer, H., *Geometric Measure Theory*, Springer-Verlag, 1969.

[43] Gallager, R.G., *Information Theory and Reliable Communication*, Wiley & Sons, 1968.

[44] Garey, M.R. and D. S. Johnson, *Computers and Intractability. A Guide to the Theory Of NP-completeness*, Freeman, 1979.

[45] Geiger, D. and A. Yuille, "A common framework for image segmentation by energy functions and nonlinear diffusion," A.I. Lab Memo, M.I.T., 1989.

[46] Geman, S. and D.Geman, "Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images," *IEEE Trans. Pattern Analysis and Machine Intelligence*, Vol. PAMI-6, pp. 721-741, 1984.

[47] Gold, I.M, "Language identification in the limit," *Information and Control*, Vol. 10, pp. 447-474, 1967.

[48] Grimson, W.E.L., "The combinatorics of object recognition in cluttered environments using constrained search," *Proceedings of the Second International Conference on Computer Vision (ICCV'88)*, pages 218–227, December, 1988.

[49] Grimson, W.E.L., "The combinatorics of local constraints in model-based recognition and localization from sparse data," *Journal of the ACM*, 33(4):658–686, 1986.

[50] Groeneboom, P., "Some current developments in density estimation," in *CWI Monographs*, North-Holland, 1986.

[51] Halmos, P.R., *Measure Theory*, Springer, 1974.

[52] Hannan, E.J. and B.G. Quinn, "The determination of the order of an autoregression," *J. Roy. Stat. Soc., Ser. B*, 41, pp. 190-195, 1979.

[53] Haralick, R.M. and L. G. Shapiro, *Computer and Robot Vision*, Addison-Wesley, 1991.

[54] Haussler, D., M. Kearns, N. Littlestone, M.K. Warmuth, "Equivalence of models for polynomial learnability," *Proc. First Workshop on Computational Learning Theory*, pp. 42-55, 1988.

[55] Haussler, D., "Generalizing the PAC model for neural net and other learning applications," Technical Report UCSC-CRL-89-30, U.C. Santa Cruz, Computer Research Laboratory, September, 1989.

[56] Haussler, D., "Decision theoretic generalizations of the PAC model for neural net and other learning applications," Technical Report UCSC-CRL-91-02, U.C. Santa Cruz, Computer Research Laboratory, December, 1990.

[57] Haussler, D. and E. Welzl. ϵ-Nets and simplex range queries. *Discrete and Comput. Geom.* 2, pp. 127-151, 1987.

[58] Hawkes, J. Hausdorff measure, entropy, and the independence of small sets. *Proc. London Math. Soc.* (3)28, pp. 700-724, 1974.

[59] Hembold, D.P. and P.M. Long, "Tracking drifting concepts using random examples," to be presented at COLT '91, Workshop on Computational Learning Theory, Santa Cruz, CA, Aug 1991.

[60] Hoeffding, W. Probability inequalities for sums of bounded random variables. *J. Amer. Statist. Assoc.* 58, pp. 13-30, 1963.

[61] Karl, W.C., "Reconstructing objects from projections," Ph.D. Thesis, Dept. of EECS, Massachusetts Institute of Technology, February, 1991.

[62] Kearns, M. and M. Li, "Learning in the presence of malicious errors," *Proc. 20th ACM Symp. on Theory of Comp.*, Chicago, Illinois, pp. 267-279, 1988.

[63] Kim, M.W., "Learning by smoothing: a morphological approach," to be presented at COLT '91, Workshop on Computational Learning Theory, Santa Cruz, CA, Aug 1991.

[64] Kirousis, L.M. and C. H. Papadimitriou, "The complexity of recognizing polyhedral scenes," *Journal of Computer and System Sciences*, 37:14–38, 1988.

[65] Kolmogorov, A.N. and V.M. Tihomirov, "ϵ-Entropy and ϵ-capacity of sets in functional spaces," *Amer. Math. Soc. Transl.*, Vol. 17, pp. 277-364, 1961.

[66] Koplowitz, J., Abstracts of papers, *Int. Symp. on Info. Theory*, Cornell Univ., Ithaca, NY, Oct. 10-14, p. 64, 1977 and Private Communication.

[67] Koplowitz, J. On the entropy and reconstruction of digitized planar curves. IEEE Int. Symp. on Info. Theory, Brighton, England, June, 1985.

[68] Koplowitz, J. and A. Bruckstein, "Design of perimeter estimators for digitized planar shapes," *SPIE Vol. 1001 Visual Communications and Image Processing*, pp. 756-763, 1988.

[69] Kulkarni, S.R., "Minkowski content and lattice approximations for a variational problem," Center for Intelligent Control Systems Report CICS-95, M.I.T., 1988.

[70] Kulkarni, S.R., "On Metric Entropy, Vapnik-Chervonenkis Dimension and Learnability for a Class of Distributions," Center for Intelligent Control Systems Report CICS-P-160, M.I.T., 1989.

[71] Kulkarni, S.R. and O. Zeitouni, "Can one decide the type of the mean from the empirical measure?" Submitted to *Statistics and Probability Letters.*

[72] Kulkarni, S.R., S.K. Mitter, J.N. Tsitsiklis, "Active learning using arbitrary binary valued queries," Center for Intelligent Control Systems Report CICS-257, M.I.T., 1990.

[73] Kulkarni, S.R., S.K. Mitter, T.J. Richardson, "An existence result and lattice approximations for a variational problem," *Signal Processing, Part I: Signal Processing Theory*, IMA Vol. 22, edt. by Auslander et al., Springer-Verlag, pp. 189-210, 1990.

[74] Lamdan, Y. and H. J. Wolfson, "Geometric hashing: A general and efficient model-based recognition scheme," *Proc. of the Second International Conference on Computer Vision (ICCV'88)*, pp. 238-249, Dec., 1988.

[75] Lee, C.N. and A. Rosenfeld, "Simple Connectivity is Not Locally Computable for Connected 3D Images," *Computer Vision, Graphics, and Image Processing*, Vol. 51, pp. 87-95, 1990.

[76] Lele, A.S., S.R. Kulkarni, and A.S. Willsky, "Convex set estimation from support line measurements and applications to target reconstruction from laser radar data," *SPIE Proc. Vol. 1222, Laser Radar V*, pp. 58-82, 1990. (Submitted to *J. Optical Soc. of Amer.*)

[77] Linial, N., Y. Mansour, R.L. Rivest, "Results on learnability and the Vapnik-Chervonenkis dimension," *Proc. First Workshop on Computational Learning Theory*, pp. 56-68, 1988.

[78] Littlestone, N., "Mistake Bounds and Logarithmic Linear-Threshold Learning," Ph.D. thesis, U.C. Santa Cruz, March, 1989.

[79] Littlestone, N., "Learning when irrelevant attributes abound: A new linear-threshold algorithm," *Machine Learning*, Vol. 2, pp. 285-318, 1988.

[80] Mandelbrot, B.B., *The Fractal Geometry of Nature*, W.H. Freeman and Company, 1982.

[81] Marroquin, J.L., "Probabilistic Solution of Inverse Problems," Ph.D. Thesis, Dept. of E.E.C.S., M.I.T., 1985.

[82] Marroquin, J., S. Mitter, T. Poggio, "Probabilistic solution of ill-posed problems in computational vision," *J. of the Amer. Stat. Soc.*, Vol. 82, No. 397, pp. 77-89, 1987.

[83] Merhav, N., M. Gutman, and J. Ziv, "On the determination of the order of a markov chain and universal data compression," *IEEE Trans. Info. Theory*, Vol. IT-35, pp. 1014-1019, 1989.

[84] Minsky, M and S. Papert, *Perceptrons: An Introduction to Computational Geometry*, The MIT Press, 1969.

[85] Montanari, U., "Continuous skeletons from digitized images," *Jornal of the ACM*, Vol. 16, No. 4, pp.534-549, 1969.

[86] Montanari, U., "A note on minimal length polygonal approximation to a digitized contour," *Communications of the ACM*, Vol. 13, No. 1, 1970.

[87] Moran, P.A.P., "Measuring the length of a curve," *Biometrika*, Vol. 53, pp. 359-364, 1966.

[88] Mumford, D. and J. Shah, "Boundary detection by minimizing functionals," *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, San Francisco, pp. 22-26, 1985.

[89] Mumford, D. and J. Shah, "Optimal Approximations by Piecewise Smooth Functions and Associated Variational Problems," Center for Intelligent Control Systems Report CICS-P-68, 1988. Submitted to *Communications on Pure and Applied Mathematics.*

[90] Natarajan, B.K., "Learning over classes of distributions," *Proc. First Workshop on Computational Learning Theory*, pp. 408-409, 1988.

[91] Natarajan, B.K., "Probably-approximate learning over classes of distributions," Carnegie-Mellon Univ., Unpublished Manuscript, 1989.

[92] Natarajan, B.K. and P.T. Tadepalli, "Two new frameworks for learning," *Proc. 5th Int. Conf. on Machine Learning*, pp. 402-415, 1988.

[93] Parent, P. and S.W. Zucker, "Trace inference, curvature consistency, and curve detection," *IEEE Trans. PAMI*, Vol. 11, pp. 823-839, 1989.

[94] Pollard, D., *Convergence of Stochastic Processes*, Springer-Verlag, 1984.

[95] Prince, J.L. and A.S. Willsky, "Estimating convex sets from noisy support line meassurements," *IEEE Trans. PAMI*, Vol. 12, pp. 377-389, 1990.

[96] Richardson, T.J., "Existence Result for a Problem Arising in Computer Vision," Center for Intelligent Control Systems Report CICS-P-63, M.I.T., 1988.

[97] Richardson, T.J., "Scale Independent Piecewise Smooth Segmentation of Images Via Variational Methods," Ph.D. Thesis, Dept. of E.E.C.S., M.I.T., 1989.

[98] Rissanen, J., "A universal prior for the integers and estimation by minimum description length," *Annals of Statistics*, Vol. 11, No. 2, pp.416-431, 1983.

[99] Rissanen, J., *Stochastic Complexity in Statistical Inquiry*, Series in Computer Science Vol. 15, World Scientific, 1989.

[100] Rivest, R.L. Learning decision lists. *Machine Learning* 2(3), pp. 229-246, 1987.

[101] Rivest, R.L., A.R. Meyer, D.J. Kleitman, K. Winklmann, and J. Spencer, "Coping with Errors in Binary Search Procedures," *J. of Computer and System Sciences*, Vol. 20, pp. 396-404, 1980.

[102] Rogers, C.A., *Hausdorff Measure*, Cambridge University Press, 1970.

[103] Rosenfeld, A., "Survey: Image Analysis and Computer Vision: 1988," *Computer Vision, Graphics, and Image Processing*, Vol. 46, pp. 196-264, 1989.

[104] Rosenfeld, A., "Survey: Image Analysis and Computer Vision: 1988," *Computer Vision, Graphics, and Image Processing*, Vol. 50, pp. 188-240, 1990.

[105] Santalo, L.A., *Integral Geometry and Geometric Probability.* Volume 1 of *Encyclopedia of Mathematics and its Applications,* Addison-Wesley, Reading, MA, 1976.

[106] Schapire, R., "The strength of weak learnability," *Machine Learning*, Vol. 5, pp.197-227, 1990.

[107] Serra, J., *Image Analysis and Mathematical Morphology*, Academic Press Inc., 1982.

[108] Shah, J., "Segmentation by Minimizing Functionals: Smoothing Properties," To be published.

[109] Sherman, S., "A comparison of linear measures in the plane," *Duke Math. J.*, Vol. 9, pp. 1-9, 1942.

[110] Skiena, S.S., "Geometric Probing," Ph.D. thesis, Dept. of Computer Science, Univ. of Illinois at Urbana-Champaign, (report no. UIUCDCS-R-88-1425), April, 1988.

[111] Sloan, R., "Types of noise in data for concept learning," *Proc. First Workshop on Computational Learning Theory*, pp. 91-96, 1988.

[112] Steele, J.M., "Subadditive Euclidean functionals and nonlinear growth in geometric probability," *The Annals of Probability*, Vol. 9, No. 3, pp. 365-376, 1981.

[113] Steinhaus, H., "Length, shape, and area," *Colloquium Mathematicum*, Vol. 3, pp. 1-13, 1954.

[114] Stoyan, D., W.S. Kendall, and J. Mecke, *Stochastic Geometry and Its Applications*, Wiley series in probability and mathematical statistics, 1987.

[115] Tikhomirov, V.M., "Kolmogorov's work on ϵ-entropy of functional classes and the superposition of functions," *Russian Math. Surveys*, Vol. k8, pp. 51-75, 1963.

[116] Tse, D.N.C., "Inductive learning from noisy examples with prior knowledge," Univ. of Waterloo, Unpublished Manuscript, 1988.

[117] Tsotsos, J.K., "Analyzing vision at the complexity level," *Behavioral and Brain Sciences*, 13:423–469, 1990.

[118] Valiant, L.G., "A theory of the learnable," *Comm. ACM*, Vol. 27, No. 11, pp. 1134-1142, 1984.

[119] Vapnik, V.N. and A.Ya. Chervonenkis, "On the uniform convergence of relative frequencies to their probabilities," *Theory of Prob. and its Appl.*, Vol. 16, No. 2, pp. 264-280, 1971.

[120] Vapnik, V.N. and A.Ya. Chervonenkis, "Necessary and and sufficient conditions for the uniform convergence of means to their expectations," *Theory of Prob. and its Appl.*, Vol. 26, No. 3, pp. 532-553, 1981.

[121] Vapnik, V.N., *Estimation of Dependeces Based on Empirical Data*, Springer-Verlag, 1982.

[122] Wang, Y., Harvard University, Unpublished Notes, 1988.

[123] Wells, W.M. III, "A statistical approach to model matching," *Proc. SPIE Conf. on Intelligent Robots and Computer Vision IX: Algorithms and Techniques*, Boston, Nov. 1990.

[124] Wenocur, R.S. and R.M. Dudley, "Some special Vapnik-Chervonenkis classes," *Discrete Math.*, Vol. 33, pp. 313-318, 1981.