NONPARAMETRIC SET ESTIMATION PROBLEMS IN STATISTICAL INFERENCE AND LEARNING

by

Aarti Singh

A dissertation submitted in partial fulfillment of
the requirements for the degree of

Doctor of Philosophy

(Electrical Engineering)

at the

UNIVERSITY OF WISCONSIN – MADISON

2008

To my dad on his 60^{th} birthday.

ACKNOWLEDGEMENTS

This thesis is the culmination of the guidance, inspiration and support of many people to whom I would like to express my gratitude. First and foremost, I thank my advisor Professor Robert Nowak for being a constant source of inspiration and new ideas, for inculcating in me the same contagious zeal for research, and promoting both professional and personal development. He is an excellent mentor, friend and role model, and I will always continue to learn from him.

I am also grateful to my other committee members, Parmesh Ramanathan for his invaluable advice and support in all matters, Barry Van Veen for inspiring me to ask the right research questions when I just started, Jerry Zhu for suggesting an exciting research problem and very helpful discussions, and Stark Draper for his careful review and stimulating remarks.

I would also like to thank Clay Scott and Rebecca Willett for being excellent collaborators, and for their very useful advice and help in exploring future career options. Rui Castro deserves a special mention; I have always enjoyed discussions with him, and appreciate his enthusiasm for discussing new ideas and patience for explaining things over and over. I thank my colleagues - Waheed Bajwa, Laura Balzano, Brian Eriksson, Jarvis Haupt and Mike Rabbat - for their support and help in both professional and personal matters, and for a wonderful graduate experience. I also thank all my friends in Madison for their companionship and making my stay here fun and enjoyable.

A very special thanks to my family - my dad for nurturing in me a scientific and rational vision, my mom for her selfless and unconditional support, my sister for the long chats that

would cheer me up on a grey day and my brother for always reminding me of the lighter side of life.

And to my husband Satbir, for always believing in me and helping me through my doubts, indecision and cribbing. I can not imagine completing this or any other journey without his incessant love and support.

TABLE OF CONTENTS

Page

ACKNOWLEDGEMENTS ii

LIST OF TABLES vii

LIST OF FIGURES viii

ABSTRACT xi

1 Introduction 1

1.1 Motivation 2
1.2 Preliminaries 6
1.3 Contribution 8
1.4 Organization 11

2 Adaptive Hausdorff Estimation of Density Level Sets 12

2.1 Introduction 13
2.2 Density assumptions 18
2.3 Motivating an Error Measure for Hausdorff control 22
2.4 Hausdorff accurate Level Set Estimation using Histograms 26
2.4.1 Adaptivity to unknown density regularity 28
2.4.2 Multiple level set estimation 30
2.4.3 Support set estimation 31
2.4.4 Addressing jumps in the density at the level of interest 33
2.5 Concluding Remarks 34
2.6 Proofs 36
2.6.1 Proof of Lemma 1 36
2.6.2 Proof of Proposition 1 40
2.6.3 Proof of Proposition 2 40
2.6.4 Proof of Theorem 1 43
2.6.5 Proof of Theorem 2 52
2.6.6 Proof of Proposition 3 58

Page

2.6.7 Proof sketch of Theorem 3 . . . 63
2.6.8 Proof sketch for $\alpha \geq 0$ . . . 65

3 Quantitative Analysis of Semi-Supervised Learning . . . 69

3.1 Introduction . . . 69
3.2 Characterization of model distributions under the cluster assumption . . . 74
3.3 Learning Decision Regions . . . 76
3.4 SSL Performance and the Value of Unlabeled Data . . . 78
3.5 Density-adaptive Regression . . . 80
3.6 Concluding Remarks . . . 83
3.7 Proofs . . . 84
3.7.1 Proof of Lemma 8 . . . 85
3.7.2 Proof of Corollary 5 . . . 92
3.7.3 Semi-Supervised Learning Upper Bound . . . 92
3.7.4 Supervised Learning Lower Bound . . . 95

4 Level Set based Approach to fMRI Data Analysis . . . 104

4.1 fMRI Analysis . . . 105
4.1.1 Hypothesis Testing for fMRI . . . 105
4.1.2 Level Set Estimation for fMRI . . . 107
4.1.3 The Level Set Approach vs. Gaussian Field Approaches . . . 109
4.2 Level Set Activation Detection in fMRI . . . 109
4.2.1 Error metrics . . . 109
4.2.2 Estimation via Trees . . . 111
4.2.3 Performance Analysis . . . 113
4.3 Improved Estimation via Cross-Validation . . . 114
4.4 Experimental Results . . . 115
4.4.1 Simulated Data . . . 115
4.4.2 fMRI Data . . . 116
4.5 Concluding Remarks . . . 117

5 Adaptive Mobile Sensing for Environmental Monitoring . . . 120

5.1 Introduction . . . 121
5.2 Active Learning for Spatial Mapping . . . 124
5.3 Adaptive Sampling Paths for Mobile Sensing . . . 130
5.3.1 Single mobile sensor . . . 131
5.3.2 Mobile Sensing Network . . . 139

Page

5.4 Case Study: Lake Mapping 140
5.5 Concluding Remarks 145

6 Summary and Future Directions 146

APPENDIX Proof Sketch of Theorem 13 153

REFERENCES 155

LIST OF TABLES

Table Page

3.1 Comparison of finite sample lower bounds on the mean square error for supervised learning, with finite sample upper bounds on the mean square error for semi-supervised learning, for the margin based model distributions. These bounds hold for $m \gg n^{2d}$ and $d \geq 2\alpha/(2\alpha - 1)$, and suppress constants and log factors. 82

LIST OF FIGURES

Figure Page

1.1 Scatterplot of chemical composition of olive oil samples with the data grouped into hierarchical density clusters, and corresponding cluster tree. (source: [1]) . 3

1.2 (a) Scatterplot of two optical properties of biological cells in flow cytometry data and (b) corresponding likelihood depth contours. (source: [2]) 4

1.3 Two dimensional projection of hand-written digit features. If unlabeled data (without corresponding digit information, can be used to learn the clusters of the feature density, a few labeled examples are needed to identify the digit corresponding to each cluster. 5

2.1 (a) The γ-level set G^*_γ of a density function $f(x)$, (b) Two candidate set estimates G_A and G_B with the same volume of symmetric difference error $\mathrm{vol}(G_A \Delta G^*_\gamma) = \mathrm{vol}(G_B \Delta G^*_\gamma)$, however G_A does not preserve the topological properties (non-connectivity) and has large Hausdorff error $d_\infty(G_A, G^*_\gamma)$, while G_B preserves non-connectivity and has small Hausdorff error $d_\infty(G_B, G^*_\gamma)$. 15

2.2 Densities used in the lower bound construction for Hausdorff accurate support set estimation. 61

3.1 (a) Two separated high density regions with different labels that (b) cannot be discerned if the sample size is too small, but (c) can be estimated if sample density is high enough. 73

3.2 The margin ξ measures the minimal width of a decision region, or separation between support sets of the marginal mixture component densities. The margin is positive if there is no overlap between the component support sets, and negative otherwise. 75

3.3 Examples of two sets of marginal density functions $p^\omega, p^{\omega'}$ for (a) $\xi < 0$, (b) $\xi > 0$ and regression functions $f^\omega, f^{\omega'}$ used for minimax construction. 98

Figure Page

4.1 (a) An example Recursive Dyadic Partition (RDP) of a 2-d function domain and (b) associated tree structure. 111

4.2 Comparison of $\gamma = 0.7$ level set estimation, thresholding with clairvoyant pre-smoothing, and thresholding with typical pre-smoothing for a simulated activation pattern. Black regions are false positives, light gray regions indicate false negatives, their union comprises the erroneous region $\widehat{G}\Delta G^*_\gamma$. (a) Level set estimate; $\mathrm{vol}(\widehat{G}\Delta G^*_\gamma) = 0.024$. (b) Voxel-wise threshold estimate with clairvoyant pre-smoothing; $\mathrm{vol}(\widehat{G}\Delta G^*_\gamma) = 0.030$. (c) Voxel-wise threshold estimate with typical pre-smoothing; $\mathrm{vol}(\widehat{G}\Delta G^*_\gamma) = 0.075$. 117

4.3 Estimates of neural activation regions. The dark gray regions overlaid onto the anatomic brain image represent the declared activity. With $\gamma = 0.95$: (a) level set estimate, (b) voxel-wise threshold estimate with pre-smoothing. With $\gamma = 0.90$: (c) level set estimate, (d) voxel-wise threshold estimate with pre-smoothing. . . 118

5.1 Probabilistic bisection sampling strategy (Horstein [3]) 126

5.2 Multiscale adaptive sampling strategy, Castro et al. [4, 5] 129

5.3 A toy example of estimating a 1-d solar intensity map containing a single change-point at 70 m due to a shadow, using a NIMS type mobile sensor. The sample locations and observations are shown in (a) for adaptive survey and (b) for passive survey. 132

5.4 Multiscale adaptive path of a mobile sensor for mapping a field containing a boundary. In the first step, the mobile sensor follows a coarse survey path (a) and produces a rough estimate of the field (b). In the refinement pass (c), the mobile follows a path adapted to the interesting regions of the field and produces a fine resolution estimate (d). 135

5.5 Cooperative strategy for k mobile sensors 141

5.6 Simulated low resolution water current velocity profile in Lake Wingra. Notice there are two distinct regions characterized by low or high velocity, with significant gradient between them. 142

Figure Page

5.7 Comparison of passive and adaptive path planning approaches for water current velocity mapping in a freshwater lake. The adaptive strategy requires only 14 hrs for mapping the nearly 2 km× 1 km lake to a resolution of < 10 m, as opposed to 48 hrs using the passive method. 144

NONPARAMETRIC SET ESTIMATION PROBLEMS IN STATISTICAL INFERENCE AND LEARNING

Aarti Singh

Under the supervision of McFarland-Bascom Professor in Engineering Robert D. Nowak

At the University of Wisconsin - Madison

Set estimation is a problem that arises in myriad applications where a region of interest needs to be estimated based on a finite number of observations. This can involve identifying the support of a function, where a function exceeds a certain level, or where a function exhibits a discontinuity or changepoint. Furthermore, set estimation arises as a subproblem in other learning problems. For example, classification and piecewise-smooth regression require identification of subregions called *decision sets* over which the classification label is constant or the regression function is homogeneously smooth. This thesis addresses some open theoretical questions in nonparametric learning using level sets and decision sets. It also discusses applications of set estimation to inference problems in neuroimaging and wireless sensor networks.

The first problem we investigate is estimation of the level set of a density under the Hausdorff error metric. Control of this error metric guarantees a spatially uniform confidence interval around the set estimate, that is desirable in many applications. A data-adaptive method is proposed that can provide minimax optimal Hausdorff guarantees over a very general class of densities, without assuming a priori knowledge of density parameters, and requiring only local regularity of the density in the vicinity of the desired level. The second problem we address is semi-supervised learning, that capitalizes on the abundance of unlabeled data to learn the decision sets and simplify supervised learning tasks such as regression and classification. We develop a rigorous framework to characterize the amount of

improvement possible in a semi-supervised setting, and quantify the relative value of unlabeled and labeled data using finite sample error bounds. Finally, this thesis also contributes to two critical applications - we propose a novel level set based approach to fMRI data analysis that can adaptively aggregate neighboring voxels to exploit structural information and boost detection of weak brain activity patterns. Secondly, we develop feedback-driven adaptive mobile sensing paths for environmental monitoring that focus on regions where the environmental phenomenon exhibits a changepoint and achieve optimal tradeoffs between path length, latency, and fidelity.

Approved:

McFarland-Bascom Professor in Engineering Robert D. Nowak
Department of Electrical and Computer Engineering
University of Wisconsin - Madison

ABSTRACT

Set estimation is a problem that arises in myriad applications where a region of interest needs to be estimated based on a finite number of observations. This can involve identifying the support of a function, where a function exceeds a certain level, or where a function exhibits a discontinuity or changepoint. Furthermore, set estimation arises as a subproblem in other learning problems. For example, classification and piecewise-smooth regression require identification of subregions called *decision sets* over which the classification label is constant or the regression function is homogeneously smooth. This thesis addresses some open theoretical questions in nonparametric learning using level sets and decision sets. It also discusses applications of set estimation to inference problems in neuroimaging and wireless sensor networks.

The first problem we investigate is the estimation of the support set and level set of a density under the Hausdorff error metric. The Hausdorff error metric controls the maximal deviation between points in the estimated set and true set, and thus guarantees a spatially uniform confidence interval around the set estimate that is desirable in many applications. Most current methods, however, optimize a global error criterion based on the symmetric set difference measure which can lead to estimates that veer greatly from the desired level set. We present a histogram-based support and level set estimation method that can provide minimax optimal rates of Hausdorff error convergence over a very general class of densities than considered in previous literature. Moreover, the proposed method is adaptive to unknown local density regularity.

The second problem we address is semi-supervised learning. Since labeled data is expensive and time-consuming to obtain, semi-supervised learning methods focus on the use

of unlabeled data to help in supervised learning tasks such as regression and classification. If the decision sets in these problems can be learnt using unlabeled data, the supervised learning task reduces to a simple majority vote or averaging on each decision set. We develop a rigorous framework to characterize the amount of improvement possible in a semi-supervised setting and the relative value of unlabeled and labeled data using finite sample error bounds. The resulting analysis provides a largely missing quantitative analysis of semi-supervised learning that explains the empirical success of semi-supervised learning in favorable situations.

Finally, we explore two critical applications where statistical inference relying on set estimation is particularly useful. The first is brain activation detection in fMRI studies. We propose a level set based approach to fMRI data analysis that optimizes a localization error and can adaptively aggregate neighboring voxels based on the correlation map. Thus, it exploits structural information, as opposed to conventional approaches that assume independent activation at each voxel. The second application is to environmental monitoring using mobile sensors, where energy constraints necessitate design of adaptive sensing paths that focus on regions where the environmental phenomenon exhibits a changepoint. Based on recent advances in active learning, we propose a feedback driven path planning scheme that provides optimal tradeoffs between path length, latency and fidelity.

Chapter 1

Introduction

The advent of technology has made it easier to generate, collect and store data. Data processing and analysis typically relies on estimation of an underlying data generating function, for example the density or regression function. As a result, there is a rich and well-developed body of literature in mathematical statistics on parametric and non-parametric density estimation and regression [6–11]. However, often it suffices to focus on a useful summary of the data rather than estimating the entire data generating function. For example, in many applications it is adequate to estimate regions or sets over which the underlying data generating function exhibits some characteristics of interest. Set estimation is an important learning problem that is concerned with accurate recovery and localization of a region of interest based on a finite number of (possibly noisy) observations. Some important set learning problems include identifying the support of a function, where a function exceeds a certain level, or where a function exhibits a discontinuity or changepoint. These problems arise in many applications, for example, hierarchical clustering [1, 12–14], data ranking [15], anomaly detection [2, 16, 17], detecting regions of brain activity [18, 19], mapping environmental contamination [20, 21], estimating contour levels in digital elevation mapping [22] and identification of genes with significant expression levels in microarray analysis [23, 24]. Furthermore, set estimation arises as a subproblem in other learning problems, where it may be solved either explicitly or implicitly. Two canonical learning problems that rely on set estimation are classification and piecewise smooth regression. Both problems involve learning a mapping from an input or feature space $\mathcal{X}$ to an output or label space $\mathcal{Y}$. In binary

classification, $\mathcal{Y} = \{0, 1\}$ and the objective is to learn the mapping (or classifier) f^* that minimizes the probability of error $P(f(X) \neq Y)$ over all classifiers f. Thus the target classifier $f^*(x) = \mathbf{1}_{\eta(x) \geq 1/2}$, where $\eta(x) = P(Y = 1|X = x)$, and the problem essentially corresponds to learning the $1/2$ level set of the regression function $\eta(x)$. In piecewise-smooth regression, $\mathcal{Y}$ is typically continuous and the objective is to learn the mapping (or estimator) f^* that minimizes the mean squared error $\mathbb{E}_{XY}[(f(X) - Y)^2]$ over all estimators f. If the target or regression function $f^*(x) = \mathbb{E}_{Y|X}[Y|X = x]$ is homogeneously smooth, the optimal estimator consists of averaging (or fitting a polynomial) in a local neighborhood. However, if the target function is inhomogeneous and exhibits sharp edges or changepoints, then the averaging needs to be confined to sets over which the target function is smooth. Thus, classification and piecewise-smooth regression require identification of regions over which the classification label is a constant or the regression function is homogeneously smooth. We refer to these regions in classification or regression problems as *decision sets*.

A straightforward approach to set estimation is to first estimate the complete function accurately and then extract the desired "plug-in" set estimate. However, set estimation requires less information (where the function exhibits a certain property of interest) than function estimation (the value of the function at all locations), and hence is an intrinsically easier problem. As a result, methods that directly address set estimation, instead of optimizing a function estimation error, tend to be more efficient in terms of both statistical and computational complexity. Because of their unique objectives, the two problems require distinct estimators, evaluation metrics and analysis tools. This thesis addresses some open theoretical questions in nonparametric learning using level sets and decision sets. It also discusses applications of set estimation to inference problems in neuroimaging and wireless sensor networks.

1.1 Motivation

As mentioned above, set estimation is relevant to many scientific and engineering applications. Here we elaborate on some of these applications that motivate the research presented

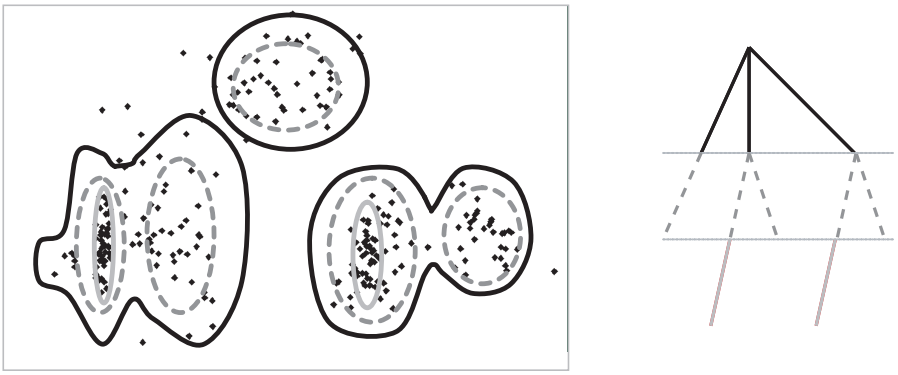


Figure 1.1 Scatterplot of chemical composition of olive oil samples with the data grouped into hierarchical density clusters, and corresponding cluster tree. (source: [1])

in this dissertation. However, this list is not all-inclusive, and the issues we will address pertain to many other application scenarios.

Clustering: Density levels set estimators are used by many data clustering procedures [1, 12–14]. In the level set based approach, clusters are identified as connected components of one or more density level sets. This approach to clustering is especially useful in exploratory data analysis as it does not require any a priori assumptions on the shape, number or location of clusters, unlike other algorithms like k-means clustering. Correct identification of connectivity of the density level set components is crucial to recover clusters with arbitrary shapes. Moreover, connected components of level sets at multiple successive levels of interest can be used to generate a hierarchy of nested clusters. This hierarchical structure can be represented as a *cluster tree* [1] (See Figure 1.1) and provides a useful summary of the data.

Anomaly detection and data ranking: Anomalies are often defined as outliers that do not follow a nominal distribution. Thus, a common approach to anomaly detection is to learn a (high) density level set of the nominal data distribution [2, 16]. Observations that fall outside the level set, in the low density region, are tagged as anomalies. Robustness of the level set estimate is highly desirable for anomaly detection to ensure that outliers lying in regions of low nominal data density are easily identified.

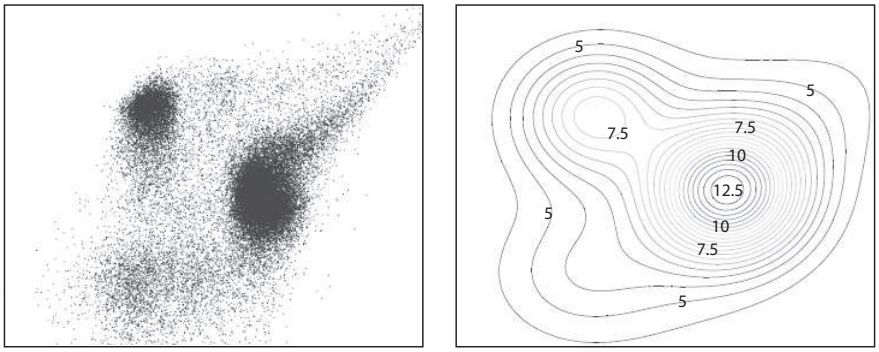


Figure 1.2 (a) Scatterplot of two optical properties of biological cells in flow cytometry data and (b) corresponding likelihood depth contours. (source: [2])

A related application is data ranking or ordering of data points. The notion of data-depth [15] is very useful in multivariate data analysis to provide an ordering of data points based on their degree of outlyingness. As an example, consider the flow cytometry data shown in Figure 1.2(a) that depict two optical properties of biological cells. Cell specimens are often contaminated with air bubbles and cell debris that result in outliers. Since these contaminants occur in small quantities, an approach to identifying these is to rank or order the data points based on their likelihood. This can be done by estimating the likelihood-depth contours (see Figure 1.2(b)), which correspond to density level sets. Again, a robust measure of accuracy is desirable for estimating the data-depth contours that is less susceptible to severe misrankings.

Statistical Learning: Set estimation has a close connection to supervised and semi-supervised statistical learning. Binary classification is essentially concerned with finding the region of feature space where the label $Y = 1$ is more probable than $Y = 0$, that is, the $1/2$ level set of $P(Y = 1|X = x)$ [25].

Moreover, recent work in semi-supervised learning [26–28] indicates that learning the decision sets, over which the label is constant in classification or the regression function is homogeneous in piecewise-smooth regression problems, using unlabeled data can

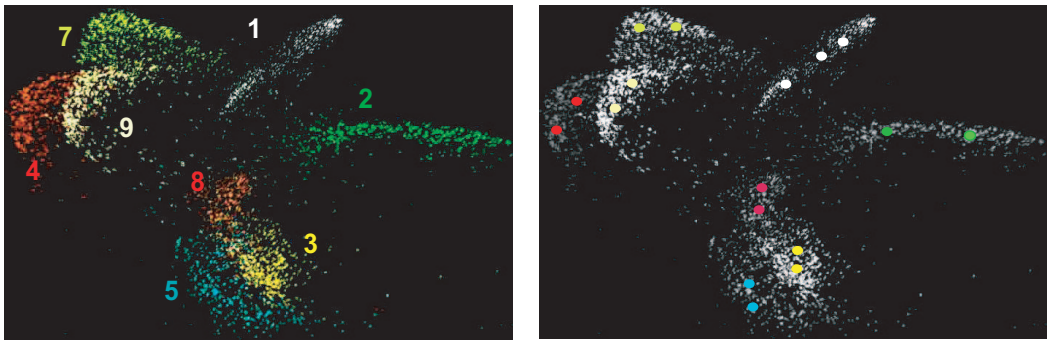


Figure 1.3 Two dimensional projection of hand-written digit features. If unlabeled data (without corresponding digit information, can be used to learn the clusters of the feature density, a few labeled examples are needed to identify the digit corresponding to each cluster.

greatly simplify the supervised learning task. For example, consider the task of hand-written digit recognition. Training a classifier in the supervised learning setting requires a large number of labeled instances of hand-written digits. Since obtaining labels requires a human to manually identify each character, the labeling process is very time-consuming and expensive. On the other hand, large databases of unlabeled hand-written input images are available. Inspection of the two-dimensional projection of hand-written digit features (first two components obtained by principal component analysis) in Figure 1.3 suggests that unlabeled data can be used to identify the high density regions or clusters of the feature density. Once these clusters, also called decision sets since the label (corresponding digit) is the same over a cluster, are learnt using unlabeled data, the supervised learning task reduces to taking a majority vote in classification (or taking an average in regression) over these sets. Thus, few labeled examples are required to learn the label of each cluster.

Functional Neuroimaging: Accurate identification of regions of brain activity from noisy neuroimaging data, obtained for example using functional magnetic resonance imaging (fMRI), is important for understanding the functioning of the brain. Level set estimation can be used to detect and localize regions of brain activity [18, 19]. Unlike conventional approaches to fMRI data analysis that are based on some form

of thresholding of the activation map at each brain voxel (volumetric element), a set estimation approach has the potential to exploit the correlation amongst neighboring voxels to boost the signal-to-noise ratio and provide better detection performance.

Environmental monitoring using sensor networks: Sensor networks are widely used for spatial mapping of various environmental phenomena, for example solar radiation intensity, rainfall, temperature and humidity, that provide useful information for understanding an ecosystem such as a forest canopy [20, 21, 29, 30]. As sensors are energy and bandwidth constrained, attention is often focused on detection of boundaries or contours where the value of the environmental variable exhibits a changepoint or discontinuity. Examples include monitoring of atmospheric pressure isolines for predicting weather patterns and identifying boundaries of oil-spills or other contaminants.

1.2 Preliminaries

In this section, we introduce the minimax learning framework that serves as a basis for the theoretical aspects of the research presented in this thesis. Statistical learning aims at learning a target function f^* based on n independent, identically distributed observations $\{Z_i\}_{i=1}^n$. The output of a learning algorithm is denoted by $\widehat{f}$. For supervised learning tasks such as classification or regression, $Z_i = (X_i, Y_i)$ denote feature-label pairs that are distributed according to some unknown joint distribution P_{XY}, whereas for unsupervised learning tasks such as density estimation or density level set estimation, $Z_i = X_i$ are distributed according to an unknown marginal distribution P_X. The performance of a candidate mapping or learner f is evaluated using an error measure or risk function denoted as $\mathcal{R}(f)$, which is the expected value of some loss function $\ell(f)$. For example, in classification the risk is typically defined as the probability of error $\mathcal{R}(f) = P(f(X) \neq Y) = \mathbb{E}[\mathbb{I}_{f(X)\neq Y}]$ and the loss function corresponds to the 0-1 loss, in regression the mean square error (MSE) risk is popular $\mathcal{R}(f) = \mathbb{E}[(f(X) - Y)^2]$, where $\mathbb{E}$ denotes the expectation with respect to the joint distribution P_{XY}, and in density estimation the risk is defined as the expected value of the

negative log likelihood loss function $\mathbb{E}[-\log f(X)]$, where $\mathbb{E}$ denotes the expectation with respect to P_X. The excess risk of f is defined as $\mathcal{E}(f) = \mathcal{R}(f) - \mathcal{R}^*$, where $\mathcal{R}^*$ denotes infimum of the risk over all possible learners. Notice that the excess risk is a random variable that depends on the random instantiation of the data $\{Z_i\}_{i=1}^n$ used to build the learner. Hence, to evaluate the performance of a learning algorithm that takes as input the data $\{Z_i\}_{i=1}^n$ and outputs the learner $\widehat{f}$, we are interested in the expected excess risk $\mathbb{E}_n[\mathcal{E}(\widehat{f})]$. Here $\mathbb{E}_n$ denotes the expected value with respect to the training sample $\{Z_i\}_{i=1}^n$. We will drop the subscript n for notational convenience.

Since it is not possible to learn any arbitrary function based on a finite training sample of size n, a well-posed learning problem is concerned with learning a target function f^* that belongs to a class of functions $\mathcal{F}^*$ that satisfies certain (mild) assumptions, for example, the class of smooth regression functions. We are interested in controlling the maximum expected excess risk of a learner f over the class of target functions under consideration, $\sup_{f^* \in \mathcal{F}^*} \mathbb{E}[\mathcal{E}(f)]$. The minimum value of the maximum expected excess risk for this class over all possible learners f_n based on n observations yields a finite sample *minimax* lower bound for the learning problem

$$\inf_{f_n} \sup_{f^* \in \mathcal{F}^*} \mathbb{E}[\mathcal{E}(f_n)] \geq \phi_1(n),$$

where $\phi_1(n)$ depends on the size or complexity of the class $\mathcal{F}^*$. As a typical example, $\phi_1(n) = C_1 n^{-r}$, where $C_1, r > 0$ are constants. A minimax lower bound determines the statistical complexity of a learning problem. A learning algorithm is said to be minimax optimal if it yields a learner $\widehat{f}$ that can achieve the minimax lower bound to within a constant factor, that is, there exists a finite sample upper bound on the performance of $\widehat{f}$ given by

$$\sup_{f^* \in \mathcal{F}^*} \mathbb{E}[\mathcal{E}(\widehat{f})] \leq \phi_2(n),$$

where $0 < \liminf_{n\to\infty} \phi_1(n)/\phi_2(n) \leq \limsup_{n\to\infty} \phi_1(n)/\phi_2(n) < \infty$. For example, $\phi_2(n) = C_2 n^{-r}$, where $C_2 \geq C_1 > 0$ is a constant. If the class $\mathcal{F}^*$ is fixed, r characterizes the *uniform rate of error convergence* at which the maximum expected excess risk over the class $\mathcal{F}^*$

converges to zero as the number of training samples n tend to infinity. For example, the minimax rate of mean square error convergence is given as $n^{-\frac{2\alpha}{2\alpha+d}}$ for estimation of Hölder-α smooth d-dimensional regression functions [5, 31]. This rate can be achieved by piecewise polynomial fits over certain number of bins that scale with the number of samples n and smoothness of the function characterized by α.

In this thesis, we will develop and analyze learning algorithms under the minimax framework introduced in this section.

1.3 Contribution

The first part of this thesis attempts to answer two open theoretical questions related to statistical learning using sets under the nonparametric framework. The second part aims to bridge the gap between theoretically optimal and practically useful estimators by investigating two critical applications in biology and networks.

The first problem we address is the estimation of the support set or level set of a density under the Hausdorff metric. In many applications, it is desirable to have a spatially uniform confidence interval around the estimated set. The Hausdorff error controls the maximal deviation between points in the estimated set and true set, and hence guarantees a local and uniform convergence of the set estimate. However, most level set estimation methods [16,22,32–37] optimize a global error criterion based on the symmetric set difference measure that is a measure of the total erroneous region, and does not penalize where the error is incurred. This can result in estimates which veer greatly from the desired level set. There are a few research papers [34,35,38] that address Hausdorff accurate level set estimation but these either focus on very restricted classes of densities or require that the density possesses some smoothness at all points in the domain. Moreover, these methods rely on a priori knowledge of the density regularity in the vicinity of the level of interest. In this thesis, we present a histogram-based level set estimation method that can provide minimax optimal rates of Hausdorff error convergence over a more general class of densities than considered in previous literature, while imposing no smoothness requirements on the density away from the level of

interest. Furthermore, the proposed method is adaptive to unknown local density regularity. To achieve adaptivity, we introduce a novel data-dependent procedure that is specifically tailored to the level set estimation problem. In the context of support set estimation, previous work on Hausdorff control has been confined to uniform densities [39]. For uniform densities, as well as densities that are bounded from below, support set estimation is equivalent to level set estimation at an appropriate level (greater than zero). We characterize minimax optimal rates (lower and upper bounds) of support set estimation for the class of densities that gradually transition to zero. The derived rate of Hausdorff accurate support set estimation is faster than level set estimation (except for densities that transition sharply to zero), as is known to be the case under the symmetric difference measure as well [35].

The second problem we investigate is semi-supervised learning. Since labeled data is expensive and time-consuming to obtain, semi-supervised learning methods focus on the use of unlabeled data to learn the decision sets in supervised learning tasks such as regression and classification. Once the decision sets are learnt, the supervised learning task reduces to a simple majority vote or averaging on each decision set. While semi-supervised methods have enjoyed empirical success in favorable situations, recent attempts at developing a theoretical basis for semi-supervised learning have been mostly pessimistic [26, 27, 40], and only provide a partial and sometimes apparently conflicting ([27] vs. [41]) explanations of whether or not, and to what extent, unlabeled data can help in learning. In this thesis, we develop a rigorous minimax framework to identify situations in which unlabeled data help to improve learning, and characterize the amount of improvement possible, using finite sample error bounds. We demonstrate that there exist general situations under which semi-supervised learning can be significantly better than supervised learning in terms of achieving smaller finite sample error bounds than any supervised learner, and sometimes in terms of an improved rate of error convergence. Moreover, our results also provide a quantification of the relative value of unlabeled to labeled data. The resulting analysis provides a largely missing statistical theory for semi-supervised learning that explains the empirical success of semi-supervised learning in favorable situations.

In the second half of this thesis, we explore two critical applications where statistical inference relying on set estimation is particularly useful. The first application arises in statistical analysis of fMRI data. Current techniques for fMRI data analysis view the brain activity detection problem as a hypothesis testing problem at each brain voxel, and employ some form of voxel-wise thresholding. Since multiple hypotheses (equal to the number of voxels) need to be tested simultaneously, a Bonferroni correction (union bound) is applied to raise the threshold and guarantee an overall false alarm control. However, this results in overly conservative thresholds with very weak detection power. An alternative is to use a data-adaptive threshold to control the expected proportion of false discoveries [42, 43]. This accounts for sparsity of the activation, however voxel based methods assume that activations in the brain are independent from voxel to voxel and do not take advantage of the correlation structure of the activation regions in the brain. We propose a level set based approach to fMRI data analysis that optimizes a set based localization error. The proposed method can adaptively aggregate neighboring voxels based on the correlation map and thus exploits structural information. We present simulation results on synthetic and real fMRI data that reflect the capabilities of level set based techniques to exploit structural information and boost detection compared to standard techniques being used for fMRI neuroimaging.

The second application we address is efficient path planning for mobile sensors that are used in environmental monitoring, such as aquatic and terrestrial ecosystem studies [20], detection of toxic biological and chemical spreads, oil spills, and weather patterns. As mobile sensors are energy constrained, adaptive sampling paths need to be designed that focus on regions where the environmental phenomenon exhibits a changepoint. Based on recent advances in active learning [4], we characterize the optimal tradeoffs between pathlength, latency, and fidelity. We also present a feedback driven path planning scheme that can achieve these optimal tradeoffs and guides mobile sensors along interesting regions of the domain, thus conducting fast and accurate spatial surveys.

1.4 Organization

This thesis is comprised of four main chapters. Each chapter starts with a summary of its contents and is self-contained. The first two chapters present the theoretical aspects of this research, and the last two chapters focus on practical applications.

Chapter 2 presents the Hausdorff accurate level set estimation procedure and establishes its minimax optimality and adaptivity to unknown local density regularity. In this chapter we also set up directions for making the procedure more practical and discuss some open questions.

In Chapter 3 we develop a general rigorous statistical theory for semi-supervised learning that can be used to analyze situations under which unlabeled data can help improve the performance of a supervised learning task, and also quantify the relative value of labeled and unlabeled data.

Chapters 4 and 5 focus on the two applications - fMRI data analysis and path planning for mobile sensors, respectively. Chapter 4 proposes the level set based approach to detecting brain activation in fMRI data and demonstrates the effectiveness of the proposed approach using synthetic and real data sets. Chapter 5 presents the adaptive path planning scheme for mobile sensors. Recent developments in active learning enable a theoretical characterization of the tradeoffs between pathlength, latency and accuracy. The theory and methods are illustrated using a simulation study of water current mapping in a freshwater lake.

Finally, Chapter 6 summarizes the research work presented in this thesis and explores directions for future work.

Chapter 2

Adaptive Hausdorff Estimation of Density Level Sets

Consider the problem of estimating the γ-level set $G_\gamma^* = \{x : f(x) \geq \gamma\}$ of an unknown d-dimensional density function f based on n independent observations $X_1, \ldots, X_n$ from the density. This problem has been addressed under global error criteria related to the symmetric set difference. However, in certain applications such as anomaly detection and clustering, a spatially uniform confidence interval is desired to ensure that the estimated set is close to the target set everywhere. The Hausdorff error criterion provides this degree of uniformity and hence is more appropriate in such situations. The minimax optimal rate of Hausdorff error convergence is known to be $(n/\log n)^{-1/(d+2\alpha)}$ for level sets with boundaries that have a Lipschitz functional form, and where the parameter α characterizes the regularity of the density around the level of interest. However, previously developed estimators are non-adaptive to the density regularity and assume knowledge of α. Moreover, the estimators proposed in previous work achieve the minimax optimal rate for rather restricted classes of sets (for example, the boundary fragment and star-shaped sets) that effectively reduce the set estimation problem to a function estimation problem. This characterization precludes level sets with multiple connected components, which are fundamental to many applications. This chapter presents a fully data-driven procedure that is adaptive to unknown local density regularity, and achieves minimax optimal Hausdorff error control for a class of level sets with very general shapes and multiple connected components.

2.1 Introduction

Level sets provide useful summaries of a function for many applications including clustering [1,12], anomaly detection [16,17,32], functional neuroimaging [18,44], bioinformatics [24], digital elevation mapping [22, 45], and environmental monitoring [29]. In practice, however, the function itself is unknown a priori and only a finite number of observations related to f are available. Here we focus on the density level set problem; extensions to general regression level set estimation should be possible using a similar approach. Let $X_1, \ldots, X_n$ be independent, identically distributed observations drawn from an unknown probability measure P, having density f with respect to the Lebesgue measure, and defined on the domain $\mathcal{X} \subseteq \mathbb{R}^d$. Given a desired density level γ, consider the γ-level set of the density f:

$$G_\gamma^* := \{x \in \mathcal{X} : f(x) \geq \gamma\}$$

The goal of the density level set estimation problem is to generate an estimate $\widehat{G}$ of the level set based on the n observations $\{X_i\}_{i=1}^n$, such that the error between the estimator $\widehat{G}$ and the target set G_γ^*, as assessed by some performance measure which gauges the closeness of the two sets, is small.

Most literature available on level set estimation methods [16, 22, 32–37] considers error measures related to the symmetric set difference,

$$G_1 \Delta G_2 = (G_1 \setminus G_2) \cup (G_2 \setminus G_1). \tag{2.1}$$

Here $G_1 \setminus G_2 = G_1 \cap G_2^c$, where G^c denotes the complement of the set G. For example, in [16, 34, 35, 37] a probability measure of the symmetric set difference is considered, and in [22, 33, 37] a probability measure of weighted symmetric set difference is considered, the weight being proportional to the deviation of the function from the desired level. Symmetric difference error based performance measures are global measures of *average* closeness between two sets and hence may produce estimates that deviate significantly from the desired level set at certain places. However, some applications such as anomaly detection and clustering may require a more local or spatially uniform error measure as provided by the Hausdorff metric,

for example, to preserve topological properties of the level set as in clustering [1, 12, 14], or ensure robustness to outliers in level set based anomaly detection [16, 17, 32] and data ranking [15]. Controlling a measure of the symmetric difference error does not provide this kind of control and does not ensure accurate recovery of the topological features. To see this, consider a level set with two components as depicted in Figure 2.1 as an example. The figure also shows two candidate estimates, one estimate connects the two components by a "bridge" (resulting in a dumbbell shaped set), while the other preserves the (non)-connectivity. However, both candidate sets have the same Lebesgue measure (volume) of symmetric difference, and hence a method that controls the volume of the symmetric set difference may not favor the one that preserves topological properties over the other. Thus, a uniform measure of closeness between sets is necessary in such situations. The Hausdorff error metric is defined as follows between two non-empty sets:

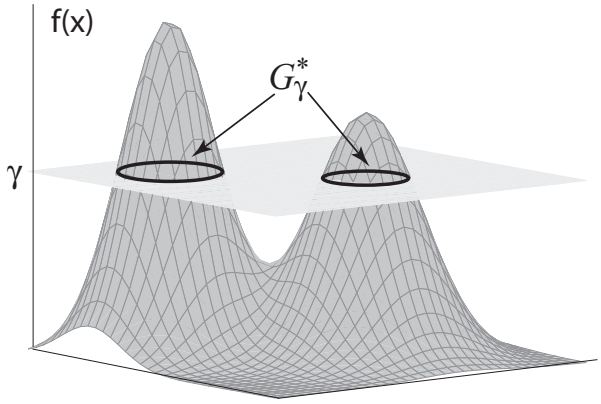
$$d_\infty(G_1, G_2) = \max\{\sup_{x \in G_2} \rho(x, G_1), \sup_{x \in G_1} \rho(x, G_2)\} \tag{2.2}$$

where

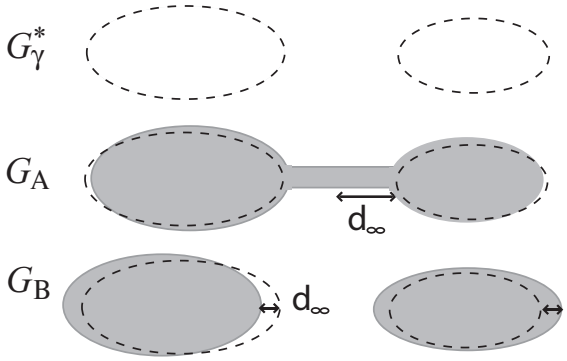
$$\rho(x, G) = \inf_{y \in G} ||x - y||, \tag{2.3}$$

the smallest Euclidean distance of a point in G to the point x. If G_1 or G_2 is empty, then let $d_\infty(G_1, G_2)$ be defined as the largest distance between any two points in the domain. Control of this error measure provides a uniform mode of convergence as it controls the deviation of even a single point from the desired set. In the dumbbell shaped set in Figure 2.1, the Hausdorff error $d_\infty(G_A, G_\gamma^*)$ is proportional to the distance between the clusters (i.e., the length of the bridge). Thus, a set estimate can have very small measure of symmetric set difference but large Hausdorff error. Conversely, as long as the set boundary is not space-filling, small Hausdorff error implies small measure of symmetric difference error.

Existing results pertaining to nonparametric level set estimation using the Hausdorff metric [34, 35, 38] focus on rather restrictive classes of level sets (for example, the boundary fragment and star-shaped set classes). These restrictions, which effectively reduce the set estimation problem to a boundary function estimation problem (in rectangular or polar



(a)



(b)

Figure 2.1 (a) The γ-level set G^*_γ of a density function $f(x)$, (b) Two candidate set estimates G_A and G_B with the same volume of symmetric difference error $\text{vol}(G_A \Delta G^*_\gamma) = \text{vol}(G_B \Delta G^*_\gamma)$, however G_A does not preserve the topological properties (non-connectivity) and has large Hausdorff error $d_\infty(G_A, G^*_\gamma)$, while G_B preserves non-connectivity and has small Hausdorff error $d_\infty(G_B, G^*_\gamma)$.

coordinates, respectively), are typically not met in practical applications. In particular, the characterization of level set estimation as a boundary function estimation problem requires prior knowledge of a reference coordinate or interior point (in rectangular or polar coordinates, respectively) and precludes level sets with multiple connected components. Moreover, the estimation techniques proposed in [34, 35, 38] require precise knowledge of the local regularity of the density (quantified by the parameter α, to be defined below) in the vicinity of the desired level in order to achieve minimax optimal rates of convergence. Such prior knowledge is unavailable in most practical applications. Recently, a plug-in method based on sup-norm density estimation was put forth in [46] that can handle more general classes than boundary fragments or star-shaped sets, however sup-norm density estimation requires the density to behave smoothly everywhere to ensure that the estimate is close to the true density at all points. Also, the method only deals with a special case of the density regularity condition considered here ($\alpha = 1$), and is therefore not adaptive to unknown density regularity.

In this chapter, we propose a plug-in procedure based on a regular histogram partition that can adaptively achieve minimax optimal rates of Hausdorff error convergence over a broad class of level sets with very general shapes and multiple connected components, without assuming *a priori* knowledge of the density regularity parameter α. Adaptivity is achieved by a new data-driven procedure for selecting the histogram resolution. The procedure is reminiscent of Lepski-type methods [47], however it is specifically designed for the level set estimation problem and only requires local regularity of the density in the vicinity of the desired level. While the basic approach is illustrated through the use of histogram-based estimators, extensions to more general partitioning schemes such as spatially adaptive partitions [48–51] might be possible. The theory and method may also provide a useful starting point for future investigations into alternative schemes, such as kernel-based approaches [17], that may be better suited for higher dimensional settings.

To motivate the importance of Hausdorff accurate level set estimation, let us briefly discuss its relevance in some applications.

Clustering - Density levels set estimators are used by many data clustering procedures [1,12,14], and the correct identification of connected level set components (i.e., clusters) is crucial to their success. The Hausdorff criterion can be used to provide theoretical guarantees regarding clustering since the connected components of a level set estimate that is ϵ-accurate in the Hausdorff sense, characterize the true level set clusters (in number, shapes, and locations), provided the true clusters remain topologically distinct upon erosion or dilation by an ϵ-ball. The last statement holds since

$$d_\infty(G_1, G_2) \leq \epsilon \quad \Longrightarrow \quad G_1 \subseteq G_2^\epsilon, \;\; G_2 \subseteq G_1^\epsilon,$$

where G^ϵ denotes the set obtained by dilation of set G by an ϵ-ball.

Data Ranking - Hausdorff accurate level set estimation is also relevant for ranking or ordering data using the notion of data-depth [15]. Density level sets correspond to likelihood-depth contours and Hausdorff distance offers a robust measure of accuracy in estimating the data-depth as it is less susceptible to severe misranking, as compared to symmetric set difference based measures.

Anomaly detection - A common approach to anomaly detection is to learn a (high) density level set of the nominal data distribution [16,17,32]. Samples that fall outside the level set, in the low density region, are considered anomalies. Level set methods based on a symmetric difference error measure may produce estimates that veer greatly from the desired level set at certain places and potentially include regions of low density, since the symmetric difference is a global error measure. Anomalous distributions concentrated in such places would elude detection. On the other hand, level set estimators based on the Hausdorff metric are guaranteed to be uniformly close to the desired level set, and therefore are more robust to anomalies in such situations.

Semi-supervised learning - Unlabeled data can be used, along with labeled data, to improve the performance of a supervised learning task in certain favorable situations. One such situation, commonly called the cluster assumption, is where the regression

function is constant or smooth in high density regions [26, 28]. As we will discuss in Chapter 3, improved error bounds can be obtained if these decision regions (corresponding to connected components of the support set) can be learnt using unlabeled data, followed by simple averaging or majority vote on each component to predict the label which requires few labeled examples. Correct identification of the connected components of the support set is crucial to obtaining improved error bounds, and hence a uniform control provided by the Hausdorff error is needed.

Thus, Hausdorff accurate estimation of density level sets is an important problem with many potential applications. However, in all these applications there are other issues, for example, selection of the density levels of interest, that are beyond the scope of this dissertation.

This chapter is organized as follows. Section 2.2 states our basic assumptions which allow Hausdorff accurate level set estimation and also presents a minimax lower bound on the Hausdorff performance of any level set estimator for the class of densities under consideration. Section 2.3 discusses the issue with direct Hausdorff estimation and provides motivation for an alternate error measure. In Section 2.4, we present the proposed histogram-based approach to Hausdorff accurate level set estimation that can achieve the minimax optimal rate of convergence, given knowledge of the density regularity parameter α. Subsection 2.4.1 extends the proposed estimator to achieve adaptivity to unknown density regularity. Subsections 2.4.2-2.4.4 present extensions that address simultaneous estimation of multiple level sets, support set estimation, and discontinuity in the density around the level of interest. Concluding remarks are given in Section 2.5 and Section 2.6 contains the proofs.

2.2 Density assumptions

We assume that the domain of the density f is the unit hypercube in d-dimensions, i.e. $\mathcal{X} = [0,1]^d$. Extensions to other compact domains are straightforward. Furthermore, the density is assumed to be bounded with range $[0, f_{\max}]$, though knowledge of $f_{\max}$ is not assumed. Controlling the Hausdorff accuracy of level set estimates requires some smoothness

assumptions on the density and the level set boundary, which are stated below. But before that we introduce some definitions:

- **ϵ-Ball:** An ϵ-ball centered at a point $x \in \mathcal{X}$ is defined as

$$B(x, \epsilon) = \{y \in \mathcal{X} : ||x - y|| \leq \epsilon\}.$$

  Here $|| \cdot ||$ denotes the Euclidean distance.

- **Inner ϵ-cover:** An inner ϵ-cover of a set $G \subseteq \mathcal{X}$ is defined as the union of all ϵ-balls contained in G. Formally,

$$\mathcal{I}_\epsilon(G) = \bigcup_{x : B(x,\epsilon) \subseteq G} B(x, \epsilon)$$

We are now ready to state the assumptions. The most crucial one is the first, which characterizes the relationship between distances and changes in density, and the second one is a topological assumption on the level set boundary that essentially generalizes the notion of Lipschitz functions to closed hypersurfaces.

[A] *Local density regularity:* The density is α-regular around the γ-level set, $0 < \alpha < \infty$ and $0 < \gamma < f_{\max}$, if

[A1] there exist constants $C_1, \delta_1 > 0$ such that for all $x \in \mathcal{X}$ with $|f(x) - \gamma| \leq \delta_1$,

$$|f(x) - \gamma| \geq C_1 \rho(x, \partial G^*_\gamma)^\alpha,$$

where ∂G^*_γ denotes the boundary of the true level set G^*_γ and $\rho(\cdot, \cdot)$ is as defined in (2.3).

[A2] there exist constants $C_2, \delta_2 > 0$ and a point $x_0 \in \partial G^*_\gamma$ such that for all $x \in B(x_0, \delta_2)$,

$$|f(x) - \gamma| \leq C_2 \rho(x, \partial G^*_\gamma)^\alpha.$$

This condition characterizes the behavior of the density around the level γ. **[A1]** states that the density cannot be arbitrarily "flat" around the level, and in fact the deviation

of the density from level γ is at least the α-th power of the distance from the level set boundary. [**A2**] states that there exists a fixed neighborhood around some point on the boundary where the density changes no faster than the α-th power of the distance from the level set boundary. The latter condition is only required for adaptivity, as we discuss later. The regularity parameter α determines the rate of error convergence for level set estimation. Accurate estimation is more difficult at levels where the density is relatively flat (large α), as intuition would suggest. It is important to point out that we do not assume knowledge of α unlike previous investigations into Hausdorff accurate level set estimation [34, 35, 38, 46]. Therefore, here the assumption simply states that there is a relationship between distance and density level, but the precise nature of the relationship is unknown. We discuss extensions to address support set estimation ($\gamma = 0$) in Subsection 2.4.3 and the case $\alpha = 0$ (which corresponds to a jump in the density at level γ) in Subsection 2.4.4.

[**B**] *Level set regularity:* There exist constants $\epsilon_o > 0$ and $C_3 > 0$ such that for all $\epsilon \leq \epsilon_o$, $\mathcal{I}_\epsilon(G^*_\gamma) \neq \emptyset$ and $\rho(x, \mathcal{I}_\epsilon(G^*_\gamma)) \leq C_3\epsilon$ for all $x \in \partial G^*_\gamma$.

This assumption states that the level set is not arbitrarily narrow anywhere. It precludes features like cusps and arbitrarily thin ribbons, as well as isolated connected components of arbitrarily small size. This condition is necessary since arbitrarily small features cannot be detected and resolved from a finite sample. However, from a practical perspective, if the assumption fails to hold then it simply means that it is not possible to theoretically guarantee that such small features will be recovered.

For a fixed set of positive numbers C_1, C_2, C_3, ϵ_0, δ_1, δ_2, $f_{\max}$, $\gamma < f_{\max}$, d and α, we consider the following classes of densities:

Definition 1. *Let $\mathcal{F}^*_1(\alpha)$ denote the class of densities satisfying assumptions **[A1,B]**.*

Definition 2. *Let $\mathcal{F}^*_2(\alpha)$ denote the class of densities satisfying assumptions **[A1,A2,B]**.*

The dependence on other parameters is omitted as these do not influence the minimax optimal rate of convergence (except the dimension d). We present a method that provides minimax optimal rates of convergence for the class $\mathcal{F}_1^*(\alpha)$, given knowledge of the density regularity parameter α. We also extend the method to achieve adaptivity to α for the class $\mathcal{F}_2^*(\alpha)$, while preserving the minimax optimal performance.

Assumption **[A]** is similar to the one employed in [35, 38], except that the upper bound assumption on the density deviation in [35,38] holds provided that the set $\{x : |f(x)-\gamma| \leq \delta_1\}$ is non-empty. This implies that the densities either jump across the level γ at any point on the level set boundary (that is, the deviation is greater than δ_1) or change exactly as the α^{th} power of the distance from the boundary. Our formulation allows for densities with regularities that vary spatially along the level set boundary - it requires that the density changes no slower than the α^{th} power of the distance from the boundary, except in a fixed neighborhood of one point where the density changes exactly as the α^{th} power of the distance from the boundary. While the formulation in [35,38] requires the upper bound on the density deviation to hold for at least one point on the boundary, our assumption **[A2]** requires the upper bound to hold for a fixed neighborhood about at least one point on the boundary. This is necessary for adaptivity since a procedure cannot sense the regularity as characterized by α if the regularity only holds in an arbitrarily small region. Assumption **[B]** basically implies that the boundary looks locally like a Lipschitz function and allows for level sets with multiple connected components and arbitrary locations. Thus, these restrictions are quite mild and less restrictive than those considered in the previous literature on Hausdorff level set estimation. In fact **[B]** is satisfied by a Lipschitz boundary fragment or star-shaped set as considered in [34, 35, 38] as the following lemma states.

Lemma 1. *Consider the γ level set G_γ^* of a density $f \in \mathcal{F}_{SL}(\alpha)$, where $\mathcal{F}_{SL}(\alpha)$ denotes the class of α-regular densities with Lipschitz star-shaped level sets as defined in [35]. Then G_γ^* satisfies the level set regularity assumption **[B]**.*

The proof is given in Section 2.6.1. Thus, the classes under consideration here are more general, except for the exclusion of densities for which the upper bound on the local density regularity assumption **[A2]** only holds in a region of arbitrarily small Lebesgue measure.

Tsybakov establishes a minimax lower bound of $(n/\log n)^{-1/(d+2\alpha)}$ in Theorem 4 of [35] for the class of star-shaped sets with Lipschitz boundaries, which as per Lemma 1 also satisfy assumption **[B]**. His proof uses Fano's lemma to derive the lower bound for a discrete subset of densities from this class. It is easy to see that the discrete subset of densities used in his construction also satisfy our form of assumption **[A]**. Hence, the same minimax lower bound holds for the classes $\mathcal{F}_1^*(\alpha), \mathcal{F}_2^*(\alpha)$ under consideration as well and we have the following proposition. Proof of the proposition is given in Section 2.6.2. Here $\mathbb{E}$ denotes expectation with respect to the random data sample.

Proposition 1. *There exists $c > 0$ such that, for large enough n,*

$$\inf_{G_n} \sup_{f \in \mathcal{F}_1^*(\alpha)} \mathbb{E}[d_\infty(G_n, G_\gamma^*)] \geq \inf_{G_n} \sup_{f \in \mathcal{F}_2^*(\alpha)} \mathbb{E}[d_\infty(G_n, G_\gamma^*)] \geq c \left(\frac{n}{\log n}\right)^{-\frac{1}{d+2\alpha}}.$$

The inf *is taken over all set estimators G_n based on the n observations.*

2.3 Motivating an Error Measure for Hausdorff control

Direct Hausdorff accurate set estimation is challenging as there exists no natural empirical measure that can be used to gauge the Hausdorff error of a set estimate. In this section, we investigate how Hausdorff control can be obtained indirectly using an alternate error measure that is based on density deviation error rather than distance deviation. While the first alternate error measure we introduce is easily motivated and arises naturally, it requires the density to have some smoothness everywhere, whereas only local smoothness in the vicinity of the level set is required for accurate level set estimation. Based on these insights, we propose our final alternate measure. If we focus on candidate set estimates based on a regular histogram, minimizing this final alternate error measure leads to a simple plug-in level set estimator that is the focus of this chapter. However, we introduce the general

error measure since it offers the potential to extend the proposed technique to more general estimators based on spatially adapted partitions or kernel based methods.

The density regularity condition [**A**] suggests that control over the deviation of any point in the estimate from the true level set boundary $\rho(x, \partial G^*_\gamma)$ can be obtained by controlling the deviation from the desired density level. In other words, a change in density level reflects change in distance. Moreover, in order to obtain a sense of distance from an estimate of density variation based on a small sample, the level set boundary cannot vary too irregularly. Specifically, the boundary should not have arbitrarily small features (e.g., cusps) that cannot be reliably detected from a small sample. Such features are ruled-out by assumption [**B**]. Thus, under regularity conditions on the function and level set boundary, the deviation of the density function from the desired level can be used as a surrogate for the Hausdorff error. Consider the following error measure:

$$\begin{aligned} \mathcal{E}(G) &= \max\{ \sup_{x \in G^*_\gamma \setminus G} (f(x) - \gamma), \sup_{x \in G \setminus G^*_\gamma} (\gamma - f(x))\} && (2.4) \\ &= \sup_{x \in \mathcal{X}} (\gamma - f(x))[\mathbf{1}_{x \in G} - \mathbf{1}_{x \notin G}] && (2.5) \end{aligned}$$

where $\mathbf{1}$ denotes the indicator function and by convention $\sup_{x \in \emptyset} g(x) = 0$ for any non-negative function $g(\cdot)$. The error measure $\mathcal{E}(G)$ has a natural empirical counterpart, $\widehat{\mathcal{E}}(G)$, obtained by simply replacing $f(x)$ by a density estimator $\widehat{f}(x)$. Notice that the set $\widehat{G}$ minimizing the empirical error corresponds to a plug-in level set estimator, that is $\widehat{G} = \{x : \widehat{f}(x) \geq \gamma\}$ [1]. Also

$$\mathcal{E}(\widehat{G}) = \max\{ \sup_{x \in G^*_\gamma \setminus \widehat{G}} (f(x) - \gamma), \sup_{x \in \widehat{G} \setminus G^*_\gamma} (\gamma - f(x))\} \leq \sup_{x \in \widehat{G} \Delta G^*_\gamma} |f(x) - \widehat{f}(x)|.$$

The last step follows since a point $x \in \widehat{G} \Delta G^*_\gamma$ is erroneously included or excluded from the level set estimate, and hence for $x \in \widehat{G} \setminus G^*_\gamma$, $\gamma - f(x) \leq |f(x) - \widehat{f}(x)|$ and for $x \in G^*_\gamma \setminus \widehat{G}$, $f(x) - \gamma \leq |f(x) - \widehat{f}(x)|$. Using this error measure, we have the following Hausdorff control.

[1] Actually the set $\widehat{G}$ is not unique since the points x with $\widehat{f}(x) = \gamma$ may or may not be included in the estimate.

Proposition 2. *If the sup norm error between $\widehat{f}(x)$ and the true density $f(x)$ converges in probability to zero and $\widehat{G}$ denotes the corresponding plug-in level set estimate, then under assumptions **[A]** and **[B]**, there exists a constant $C > 0$ such that for large enough n, with high probability*

$$d_\infty(\widehat{G}, G^*_\gamma) \leq C\, \mathcal{E}(\widehat{G})^{1/\alpha} \leq C \left(\sup_{x \in \widehat{G} \Delta G^*_\gamma} |f(x) - \widehat{f}(x)| \right)^{1/\alpha} \leq C \left(\sup_{x \in \mathcal{X}} |f(x) - \widehat{f}(x)| \right)^{1/\alpha}.$$

The proof is given in Section 2.6.3. This result shows that the sup-norm error of a density estimate gives an upper bound on the Hausdorff error of a plug-in level set estimate, which agrees with Cuevas' result [46] for $\alpha = 1$. However, arbitrarily rough and complicated behavior of the density away from the level of interest can cause a large sup-norm density error, whereas the Hausdorff accuracy of a level set estimate should only depends on the accuracy of the density estimate around the level of interest. Therefore, we follow Vapnik's maxim: *When solving a given problem, try to avoid solving a more general problem as an intermediate step* [52], and instead of solving the harder intermediate problem of sup-norm density estimation (which requires some smoothness of the density at all points), we approach the set estimation problem directly.

We now consider a modified version of the error measure introduced above. Let Π denote a partition of $[0,1]^d$ and let G be any set defined in terms of this partition (i.e., the union of any collection of cells of the partition). We will consider a hierarchy of partitions with increasing complexity and the sets G, defined in terms of the partitions, form candidate representations of the γ level set of the density f. The partition could, for example, correspond to a decision tree or regular histogram. Define the error of G as

$$\mathcal{E}_\gamma(G) = \sup_{A \in \Pi(G)} (\gamma - \bar{f}(A))[\mathbf{1}_{A \subseteq G} - \mathbf{1}_{A \nsubseteq G}].$$

Here $\Pi(G)$ denotes the partition associated with set G and $\bar{f}(A) = P(A)/\mu(A)$ denotes average of the density function on the cell A, where P is the unknown probability measure and μ is the Lebesgue measure. Note the analogy between this error and that defined in (2.4). We would like to point out that even though this error depends on the class of candidate

sets being considered, it can be used to establish control over the Hausdorff error which is independent of the candidate class. This performance measure evaluates a set based on the maximum deviation of the average density in a cell of the partition from the γ level. Note that $\left(\gamma - \bar{f}(A)\right)\left[\mathbf{1}_{A\subseteq G} - \mathbf{1}_{A\nsubseteq G}\right] > 0$ whenever a cell with average density $\bar{f}(A) < \gamma$ is included in the set G or a cell with $\bar{f}(A) > \gamma$ is excluded. A natural empirical error, $\widehat{\mathcal{E}}(G)$, is obtained by replacing $\bar{f}(A)$ with its empirical counterpart.

$$\widehat{\mathcal{E}}_\gamma(G) = \max_{A\in\Pi(G)} \left(\gamma - \widehat{f}(A)\right)\left[\mathbf{1}_{A\subseteq G} - \mathbf{1}_{A\nsubseteq G}\right]$$

Here $\widehat{f}(A) = \frac{\widehat{P}(A)}{\mu(A)}$, where $\widehat{P}(A) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{\{X_i\in A\}}$ denotes the empirical probability of an observation occurring in A. Among all sets that are based on the same partition, the one minimizing the empirical error $\widehat{\mathcal{E}}_\gamma$ is a natural candidate:

$$\widehat{G}_{\Pi_o} = \arg\min_{G:\Pi(G)=\Pi_o} \widehat{\mathcal{E}}_\gamma(G) \tag{2.6}$$

This rule selects the set that includes all cells with empirical density $\widehat{f}(A) > \gamma$ and excludes all cells with $\widehat{f}(A) < \gamma$, hence it is essentially a plug-in level set estimator. We focus on sets based on a uniform histogram partition and establish that minimizing the empirical error $\widehat{\mathcal{E}}_\gamma(G)$, along with appropriate choice of the histogram resolution, is sufficient for Hausdorff control. The appropriate histogram resolution depends only on the *local* regularity of the density around the level of interest. Furthermore, we show that the histogram resolution can be chosen adaptively in a purely data-driven way without assuming knowledge of the local density regularity. The performance of the regular histogram-based level set estimator is shown to be minimax optimal for the class of densities $\mathcal{F}_1^*(\alpha)$ (assuming knowledge of the local density regularity parameter α) and $\mathcal{F}_2^*(\alpha)$ (using an adaptive procedure, to be defined later).

Remark 1: In practice, estimators based on spatially adapted partitions can provide better performance since they can adapt to the spatial variations in density regularity to yield better estimate of the boundary where the density changes sharply, though the overall Hausdorff error is dominated by the accuracy achievable in estimating the boundary where the density

changes slowly. Thus, it is of interest to develop spatially adapted estimators. While, in the context of histogram based set estimators, only an appropriate choice of the resolution is needed, spatially adapted estimators require a more sophisticated procedure for selecting the appropriate partition. We do not address this aspect here, however the set up described above can serve as a useful starting point.

2.4 Hausdorff accurate Level Set Estimation using Histograms

Let $\mathcal{A}_j$ denote the collection of cells in a regular partition of $[0,1]^d$ into hypercubes of dyadic sidelength 2^{-j}, where j is a non-negative integer. The level set estimate at this resolution is given as

$$\widehat{G}_j = \bigcup_{A \in \mathcal{A}_j : \widehat{f}(A) \geq \gamma} A \tag{2.7}$$

Here $\widehat{f}(A) = \widehat{P}(A)/\mu(A)$, where $\widehat{P}(A) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{\{X_i \in A\}}$ denotes the empirical probability of an observation occurring in A and μ is the Lebesgue measure.

The appropriate resolution for accurate level set estimation depends on the local density regularity, as characterized by α, near the level of interest. If the density varies sharply (small α) near the level of interest, then accurate estimation is easier and a fine resolution suffices. Identifying the level set is more difficult if the density is very flat (large α) and hence a lower resolution (more averaging) is required. Our first result shows that, if the local density regularity parameter α is known, then the correct resolution j can be chosen (as in [35,38]), and the corresponding estimator achieves near minimax optimal rate over the class of densities given by $\mathcal{F}_1^*(\alpha)$. Notice that even though the proposed method is a plug-in level set estimator based on a histogram density estimate, the histogram resolution is chosen to specifically target the level set problem and is not optimized for density estimation. Thus, we do not require that the density exhibits some smoothness everywhere. We introduce the notation $a_n \asymp b_n$ to denote that $a_n = O(b_n)$ and $b_n = O(a_n)$.

Theorem 1. *Assume that the local density regularity α is known. Pick resolution $j \equiv j(n)$ such that $2^{-j} \asymp s_n(n/\log n)^{-\frac{1}{(d+2\alpha)}}$, where s_n is a monotone diverging sequence. Then*

$$\sup_{f \in \mathcal{F}_1^*(\alpha)} \mathbb{E}[d_\infty(\widehat{G}_j, G_\gamma^*)] \leq C s_n \left(\frac{n}{\log n}\right)^{-\frac{1}{d+2\alpha}}$$

for all n, where $C \equiv C(C_1, C_3, \epsilon_o, f_{\max}, \delta_1, d, \alpha) > 0$ is a constant.

The proof is given in Section 2.6.4 and relies on two key facts. First, the density regularity assumption [**A1**] implies that the distance of any point in the level set estimate is controlled by its deviation from the level of interest γ. This implies that, with high probability, only the cells near the boundary are erroneously included or excluded in the level set estimate. Second, the level set boundary does not have very narrow features that cannot be detected by a finite sample and is locally Lipschitz as per assumption [**B**]. Using these facts, it follows that the Hausdorff error scales as the histogram sidelength.

Theorem 1 provides an upper bound on the Hausdorff error of our estimate. If s_n is slowly diverging, for example if $s_n = (\log n)^\epsilon$ where $\epsilon > 0$, this upper bound agrees with the minimax lower bound of Proposition 1 up to a $(\log n)^\epsilon$ factor. Hence the proposed estimator can achieve near minimax optimal rates, given knowledge of the density regularity. We would like to point out that if the parameter δ_1 characterizing assumption [**A**] and the density bound $f_{\max}$ are also known, then the appropriate resolution can be chosen as $j = \lfloor \log_2 \left(c^{-1}(n/\log n)^{1/(d+2\alpha)}\right) \rfloor$, where the constant $c \equiv c(\delta_1, f_{\max})$. With this choice, the optimal sidelength scales as $2^{-j} \asymp (n/\log n)^{-1/(d+2\alpha)}$, and the estimator $\widehat{G}_j$ exactly achieves the minimax optimal rate.

Remark 2: A dyadic sidelength is not necessary for Theorem 1 to hold, however the adaptive procedure described next is based on a search over dyadic resolutions. Thus, to present a unified analysis, we consider a dyadic sidelength here too.

2.4.1 Adaptivity to unknown density regularity

In this section we present a procedure that automatically selects the appropriate resolution in a purely data-driven way without assuming prior knowledge of α. The proposed procedure is a complexity regularization approach that is reminiscent of Lepski-type methods for function estimation [47], which are spatially adaptive bandwidth selectors. In Lepski methods, the appropriate bandwidth at a point is determined as the largest bandwidth for which the estimate does not deviate significantly from estimates generated at finer resolutions. Our procedure is similar in spirit, however it is tailored specifically for the level set problem and hence the chosen resolution at any point depends only on the local regularity of the density around the level of interest.

The histogram resolution search is focused on regular partitions of dyadic sidelength 2^{-j}, $j \in \{0, 1, \ldots, J\}$. The choice of J will be specified below. Since the selected resolution needs to be adapted to the local regularity of the density around the level of interest, we introduce the following *vernier*:

$$\mathcal{V}_{\gamma,j} = \min_{A \in \mathcal{A}_j} \max_{A' \in \mathcal{A}_{j'} \cap A} |\gamma - \bar{f}(A')|.$$

Here $\bar{f}(A) = P(A)/\mu(A)$, $j' = \lfloor j + \log_2 s_n \rfloor$, where s_n is a slowly diverging monotone sequence, for example $\log n$, $\log \log n$, etc., and $\mathcal{A}_{j'} \cap A$ denotes the collection of subcells with sidelength $2^{-j'} \in [2^{-j}/s_n, 2^{-j+1}/s_n)$ within the cell A. Observe that the vernier value is determined by a cell $A \in \mathcal{A}_j$ that intersects the boundary ∂G^*_γ. By evaluating the deviation in average density from level γ within subcells of A, the vernier indicates whether or not the density in cell A is uniformly close to γ. Thus, the vernier is sensitive to the local density regularity in the vicinity of the desired level and leads to selection of the appropriate resolution adapted to the unknown density regularity parameter α, as we will show in Theorem 2.

Since $\mathcal{V}_{\gamma,j}$ requires knowledge of the unknown probability measure, we must work with the empirical version, defined analogously as:

$$\widehat{\mathcal{V}}_{\gamma,j} = \min_{A \in \mathcal{A}_j} \max_{A' \in \mathcal{A}_{j'} \cap A} |\gamma - \widehat{f}(A')|.$$

The empirical vernier $\widehat{\mathcal{V}}_{\gamma,j}$ is balanced by a penalty term:

$$\Psi_{j'} := \max_{A\in\mathcal{A}_{j'}} \sqrt{8\frac{\log(2^{j'(d+1)}16/\delta)}{n\mu(A)} \max\left(\widehat{f}(A), 8\frac{\log(2^{j'(d+1)}16/\delta)}{n\mu(A)}\right)}$$

where $0 < \delta < 1$ is a confidence parameter, and $\mu(A) = 2^{-j'd}$. Notice that the penalty is computable from the given observations. The precise form of Ψ is chosen to bound the deviation of true and empirical vernier with high probability (refer to Corollary 3 for a formal proof). The final level set estimate is given by

$$\widehat{G} = \widehat{G}_{\widehat{j}} \tag{2.8}$$

where

$$\widehat{j} = \arg\min_{0\le j\le J}\left\{\widehat{\mathcal{V}}_{\gamma,j} + \Psi_{j'}\right\} \tag{2.9}$$

Observe that the value of the vernier decreases with increasing resolution as better approximations to the true level are available. On the other hand, the penalty is designed to increase with resolution to penalize high complexity estimates that might overfit the given sample of data. Thus, the above procedure chooses the appropriate resolution automatically by balancing these two terms. The following theorem characterizes the performance of the proposed complexity penalized procedure.

Theorem 2. *Pick $J \equiv J(n)$ such that $2^{-J} \asymp s_n(n/\log n)^{-\frac{1}{d}}$, where s_n is a monotone diverging sequence. Let $\widehat{j}$ denote the resolution chosen by the complexity penalized method as given by (2.9), and $\widehat{G}$ denote the final estimate of (2.8). Then with probability at least $1 - 2/n$, for all densities in the class $\mathcal{F}_2^*(\alpha)$,*

$$c_1 s_n^{\frac{d}{d+2\alpha}}\left(\frac{n}{\log n}\right)^{-\frac{1}{d+2\alpha}} \le 2^{-\widehat{j}} \le c_2 s_n s_n^{\frac{d}{d+2\alpha}}\left(\frac{n}{\log n}\right)^{-\frac{1}{d+2\alpha}}$$

for n large enough (so that $s_n > c(C_3, \epsilon_o, d)$), where $c_1, c_2 > 0$ are constants. In addition,

$$\sup_{f\in\mathcal{F}_2^*(\alpha)} \mathbb{E}[d_\infty(\widehat{G}, G_\gamma^*)] \le C s_n^2 \left(\frac{n}{\log n}\right)^{-\frac{1}{d+2\alpha}}$$

for all n, where $C \equiv C(C_1, C_2, C_3, \epsilon_o, f_{\max}, \delta_1, \delta_2, d, \alpha) > 0$ is a constant.

The proof is given in Section 2.6.5. Observe that the maximum resolution $2^J \asymp s_n^{-1}(n/\log n)^{\frac{1}{d}}$ can be easily chosen, based only on n, and allows the optimal resolution for any α to lie in the search space. By appropriate choice of s_n, for example $s_n = (\log n)^{\epsilon/2}$ with ϵ a small number > 0, the bound of Theorem 2 matches the minimax lower bound of Proposition 1, except for an additional $(\log n)^{\epsilon}$ factor. Hence our method *adaptively* achieves near minimax optimal rates of convergence for the class $\mathcal{F}_2^*(\alpha)$.

Remark 3: To prove the results of Theorems 1 and 2, we do not need to assume an exact form for s_n except that it is a monotone diverging sequence. However, s_n needs to be slowly diverging for the derived rates to be near minimax optimal.

Remark 4: It should be noted that there is a price of adaptivity. To obtain a rate that is very close to the minimax optimal rate, we desire s_n to increase very slowly with n. However, the slower s_n grows, the more the number of samples required to meet the condition $s_n > c(C_3, \epsilon_o, d)$ and obtain a useful (non-trivial) bound.

Remark 5: We would like to point out that even though we state the convergence results in expectation, the proofs also establish high probability confidence bounds.

2.4.2 Multiple level set estimation

The proposed framework can easily be extended to simultaneous estimation of level sets at multiple levels $\Gamma = \{\gamma_k\}_{k=1}^K$ $(K < \infty)$. Assuming the density regularity condition **[A]** holds with parameter α_k for the γ_k level, we have the following corollary that is a direct consequence of Theorem 2.

Corollary 1. *Pick $J = J(n)$ such that $2^{-J} \asymp s_n(n/\log n)^{-1/d}$, where s_n is a monotone diverging sequence. Let $\widehat{G}_{\gamma_k}$ denote the estimate generated using the complexity penalized procedure of (2.8) for level γ_k. Then*

$$\max_{1\le k\le K} \sup_{f\in\mathcal{F}_2^*(\alpha_k)} \mathbb{E}[d_\infty(\widehat{G}_{\gamma_k}, G^*_{\gamma_k})] \le C s_n^2 \left(\frac{n}{\log n}\right)^{-1/(d+2\max_k \alpha_k)}$$

for all n, here $C \equiv C(C_1, C_2, C_3, \epsilon_o, f_{\max}, \delta_1, \delta_2, d, \{\alpha_k\}_{k=1}^K) > 0$.

Notice that, while the estimate $\widehat{G}_{\gamma_k}$ at each level is adaptive to the local density regularity as determined by α_k, the overall convergence rate is determined by the level where the density is most flat (largest α_k).

Another issue that comes up in multiple level set estimation is nestedness. If the density levels of interest Γ are sorted, $\gamma_1 \leq \gamma_2 \leq \cdots \leq \gamma_K$, then the true level sets will be nested $G^*_{\gamma_1} \supseteq G^*_{\gamma_2} \supseteq \cdots \supseteq G^*_{\gamma_K}$. However, the estimates $\{\widehat{G}_{\gamma_k}\}_{k=1}^K$ may not be nested as the resolution at each level is determined by the local density regularity (α_k). For some applications, for example hierarchical clustering, nested estimates may be desirable. We can enforce this by choosing the same resolution, corresponding to the largest α_k, at all levels. Since the largest α_k corresponds to smallest vernier $\mathcal{V}_{\gamma_k,j}$ (see Lemma 5), nested level set estimates can be generated by selecting the resolution according to

$$\widehat{j} = \arg\min_{0 \leq j \leq J} \left\{ \min_{1 \leq k \leq K} \widehat{\mathcal{V}}_{\gamma_k,j} + \Psi_{j'} \right\}.$$

This does not change the rate of convergence, however if the density is flat at one level of interest, this forces large Hausdorff error at all levels, even if the density at those levels is well-behaved (varies sharply near the level of interest).

2.4.3 Support set estimation

In the earlier analysis, we assumed that the level of interest $\gamma > 0$. The case $\gamma = 0$ corresponds to estimating the support set of the density function, which is defined as

$$G^*_0 := \{x : f(x) > 0\}.$$

In the context of symmetric difference error, it is known [35,39,53] that support set estimation is easier than level set estimation (except for the case $\alpha = 0$ when the density exhibits a discontinuity around the level of interest). We show that the same holds for Hausdorff error and the minimax rate of convergence is given as $(n/\log n)^{-1/(d+\alpha)}$ which is faster than the rate of $(n/\log n)^{-1/(d+2\alpha)}$ for level set estimation. For support set estimation the density

regularity assumption [**A**] holds only for points lying in the support of the density, that is [**A1, A2**] hold only for $x \in G_0^*$.

First, we establish the minimax lower bound. We actually establish the bound for the class of densities $\mathcal{F}_{BF}(\alpha, \zeta)$ that satisfy the local density regularity [**A1, A2**] for all $x \in G_0^*$ and whose support sets G_0^* are Hölder-ζ boundary fragments. The case $\zeta = 1$ corresponds to the class of Lipschitz boundary fragments which satisfy assumption [**B**] (this can be shown along the lines of the proof of Lemma 1). Thus, $\mathcal{F}_{BF}(\alpha, 1)$ is a subclass of the classes $\mathcal{F}_1^*(\alpha), \mathcal{F}_2^*(\alpha)$ under consideration, and hence a lower bound for $\mathcal{F}_{BF}(\alpha, 1)$ yields a corresponding lower bound for these classes.

Proposition 3. *There exists $c > 0$ such that*

$$\inf_{G_n} \sup_{f \in \mathcal{F}_{BF}(\alpha,\zeta)} \mathbb{E}[d_\infty(G_n, G_0^*)] \geq c \left(\frac{n}{\log n}\right)^{-\frac{\zeta}{d-1+\zeta(\alpha+1)}}.$$

for n large enough. This implies that

$$\inf_{G_n} \sup_{f \in \mathcal{F}_1^*(\alpha)} \mathbb{E}[d_\infty(G_n, G_0^*)] \geq \inf_{G_n} \sup_{f \in \mathcal{F}_2^*(\alpha)} \mathbb{E}[d_\infty(G_n, G_0^*)] \geq c \left(\frac{n}{\log n}\right)^{-\frac{1}{d+\alpha}}$$

for n large enough. The inf *is taken over all possible set estimators G_n based on the n observations.*

The proof is given in Section 2.6.6 and requires a construction similar to the minimax lower bound derived in [35] for level set estimation.

Next, we establish that with knowledge of the local density regularity, the following histogram based plug-in level set estimator

$$\widehat{G}_{0,j} = \bigcup_{A \in \mathcal{A}_j : \widehat{f}(A) > 0} A \tag{2.10}$$

achieves this minimax lower bound of Proposition 3 for support set estimation using an appropriate choice of the histogram resolution. This requires a modified theoretical analysis using the Craig-Bernstein inequality [54] rather than the relative VC inequalities used in the proofs of Theorems 1, 2 for level set estimation. The proof is sketched in Section 2.6.7.

Theorem 3. *Assume that the local density regularity α is known. Pick the resolution j such that $2^{-j} \asymp s_n(n/\log n)^{-\frac{1}{(d+\alpha)}}$, where s_n is a monotone diverging sequence. Then*

$$\sup_{f \in \mathcal{F}_1^*(\alpha)} \mathbb{E}[d_\infty(\widehat{G}_{0,j}, G_0^*)] \leq C s_n \left(\frac{n}{\log n}\right)^{-\frac{1}{d+\alpha}}$$

for all n, where $C \equiv C(C_1, C_3, \epsilon_o, f_{\max}, \delta_1, d, \alpha) > 0$ is a constant.

Achieving adaptivity for support set estimation requires modification of the vernier procedure. This is because to judge the local density regularity near the level of interest, the adaptivity procedure needs to focus on the cells that are close to the boundary. The vernier can achieve this for level $\gamma > 0$ under assumption **[A]** which implies that the density has no flat parts near the level of interest. However, for support set estimation, the density is flat (zero) outside the support set. Hence, observe that the vernier output is not determined by a cell intersecting the boundary and it fails to focus on the cells that are close to the boundary. One direction to rectify this is to force the vernier to investigate only those subcells which have a certain positive average density. However, if the support set boundary is too close to the cell boundary this can still cause the vernier to yield a small value even when the cell is large. To avoid such alignment artifacts, the vernier can be defined in terms of multiple shifted regular partitions, but we do not pursue this further.

2.4.4 Addressing jumps in the density at the level of interest

The case $\alpha = 0$ implies that the density jumps across the level of interest at all points around the level set boundary. In the non-adaptive setting, the histogram-based plug-in level set estimator of (2.7) achieves the minimax Hausdorff performance. To see this, the theoretical analysis needs to be modified a bit and is discussed in Section 2.6.8. The adaptive estimator can also be extended to handle the complete range $0 \leq \alpha < \infty$ by a slight modification of the vernier. Notice that the current form of the vernier may fail to select an appropriate resolution in the jump case; for example, if the density is piecewise constant on either side of the jump, the vernier output is the same irrespective of the resolution. A

slight modification of the vernier as follows

$$\mathcal{V}_{\gamma,j} = 2^{-j'/2} \min_{A \in \mathcal{A}_j} \max_{A' \in \mathcal{A}_{j'} \cap A} |\gamma - \bar{f}(A')|,$$

makes the vernier sensitive to the resolution even for the jump case, and biases a vernier minimizer towards finer resolutions. A fine resolution is needed for the jump case to approximate the density well (notice that a fine resolution implies less averaging, however the resulting instability in the estimate can be tolerated as there is a jump in the density). While it is clear why the modification is needed, the exact form of the modifying factor $2^{-j'/2}$ arises from technical considerations and is somewhat non-intuitive. Hence, we omitted the jump case in our earlier analysis to keep the presentation simple. Since the penalty is designed to control the deviation of empirical and true vernier, it also needs to be scaled accordingly:

$$\Psi_{j'} := 2^{-j'/2} \max_{A \in \mathcal{A}_{j'}} \sqrt{8 \frac{\log(2^{j'(d+1)} 16/\delta)}{n\mu(A)} \max\left(\widehat{f}(A), 8 \frac{\log(2^{j'(d+1)} 16/\delta)}{n\mu(A)}\right)}$$

This ensures that balancing the vernier and penalty leads to the appropriate resolution for the whole range of the regularity parameter, $0 \leq \alpha < \infty$. A proof sketch is given in Section 2.6.8.

2.5 Concluding Remarks

In this chapter, we developed a Hausdorff accurate level set estimation method that is adaptive to unknown density regularity and achieves nearly minimax optimal rates of error convergence over a more general class of level sets than considered in previous literature. The vernier provides the key to achieve adaptivity while requiring only local regularity of the density in the vicinity of the desired level. The complexity regularization approach based on the vernier is similar in spirit to so-called Lepski methods (for example, [47]) for function estimation which are spatially adaptive bandwidth selectors, but the vernier focuses on cells close to the desired level and thus is optimally designed for the level set problem. However, Lepski methods involve sequential testing, whereas our procedure needs the vernier to be evaluated at all resolutions to determine the appropriate resolution. It is of interest

to develop a sequential procedure based on the vernier that will only require local density regularity, but will be faster to implement.

We also discussed extensions of the proposed estimator to address simultaneous multiple level set estimation, support set estimation and discontinuity in the density around the level of interest. We provided some pointers to address adaptivity for support set estimation, however we have not solved this completely yet. While we consider level sets with locally Lipschitz boundaries, extensions to additional boundary smoothness (for example, Hölder regularity > 1) may be possible in the proposed framework using techniques such as wedgelets [55] or curvelets [56]. The earlier work on Hausdorff accurate level set estimation [34, 35, 38] does address higher smoothness of the boundary, but that follows as a straightforward consequence of assuming a functional form for the boundary. Also, we only addressed the density level set problem, extensions to general regression level set estimation should be possible using a similar approach.

Finally, we discuss and motivate estimators based on spatially adapted partitions that can offer improved performance in practice under spatial variations in the density regularity. It is well known that spatially adaptive partitions such as recursive dyadic partitions (RDPs) [48–51] may provide significant improvements over non-adaptive partitions like histograms for many set learning problems involving a weighted symmetric difference error measure, including classification [51], minimum volume set estimation [32] and level set estimation [22]. In fact, for many function classes, estimators based on adaptive, non-uniform partitions can achieve minimax optimal rates that cannot be achieved by estimators based on non-adaptive partitions. However, the results of this chapter establish that this is not the case for the Hausdorff metric. This is a consequence of the fact that symmetric difference based errors are global, whereas the Hausdorff error is sensitive to local errors and depends on the worst case error at any point. Having non-uniform cells adapted to the regularity along the boundary can lead to faster convergence rates under global measures, whereas the Hausdorff error being dominated by the worst case error is not expected to benefit from adaptivity of the partition. While spatially adaptive, non-uniform partitions do not provide an improvement in

convergence rates under the Hausdorff error metric, if the density regularity varies smoothly along the level set boundary or if the connected components of a level set have different density regularities, non-uniform partitions are capable of adapting to the local smoothness around each component and this may generate better estimates in practice. This might be possible by developing a tree-based approach based on the vernier or a modified Lepski method, and is the subject of current research.

2.6 Proofs

2.6.1 Proof of Lemma 1

We proceed by recalling the definition of $\mathcal{F}_{SL}(\alpha)$ as defined in [35]. The class corresponds to densities bounded above by $f_{\max}$, satisfying a slightly modified form of the local density regularity assumption [**A**]:

[**A"**] *Local density regularity:* The density is α-regular around the γ-level set, $0 < \alpha < \infty$ and $\gamma < f_{\max}$, if there exist constants $C_2 > C_1 > 0$ and $\delta_1 > 0$ such that

$$C_1 \rho(x, \partial G_\gamma^*)^\alpha \leq |f(x) - \gamma| \leq C_2 \rho(x, \partial G_\gamma^*)^\alpha$$

for all $x \in \mathcal{X}$ with $|f(x) - \gamma| \leq \delta_1$, where ∂G_γ^* is the boundary of the true level set G_γ^*, and the set $\{x : |f(x) - \gamma| \leq \delta_1\}$ is non-empty.

and the densities have γ level sets of the form

$$G_\gamma^* = \{(r, \boldsymbol{\phi}); \boldsymbol{\phi} \in [0, \pi)^{d-2} \times [0, 2\pi), 0 \leq r \leq g(\boldsymbol{\phi}) \leq R\},$$

where $(r, \boldsymbol{\phi})$ denote the polar/hyperspherical coordinates and $R > 0$ is a constant. g is a periodic Lipschitz function that satisfies $g(\boldsymbol{\phi}) \geq h$, where $h > 0$ is a constant, and

$$|g(\boldsymbol{\phi}) - g(\boldsymbol{\theta})| \leq L||\boldsymbol{\phi} - \boldsymbol{\theta}||_1, \quad \forall\, \boldsymbol{\phi}, \boldsymbol{\theta} \in [0, \pi)^{d-2} \times [0, 2\pi).$$

Here $L > 0$ is the Lipschitz constant, and $||\cdot||_1$ denotes the ℓ_1 norm.

We set $R = 1/2$ in the definition of the star-shaped set so that the domain is a subset of $[-1/2, 1/2]^d$. With this domain, we now show that the level set G^*_γ of a density $f \in \mathcal{F}_{SL}(\alpha)$ satisfies **[B]**. The same result holds for star-shaped sets defined on the shifted domain $[0,1]^d$.

We first present a sketch of the main ideas, and then provide a detailed proof. Consider the γ-level set G^*_γ of a density $f \in \mathcal{F}_{SL}(\alpha)$. To see that it satisfies **[B]**, divide the star-shaped set G^*_γ into sectors of width $\asymp \epsilon$ so that each sector contains at least one ϵ-ball and the inner cover $\mathcal{I}_\epsilon(G^*_\gamma)$ touches the boundary at some point(s) in each sector. Now one can argue that, in each sector, all other points on the boundary are $O(\epsilon)$ from the inner cover since the boundary is Lipschitz. Since this is true for each sector, we have $\forall x \in \partial G^*_\gamma$, $\rho(x, \mathcal{I}_\epsilon(G^*_\gamma)) = O(\epsilon)$. Hence, the result follows. We now present the proof in detail.

To see that G^*_γ satisfies **[B]**, fix $\epsilon_o \leq h/3$. Then for all $\epsilon \leq \epsilon_o$, $B(0,\epsilon) \subseteq G^*_\gamma$ (since $g(\boldsymbol{\phi}) \geq h > \epsilon_o$), and hence $\mathcal{I}_\epsilon(G^*_\gamma) \neq \emptyset$. We also need to show that $\exists C_3 > 0$ such that for all $x \in \partial G^*_\gamma$, $\rho(x, \mathcal{I}_\epsilon(G)) \leq C_3\epsilon$. For this, divide G^*_γ into M^{d-1} sectors indexed by $\boldsymbol{m} = (m_1, m_2, \ldots, m_{d-1}) \in \{1, \ldots, M\}^{d-1}$

$$S_{\boldsymbol{m}} = \Big\{ (r, \boldsymbol{\phi}) : 0 \leq r \leq g(\boldsymbol{\phi}), \frac{2\pi(m_{d-1}-1)}{M} \leq \phi_{d-1} < \frac{2\pi m_{d-1}}{M} \\ \frac{\pi(m_i - 1)}{M} \leq \phi_i < \frac{\pi m_i}{M} \quad i = 1, \ldots, d-2 \Big\},$$

where $\boldsymbol{\phi} = (\phi_1, \phi_2, \ldots, \phi_{d-1})$. Let

$$M = \left\lfloor \frac{\pi}{2 \sin^{-1} \frac{\epsilon}{h - \epsilon_o}} \right\rfloor$$

This choice of M implies that:

(i) There exists an ϵ-ball within $S_{\boldsymbol{m}} \cap B(0,h)$ for every $\boldsymbol{m} \in \{1, \ldots, M\}^{d-1}$, and hence within each sector $S_{\boldsymbol{m}}$. This follows because the minimum angular width of a sector with radius h required to fit an ϵ-ball within is

$$2 \sin^{-1} \frac{\epsilon}{h - \epsilon} \leq 2 \sin^{-1} \frac{\epsilon}{h - \epsilon_o} \leq \frac{\pi}{M}.$$

(ii) The angular-width of the sectors scales as $O(\epsilon)$.

$$\begin{aligned}\frac{\pi}{M} < \frac{\pi}{\frac{\pi}{2\sin^{-1}\frac{\epsilon}{h-\epsilon_o}} - 1} = \frac{1}{\frac{1}{2\sin^{-1}\frac{\epsilon}{h-\epsilon_o}} - \frac{1}{\pi}} &\leq 3\sin^{-1}\frac{\epsilon}{h-\epsilon_o} \\ &\leq 6\frac{\epsilon}{h-\epsilon_o} \leq \frac{9}{h}\epsilon\end{aligned}$$

The second inequality follows since

$$\frac{1}{\pi} \leq \frac{1}{6\sin^{-1}\frac{\epsilon}{h-\epsilon_o}}$$

since $\frac{\epsilon}{h-\epsilon_o} \leq \frac{\epsilon_o}{h-\epsilon_o} \leq \frac{1}{2}$ by choice of $\epsilon_o \leq h/3$. The third inequality is true since $\sin^{-1}(z/2) \leq z$ for $0 \leq z \leq \pi/2$. The last step follows by choice of $\epsilon_o \leq h/3$.

Now from (i) above, each sector contains at least one ϵ-ball. Consider any $\boldsymbol{m} \in \{1, \ldots, M\}^{d-1}$. We claim that there exists a point $x_{\boldsymbol{m}} \in \partial G_\gamma^* \cap S_{\boldsymbol{m}}$, $x_{\boldsymbol{m}} = (g(\boldsymbol{\theta}), \boldsymbol{\theta})$ for some $\boldsymbol{\theta} \in [0,\pi)^{d-2} \times [0, 2\pi)$, such that $\rho(x_{\boldsymbol{m}}, \mathcal{I}_\epsilon(G_\gamma^*)) = 0$. Suppose not. Then one can slide the ϵ-ball within the sector towards the periphery and never touch the boundary, implying that the set G_γ^* is unbounded. This is a contradiction by the definition of the class $\mathcal{F}_{SL}(\alpha)$. So now we have, $\forall y \in \partial G_\gamma^* \cap S_{\boldsymbol{m}}$, $y = (g(\boldsymbol{\phi}), \boldsymbol{\phi})$

$$\rho(y, \mathcal{I}_\epsilon(G_\gamma^*)) \leq \rho(y, x_{\boldsymbol{m}}) = ||y - x_{\boldsymbol{m}}||$$

Now recall that if $y = (y_1, \ldots, y_d) \equiv (r, \phi_1, \ldots, \phi_{d-1}) = (g(\boldsymbol{\phi}), \boldsymbol{\phi})$, then the relation between the Cartesian and hypershperical coordinates is given as:

$$\begin{aligned}
y_1 &= r\cos\phi_1 \\
y_2 &= r\sin\phi_1\cos\phi_2 \\
y_3 &= r\sin\phi_1\sin\phi_2\cos\phi_3 \\
&\vdots \\
y_{d-1} &= r\sin\phi_1 \ldots \sin\phi_{d-2}\cos\phi_{d-1} \\
y_d &= r\sin\phi_1 \ldots \sin\phi_{d-2}\sin\phi_{d-1}
\end{aligned}$$

Now since $||y-x|| = \sum_{i=1}^{d}(y_i - x_i)^2$, using the above transformation and simple algebra, we can show that:

$$\begin{aligned}
||y - x_{\boldsymbol{m}}||^2 &= ||(g(\boldsymbol{\phi}), \boldsymbol{\phi}) - (g(\boldsymbol{\theta}), \boldsymbol{\theta})||^2 \\
&= (g(\boldsymbol{\phi}) - g(\boldsymbol{\theta}))^2 + 4g(\boldsymbol{\phi})g(\boldsymbol{\theta}) \sum_{i=1}^{d-1} \sin\phi_1 \ldots \sin\phi_{i-1} \sin\theta_1 \ldots \sin\theta_{i-1} \sin^2 \frac{\phi_i - \theta_i}{2} \\
&\leq (g(\boldsymbol{\phi}) - g(\boldsymbol{\theta}))^2 + 4g(\boldsymbol{\phi})g(\boldsymbol{\theta}) \sum_{i=1}^{d-1} \sin^2 \frac{\phi_i - \theta_i}{2}
\end{aligned}$$

Using this, we have $\forall y \in \partial G_\gamma^* \cap S_{\boldsymbol{m}}$

$$\begin{aligned}
\rho(y, \mathcal{I}_\epsilon(G_\gamma^*)) &\leq \sqrt{(g(\boldsymbol{\phi}) - g(\boldsymbol{\theta}))^2 + 4g(\boldsymbol{\phi})g(\boldsymbol{\theta}) \sum_{i=1}^{d-1} \sin^2 \frac{\phi_i - \theta_i}{2}} \\
&\leq |g(\boldsymbol{\phi}) - g(\boldsymbol{\theta})| + 2\sqrt{g(\boldsymbol{\phi})g(\boldsymbol{\theta})} \sum_{i=1}^{d-1} \left| \sin \frac{\phi_i - \theta_i}{2} \right| \\
&\leq L||\boldsymbol{\phi} - \boldsymbol{\theta}||_1 + \sum_{i=1}^{d-1} \frac{|\phi_i - \theta_i|}{2} \\
&= (L + 1/2) \sum_{i=1}^{d-1} |\phi_i - \theta_i| \\
&\leq (L + 1/2) d \frac{\pi}{M} \\
&\leq \frac{9d(L + 1/2)}{h} \epsilon := C_3 \epsilon
\end{aligned}$$

where the third step follows by using the Lipschitz condition on $g(\cdot)$, $g(\cdot) \leq R = 1/2$ and since $|\sin(z)| \leq |z|$. The fifth step follows since $x, y \in S_{\boldsymbol{m}}$ and hence $|\phi_i - \theta_i| \leq \pi/M$ for $i = 1, \ldots, d-2$ and $|\phi_{d-1} - \theta_{d-1}| \leq 2\pi/M$. The sixth step invokes (ii) above.

Therefore, we have for all $y \in \partial G_\gamma^* \cap S_{\boldsymbol{m}}$ $\rho(y, \mathcal{I}_\epsilon(G_\gamma^*)) \leq C_3\epsilon$. And since the result is true for any sector, condition **[B]** is satisfied by any level set G_γ^* with density $f \in \mathcal{F}_{SL}(\alpha)$.

■

2.6.2 Proof of Proposition 1

Notice that since $\mathcal{F}_2^*(\alpha) \subset \mathcal{F}_1^*(\alpha)$, we have

$$\inf_{\widehat{G}_n} \sup_{f \in \mathcal{F}_1^*(\alpha)} \mathbb{E}[d_\infty(\widehat{G}_n, G_\gamma^*)] \geq \inf_{\widehat{G}_n} \sup_{f \in \mathcal{F}_2^*(\alpha)} \mathbb{E}[d_\infty(\widehat{G}_n, G_\gamma^*)]$$

Therefore, it suffices to establish a lower bound for the class of densities given by $\mathcal{F}_2^*(\alpha)$.

We consider the class of densities $\mathcal{F}_{SL}(\alpha)$ with star-shaped levels sets having Lipschitz boundaries, as defined in [35]. Lemma 1 establishes that all densities in $\mathcal{F}_{SL}(\alpha)$ satisfy assumption **[B]**. Further, since the discrete set of densities $\mathcal{F}_{SL}^D(\alpha) \subset \mathcal{F}_{SL}(\alpha)$ used to derive the lower bound using Fano's lemma in [35], satisfy the local density regularity as stated in assumption **[A]** [2], we have

$$\inf_{\widehat{G}_n} \sup_{f \in \mathcal{F}_2^*(\alpha)} \mathbb{E}[d_\infty(\widehat{G}_n, G_\gamma^*)] \geq \inf_{\widehat{G}_n} \sup_{f \in \mathcal{F}_{SL}^D(\alpha)} \mathbb{E}[d_\infty(\widehat{G}_n, G_\gamma^*)] \geq c \left(\frac{n}{\log n}\right)^{-\frac{1}{d+2\alpha}},$$

for n large enough. The last step follows from proof of Theorem 4 in [35].

■

2.6.3 Proof of Proposition 2

Observe that assumption **[B]** implies that G_γ^* is not empty since $G_\gamma^* \supseteq \mathcal{I}_\epsilon(G_\gamma^*) \neq \emptyset$ for $\epsilon \leq \epsilon_o$. Hence for large enough n, with high probability, the plug-in level set estimate $\widehat{G}$ is also non-empty since the sup norm error between $\widehat{f}(x)$ and $f(x)$ converges in probability to zero. Now recall that for non-empty sets

$$d_\infty(\widehat{G}, G_\gamma^*) = \max\{\sup_{x \in G_\gamma^*} \rho(x, \widehat{G}), \sup_{x \in \widehat{G}} \rho(x, G_\gamma^*)\}.$$

We now derive upper bounds on the two terms that control the Hausdorff error.

First, observe that if $\widehat{G} \Delta G_\gamma^* \neq \emptyset$, then for all points $x \in \widehat{G} \Delta G_\gamma^*$ (that is, points that are incorrectly included or excluded from the level set estimate), $|f(x) - \gamma| \leq |f(x) - \widehat{f}(x)|$

[2] All densities in $\mathcal{F}_{SL}(\alpha)$ satisfy a weaker former of assumption **[A]** that only requires density regularity to hold at (at least) one point along the boundary. However, for the discrete set of densities considered in the construction of the lower bound, density regularity holds in an open neighborhood around at least one point of the boundary, and hence these satisfy assumption **[A]**.

and hence regularity condition **[A1]** holds at x since the sup norm error between $\widehat{f}(x)$ and $f(x)$ converges in probability to zero and hence for large enough n, with high probability, $|f(x) - \widehat{f}(x)| \leq \delta_1$. So we have:

$$\sup_{x \in \widehat{G} \Delta G^*_\gamma} \rho(x, \partial G^*_\gamma) \leq \sup_{x \in \widehat{G} \Delta G^*_\gamma} \left(\frac{|f(x) - \gamma|}{C_1} \right)^{1/\alpha} \leq \left(\frac{\mathcal{E}(\widehat{G})}{C_1} \right)^{1/\alpha} =: \epsilon. \tag{2.11}$$

The last inequality follows since $\forall x \in \widehat{G} \Delta G^*_\gamma$, $|f(x) - \gamma| \leq \mathcal{E}(\widehat{G})$. Also, notice that we define ϵ equal to this upper bound. This result implies that all points whose distance to the boundary ∂G^*_γ is greater than ϵ cannot lie in $\widehat{G} \Delta G^*_\gamma$ and hence are correctly included or excluded from the level set estimate. Let $\mathcal{I}_{2\epsilon} \equiv \mathcal{I}_{2\epsilon}(G^*_\gamma)$. This implies that all points within $\mathcal{I}_{2\epsilon}$ that are greater than ϵ away from the boundary lie in $\widehat{G} \cap G^*_\gamma$ since they lie in $\mathcal{I}_{2\epsilon} \subseteq G^*_\gamma$. Hence,

$$\sup_{x \in \mathcal{I}_{2\epsilon}} \rho(x, \widehat{G} \cap G^*_\gamma) \leq \epsilon. \tag{2.12}$$

Using Eqs. (2.11) and (2.12), we now bound the two terms of the Hausdorff error. To bound the second term of the Hausdorff error, consider two cases:

(i) If $\widehat{G} \setminus G^*_\gamma = \emptyset$, then $\widehat{G} \subseteq G^*_\gamma$. Hence

$$\sup_{x \in \widehat{G}} \rho(x, G^*_\gamma) = 0.$$

(ii) If $\widehat{G} \setminus G^*_\gamma \neq \emptyset$, then $\widehat{G} \Delta G^*_\gamma \neq \emptyset$. Hence using (2.11), we get:

$$\begin{aligned} \sup_{x \in \widehat{G}} \rho(x, G^*_\gamma) = \sup_{x \in \widehat{G} \setminus G^*_\gamma} \rho(x, G^*_\gamma) &= \sup_{x \in \widehat{G} \setminus G^*_\gamma} \rho(x, \partial G^*_\gamma) \\ &\leq \sup_{x \in \widehat{G} \Delta G^*_\gamma} \rho(x, \partial G^*_\gamma) \leq \left(\frac{\mathcal{E}(\widehat{G})}{C_1} \right)^{1/\alpha}. \end{aligned}$$

Therefore, for either case

$$\sup_{x \in \widehat{G}} \rho(x, G^*_\gamma) \leq \left(\frac{\mathcal{E}(\widehat{G})}{C_1} \right)^{1/\alpha}. \tag{2.13}$$

To bound the first term of the Hausdorff error, again consider two cases:

(i) If $G^*_\gamma \setminus \widehat{G} = \emptyset$, then $G^*_\gamma \subseteq \widehat{G}$. Hence

$$\sup_{x \in G^*_\gamma} \rho(x, \widehat{G}) = 0.$$

(ii) If $G^*_\gamma \setminus \widehat{G} \neq \emptyset$, then we proceed by recalling assumption **[B]** which states that the boundary points of G^*_γ are not too far from the inner cover and using (2.12) to control the distance of the inner cover from $\widehat{G}$.

$$\begin{aligned}
\sup_{x \in G^*_\gamma} \rho(x, \widehat{G}) &\leq \sup_{x \in G^*_\gamma} \rho(x, \widehat{G} \cap G^*_\gamma) \\
&= \max\{\sup_{x \in \mathcal{I}_{2\epsilon}} \rho(x, \widehat{G} \cap G^*_\gamma), \sup_{x \in G^*_\gamma \setminus \mathcal{I}_{2\epsilon}} \rho(x, \widehat{G} \cap G^*_\gamma)\} \\
&\leq \max\{\epsilon, \sup_{x \in G^*_\gamma \setminus \mathcal{I}_{2\epsilon}} \rho(x, \widehat{G} \cap G^*_\gamma)\}.
\end{aligned}$$

The last step follows from (2.12). Now consider any $x \in G^*_\gamma \setminus \mathcal{I}_{2\epsilon}$. Then using triangle inequality, $\forall y \in \partial G^*_\gamma$ and $\forall z \in \mathcal{I}_{2\epsilon}$,

$$\begin{aligned}
\rho(x, \widehat{G} \cap G^*_\gamma) &\leq \rho(x, y) + \rho(y, z) + \rho(z, \widehat{G} \cap G^*_\gamma) \\
&\leq \rho(x, y) + \rho(y, z) + \sup_{z' \in \mathcal{I}_{2\epsilon}} \rho(z', \widehat{G} \cap G^*_\gamma) \\
&\leq \rho(x, y) + \rho(y, z) + \epsilon.
\end{aligned}$$

The last step follows from (2.12). This implies that $\forall y \in \partial G^*_\gamma$,

$$\begin{aligned}
\rho(x, \widehat{G} \cap G^*_\gamma) &\leq \rho(x, y) + \inf_{z \in \mathcal{I}_{2\epsilon}} \rho(y, z) + \epsilon \\
&= \rho(x, y) + \rho(y, \mathcal{I}_{2\epsilon}) + \epsilon \\
&\leq \rho(x, y) + \sup_{y' \in \partial G^*_\gamma} \rho(y', \mathcal{I}_{2\epsilon}) + \epsilon \\
&\leq \rho(x, y) + 2C_3\epsilon + \epsilon.
\end{aligned}$$

Here the last step invokes assumption **[B]**. This in turn implies that

$$\rho(x, \widehat{G} \cap G^*_\gamma) \leq \inf_{y \in \partial G^*_\gamma} \rho(x, y) + (2C_3 + 1)\epsilon \leq 2\epsilon + (2C_3 + 1)\epsilon$$

The second step is true for $x \in G^*_\gamma \setminus \mathcal{I}_{2\epsilon}$, because if it was not true then $\forall y \in \partial G^*_\gamma$, $\rho(x, y) > 2\epsilon$ and hence there exists a closed 2ϵ-ball around x that is in G^*_γ. This contradicts the fact that $x \notin \mathcal{I}_{2\epsilon}$.

Therefore, we have:

$$\sup_{x \in G_\gamma^* \setminus \mathcal{I}_{2\epsilon}} \rho(x, \widehat{G} \cap G_\gamma^*) \leq (2C_3 + 3)\epsilon.$$

And going back to the start of case (ii) we get:

$$\sup_{x \in G_\gamma^*} \rho(x, \widehat{G}) \leq (2C_3 + 3)\epsilon.$$

So for either case

$$\sup_{x \in G_\gamma^*} \rho(x, \widehat{G}) \leq (2C_3 + 3)\epsilon = (2C_3 + 3) \left(\frac{\mathcal{E}(\widehat{G})}{C_1} \right)^{1/\alpha}. \tag{2.14}$$

Putting together the bounds from Eqs. (2.13), (2.14) for the two terms of the Hausdorff error, we get: For large enough n, with high probability

$$d_\infty(\widehat{G}, G_\gamma^*) = \max\{ \sup_{x \in G_\gamma^*} \rho(x, \widehat{G}), \sup_{x \in \widehat{G}} \rho(x, G_\gamma^*) \} \leq (2C_3 + 3) \left(\frac{\mathcal{E}(\widehat{G})}{C_1} \right)^{1/\alpha}.$$

This concludes the proof.

■

2.6.4 Proof of Theorem 1

Before proceeding to the proof of Theorem 1, we establish three lemmas that will be used in this proof, as well as the proof of Theorem 2. The first lemma bounds the deviation of true and empirical density averages. The choice of penalty used to achieve adaptivity is motivated by this relation.

Lemma 2. *Consider $0 < \delta < 1$. With probability at least $1 - \delta$, the following is true for all $j \geq 0$:*

$$\max_{A \in \mathcal{A}_j} |\bar{f}(A) - \widehat{f}(A)| \leq \Psi_j.$$

Proof. The proof relies on a pair of VC inequalities (See [8] Chapter 3) that bound the *relative* deviation of true and empirical probabilities. For the collection $\mathcal{A}_j$ with cardinality

bounded by 2^{jd}, the relative VC inequalities state that for any $\epsilon > 0$

$$P\left(\sup_{A\in\mathcal{A}_j}\frac{P(A)-\widehat{P}(A)}{\sqrt{P(A)}}>\epsilon\right)\leq 4\cdot 2^{jd}e^{-n\epsilon^2/4}$$

and

$$P\left(\sup_{A\in\mathcal{A}_j}\frac{\widehat{P}(A)-P(A)}{\sqrt{\widehat{P}(A)}}>\epsilon\right)\leq 4\cdot 2^{jd}e^{-n\epsilon^2/4}.$$

Also observe that

$$\widehat{P}(A)\leq P(A)+\epsilon\sqrt{\widehat{P}(A)}\Longrightarrow \widehat{P}(A)\leq 2\max(P(A),2\epsilon^2) \tag{2.15}$$

and

$$P(A)\leq \widehat{P}(A)+\epsilon\sqrt{P(A)}\Longrightarrow P(A)\leq 2\max(\widehat{P}(A),2\epsilon^2). \tag{2.16}$$

To see the first statement, consider two cases:

1) $\widehat{P}(A)\leq 4\epsilon^2$. The statement is obvious.

2) $\widehat{P}(A)> 4\epsilon^2$. This gives a bound on ϵ, which implies

$$\widehat{P}(A)\leq P(A)+\widehat{P}(A)/2\Longrightarrow \widehat{P}(A)\leq 2P(A).$$

The second statement follows similarly.

Using the second statement and the relative VC inequalities for the collection $\mathcal{A}_j$, we have: With probability $> 1-8\cdot 2^{jd}e^{-n\epsilon^2/4}$, $\forall A\in\mathcal{A}_j$ both

$$P(A)-\widehat{P}(A)\leq\epsilon\sqrt{P(A)}\leq\epsilon\sqrt{2\max(\widehat{P}(A),2\epsilon^2)}$$

and

$$\widehat{P}(A)-P(A)\leq\epsilon\sqrt{\widehat{P}(A)}\leq\epsilon\sqrt{2\max(\widehat{P}(A),2\epsilon^2)}.$$

In other words, with probability $> 1-8\cdot 2^{jd}e^{-n\epsilon^2/4}$, $\forall A\in\mathcal{A}_j$

$$|P(A)-\widehat{P}(A)|\leq\epsilon\sqrt{2\max(\widehat{P}(A),2\epsilon^2)}.$$

Setting $\epsilon = \sqrt{4\log(2^{jd}8/\delta_j)/n}$, we have with probability $> 1-\delta_j$, $\forall A \in \mathcal{A}_j$

$$|P(A) - \widehat{P}(A)| \leq \sqrt{8\frac{\log(2^{jd}8/\delta_j)}{n}\max\left(\widehat{P}(A), 8\frac{\log(2^{jd}8/\delta_j)}{n}\right)}$$

Setting $\delta_j = \delta 2^{-(j+1)}$ and applying union bound, we have with probability $> 1-\delta$, for all resolutions $j \geq 0$ and all cells $A \in \mathcal{A}_j$

$$|P(A) - \widehat{P}(A)| \leq \sqrt{8\frac{\log(2^{j(d+1)}16/\delta)}{n}\max\left(\widehat{P}(A), 8\frac{\log(2^{j(d+1)}16/\delta)}{n}\right)}$$

The result follows by dividing both sides by $\mu(A)$. □

The next lemma states how the density deviation bound or penalty Ψ_j scales with resolution j and number of observations n. It essentially reflects the fact that at finer resolutions, the amount of data per cell decreases leading to larger estimation error.

Lemma 3. *There exist constants $c_3, c_4 \equiv c_4(f_{\max}, d) > 0$ such that if $j \equiv j(n)$ satisfies $2^j = O((n/\log n)^{1/d})$, then for all n, with probability at least $1 - 1/n$,*

$$c_3\sqrt{2^{jd}\frac{\log n}{n}} \leq \Psi_j \leq c_4\sqrt{2^{jd}\frac{\log n}{n}}.$$

Proof. Recall the definition of Ψ_j

$$\Psi_j := \max_{A\in\mathcal{A}_j}\sqrt{8\frac{\log(2^{j(d+1)}16/\delta)}{n\mu(A)}\max\left(\widehat{f}(A), 8\frac{\log(2^{j(d+1)}16/\delta)}{n\mu(A)}\right)}$$

We first derive the lower bound. Observe that since the total empirical probability mass is 1, we have

$$1 = \sum_{A\in\mathcal{A}_j}\widehat{P}(A) \leq \max_{A\in\mathcal{A}_j}\widehat{P}(A) \times |\mathcal{A}_j| = \max_{A\in\mathcal{A}_j}\frac{\widehat{P}(A)}{\mu(A)} = \max_{A\in\mathcal{A}_j}\widehat{f}(A).$$

Use this along with $\delta = 1/n$, $j \geq 0$ and $\mu(A) = 2^{-jd}$ to get:

$$\Psi_j \geq \sqrt{2^{jd}8\frac{\log 16n}{n}}.$$

To get an upper bound, using (2.15) from the proof of Lemma 2, we have with probability $> 1 - 8 \cdot 2^{jd} e^{-n\epsilon^2/4}$, for all $A \in \mathcal{A}_j$

$$\widehat{P}(A) \leq 2 \max(P(A), 2\epsilon^2).$$

Setting $\epsilon = \sqrt{4 \log(2^{jd} 8/\delta_j)/n}$, we have with probability $> 1 - \delta_j$, for all $A \in \mathcal{A}_j$

$$\widehat{P}(A) \leq 2 \max \left(P(A), 8 \frac{\log(2^{jd} 8/\delta_j)}{n} \right)$$

Dividing by $\mu(A) = 2^{-jd}$, using the density bound $f_{\max}$, we have with probability $> 1 - \delta_j$, for all $A \in \mathcal{A}_j$

$$\widehat{f}(A) \leq 2 \max \left(f_{\max}, 2^{jd} 8 \frac{\log(2^{jd} 8/\delta_j)}{n} \right).$$

Setting $\delta_j = \delta 2^{-(j+1)}$ and applying union bound, we have with probability $> 1 - \delta$, for all resolutions $j \geq 0$

$$\max_{A \in \mathcal{A}_j} \widehat{f}(A) \leq 2 \max \left(f_{\max}, 2^{jd} 8 \frac{\log(2^{j(d+1)} 16/\delta)}{n} \right).$$

This implies

$$\Psi_j \leq \sqrt{2^{jd} 8 \frac{\log(2^{j(d+1)} 16/\delta)}{n} \cdot 2 \max \left(f_{\max}, 2^{jd} 8 \frac{\log(2^{j(d+1)} 16/\delta)}{n} \right)}.$$

Using $\delta = 1/n$ and $2^j = O((n/\log n)^{1/d})$, we get:

$$\Psi_j \leq c_4(f_{\max}, d) \sqrt{2^{jd} \frac{\log n}{n}}.$$

□

We now analyze the performance of the plug-in histogram-based level set estimator proposed in (2.7), and establish the following lemma that bounds its Hausdorff error. The first term denotes the estimation error while the second term that is proportional to the side-length of a cell (2^{-j}) reflects the approximation error. We would like to point out that some arguments in the proofs hold for s_n large enough. This implies that some of the constants in our proofs will depend on $\{s_i\}_{i=1}^{\infty}$, the exact form that the sequence s_n takes (but not on n). However, we omit this dependence for simplicity.

Lemma 4. *Consider densities satisfying assumptions **[A1]** and **[B]**. If $j \equiv j(n)$ is such that $2^j = O(s_n^{-1}(n/\log n)^{1/d})$, where s_n is a monotone diverging sequence, and $n \geq n_0(f_{\max}, d, \delta_1, \epsilon_o, C_1, \alpha)$, then with probability at least $1 - 3/n$*

$$d_\infty(\widehat{G}_j, G^*_\gamma) \leq \max(2C_3 + 3, 8\sqrt{d}\epsilon_o^{-1}) \left[\left(\frac{\Psi_j}{C_1} \right)^{1/\alpha} + \sqrt{d}2^{-j} \right].$$

Proof. The proof follows along the lines of the proof of Proposition 2. Let $J_0 = \lceil \log_2 4\sqrt{d}/\epsilon_o \rceil$, where ϵ_o is as defined in assumption **[B]**. Also define

$$\epsilon_j := \left[\left(\frac{\Psi_j}{C_1} \right)^{1/\alpha} + \sqrt{d}2^{-j} \right].$$

Consider two cases:

I. $j < J_0$.

For this case, since the domain $\mathcal{X} = [0,1]^d$, we use the trivial bound

$$d_\infty(\widehat{G}_j, G^*_\gamma) \leq \sqrt{d} \leq 2^{J_0}(\sqrt{d}2^{-j}) \leq 8\sqrt{d}\epsilon_o^{-1}\epsilon_j.$$

The last step follows by choice of J_0 and since $\Psi_j, C_1 > 0$.

II. $j \geq J_0$.

Observe that assumption **[B]** implies that G^*_γ is not empty since $G^*_\gamma \supseteq \mathcal{I}_\epsilon(G^*_\gamma) \neq \emptyset$ for $\epsilon \leq \epsilon_o$. We will show that for large enough n, with high probability, $\widehat{G}_j \cap G^*_\gamma \neq \emptyset$ for $j \geq J_0$ and hence $\widehat{G}_j$ is not empty. Thus the Hausdorff error is given as

$$d_\infty(\widehat{G}_j, G^*_\gamma) = \max\{ \sup_{x \in G^*_\gamma} \rho(x, \widehat{G}_j), \sup_{x \in \widehat{G}_j} \rho(x, G^*_\gamma) \}, \tag{2.17}$$

and we need bounds on the two terms in the right hand side.

To prove that $\widehat{G}_j$ is not empty and obtain bounds on the two terms in the Hausdorff error, we establish a proposition and corollary. In the following analysis, if $G = \emptyset$, then we define $\sup_{x \in G} g(x) = 0$ for any function $g(\cdot)$. The proposition establishes that for large enough n, with high probability, all points whose distance to the boundary ∂G^*_γ is greater than ϵ_j are correctly excluded or included in the level set estimate.

Proposition 4. *If $j \equiv j(n)$ is such that $2^j = O(s_n^{-1}\ (n/\log n)^{1/d})$, and $n \geq n_1(f_{\max}, d, \delta_1)$, then with probability at least $1 - 2/n$,*

$$\sup_{x \in \widehat{G}_j \Delta G_\gamma^*} \rho(x, \partial G_\gamma^*) \leq \left(\frac{\Psi_j}{C_1}\right)^{1/\alpha} + \sqrt{d} 2^{-j} = \epsilon_j.$$

Proof. If $\widehat{G}_j \Delta G_\gamma^* = \emptyset$, then $\sup_{x \in \widehat{G}_j \Delta G_\gamma^*} \rho(x, \partial G_\gamma^*) = 0$ by definition, and the result of Proposition 4 holds. If $\widehat{G}_j \Delta G_\gamma^* \neq \emptyset$, consider $x \in \widehat{G}_j \Delta G_\gamma^*$. Let $A_x \in \mathcal{A}_j$ denote the cell containing x at resolution j. Consider two cases:

(i) $A_x \cap \partial G_\gamma^* \neq \emptyset$. This implies that

$$\rho(x, \partial G_\gamma^*) \leq \sqrt{d} 2^{-j}.$$

(ii) $A_x \cap \partial G_\gamma^* = \emptyset$. Since $x \in \widehat{G}_j \Delta G_\gamma^*$, it is erroneously included or excluded from the level set estimate $\widehat{G}_j$. Therefore, if $\bar{f}(A_x) \geq \gamma$, then $\widehat{f}(A_x) < \gamma$ otherwise if $\bar{f}(A_x) < \gamma$, then $\widehat{f}(A_x) \geq \gamma$. This implies that $|\gamma - \bar{f}(A_x)| \leq |\bar{f}(A_x) - \widehat{f}(A_x)|$. Using Lemma 2, we get $|\gamma - \bar{f}(A_x)| \leq \Psi_j$ with probability at least $1 - \delta$.

Now let x_1 be any point in A_x such that $|\gamma - f(x_1)| \leq |\gamma - \bar{f}(A_x)|$ (Notice that at least one such point must exist in A_x since this cell does not intersect the boundary). As argued above, $|\gamma - \bar{f}(A_x)| \leq \Psi_j$ with probability at least $1 - 1/n$ (for $\delta = 1/n$). Using Lemma 3, for resolutions satisfying $2^j = O(s_n^{-1}(n/\log n)^{1/d})$, and for large enough $n \geq n_1(f_{\max}, d, \delta_1)$, $\Psi_j \leq \delta_1$ and hence $|\gamma - f(x_1)| \leq \delta_1$, with probability at least $1 - 1/n$. Thus, the density regularity assumption **[A1]** holds at x_1 with probability $> 1 - 2/n$ and we have

$$\rho(x_1, \partial G_\gamma^*) \leq \left(\frac{|\gamma - f(x_1)|}{C_1}\right)^{1/\alpha} \leq \left(\frac{|\gamma - \bar{f}(A_x)|}{C_1}\right)^{1/\alpha} \leq \left(\frac{\Psi_j}{C_1}\right)^{1/\alpha}.$$

Since $x, x_1 \in A_x$,

$$\rho(x, \partial G_\gamma^*) \leq \rho(x_1, \partial G_\gamma^*) + \sqrt{d} 2^{-j} \leq \left(\frac{\Psi_j}{C_1}\right)^{1/\alpha} + \sqrt{d} 2^{-j}.$$

So for both cases, if $j \equiv j(n)$ is such that $2^j = O(s_n^{-1}\ (n/\log n)^{1/d})$, and $n \geq n_1(f_{\max}, d, \delta_1)$, then with probability at least $1 - 2/n$, $\forall x \in \widehat{G}_j \Delta G_\gamma^*$

$$\rho(x, \partial G_\gamma^*) \leq \left(\frac{\Psi_j}{C_1}\right)^{1/\alpha} + \sqrt{d}2^{-j} = \epsilon_j.$$

□

Based on Proposition 4, the following corollary argues that for large enough n and $j \geq J_0 = \lceil \log_2 4\sqrt{d}/\epsilon_o \rceil$, with high probability, all points within the inner cover $\mathcal{I}_{2\epsilon_j}(G_\gamma^*)$ that are at a distance greater than ϵ_j are correctly included in the level set estimate, and hence lie in $\widehat{G}_j \cap G_\gamma^*$. This also implies that $\widehat{G}_j$ is not empty.

Corollary 2. *Recall assumption **[B]** and denote the inner cover of G_γ^* with $2\epsilon_j$-balls, $\mathcal{I}_{2\epsilon_j}(G_\gamma^*) \equiv \mathcal{I}_{2\epsilon_j}$ for simplicity. For any $n \geq n_0 \equiv n_0(f_{\max}, d, \delta_1, \epsilon_o, C_1, \alpha)$, if $j \equiv j(n)$ is such that $2^j = O(s_n^{-1}(n/\log n)^{1/d})$ and $j \geq J_0$, then, with probability at least $1 - 3/n$,*

$$\widehat{G}_j \neq \emptyset \quad \textit{and} \quad \sup_{x \in \mathcal{I}_{2\epsilon_j}} \rho(x, \widehat{G}_j \cap G_\gamma^*) \leq \epsilon_j.$$

Proof. Observe that for $j \geq J_0$, $2\sqrt{d}2^{-j} \leq 2\sqrt{d}2^{-J_0} \leq \epsilon_o/2$. By Lemma 3, for resolutions satisfying $2^j = O(s_n^{-1}(n/\log n)^{1/d})$, and for large enough $n \geq n_2(\epsilon_o, f_{\max}, C_1, \alpha)$, $2(\Psi_j/C_1)^{1/\alpha} \leq \epsilon_o/2$, with probability at least $1 - 1/n$. Therefore for resolutions satisfying $2^j = O(s_n^{-1}(n/\log n)^{1/d})$ and $j \geq J_0$, and for $n \geq n_2$, with probability at least $1 - 1/n$, $2\epsilon_j \leq \epsilon_o$ and hence $\mathcal{I}_{2\epsilon_j} \neq \emptyset$.

Now consider any $2\epsilon_j$-ball in $\mathcal{I}_{2\epsilon_j}$. Then the distance of all points in the interior of the concentric ϵ_j-ball from the boundary of $\mathcal{I}_{2\epsilon_j}$, and hence from the boundary of G_γ^* is greater than ϵ_j. As per Proposition 4 for $n \geq n_0 = \max(n_1, n_2)$, with probability $> 1 - 3/n$, none of these points can lie in $\widehat{G}_j \Delta G_\gamma^*$, and hence must lie in $\widehat{G}_j \cap G_\gamma^*$ since they are in $\mathcal{I}_{2\epsilon_j} \subseteq G_\gamma^*$. Thus, $\widehat{G}_j \neq \emptyset$ and for all $x \in \mathcal{I}_{2\epsilon_j}$,

$$\rho(x, \widehat{G}_j \cap G_\gamma^*) \leq \epsilon_j.$$

□

We now resume the proof of Lemma 4. Assume the conclusions of Proposition 4 and Corollary 2 hold. Thus all the following statements hold for resolutions satisfying $2^j = O(s_n^{-1}(n/\log n)^{1/d})$, $j \geq J_0$ and $n \geq n_0 \equiv n_0(f_{\max}, d, \delta_1, \epsilon_o, C_1, \alpha)$, with probability at least $1 - 3/n$. Since G^*_γ and $\widehat{G}_j$ are non-empty sets, we now bound the two terms that contribute to the Hausdorff error:

$$\sup_{x \in G^*_\gamma} \rho(x, \widehat{G}_j) \qquad \text{and} \qquad \sup_{x \in \widehat{G}_j} \rho(x, G^*_\gamma)$$

To bound the second term, observe that

$$\sup_{x \in \widehat{G}_j} \rho(x, G^*_\gamma) = \sup_{x \in \widehat{G}_j \setminus G^*_\gamma} \rho(x, G^*_\gamma) = \sup_{x \in \widehat{G}_j \setminus G^*_\gamma} \rho(x, \partial G^*_\gamma) \leq \sup_{x \in \widehat{G}_j \Delta G^*_\gamma} \rho(x, \partial G^*_\gamma) \leq \epsilon_j,$$

where the last step follows from Proposition 4. Thus,

$$\sup_{x \in \widehat{G}_j} \rho(x, G^*_\gamma) \leq \epsilon_j. \tag{2.18}$$

To bound the first term, we recall assumption **[B]** which states that the boundary points of G^*_γ are $O(\epsilon_j)$ from the inner cover $\mathcal{I}_{2\epsilon_j}(G^*_\gamma)$, and using Corollary 2 to bound the distance of the inner cover from $\widehat{G}_j$.

$$\begin{aligned}
\sup_{x \in G^*_\gamma} \rho(x, \widehat{G}_j) &\leq \sup_{x \in G^*_\gamma} \rho(x, \widehat{G}_j \cap G^*_\gamma) \\
&= \max\{ \sup_{x \in \mathcal{I}_{2\epsilon_j}} \rho(x, \widehat{G}_j \cap G^*_\gamma), \sup_{x \in G^*_\gamma \setminus \mathcal{I}_{2\epsilon_j}} \rho(x, \widehat{G}_j \cap G^*_\gamma)\} \\
&\leq \max\{\epsilon_j, \sup_{x \in G^*_\gamma \setminus \mathcal{I}_{2\epsilon_j}} \rho(x, \widehat{G}_j \cap G^*_\gamma)\},
\end{aligned}$$

where the last step follows using Corollary 2.

Now consider any $x \in G^*_\gamma \setminus \mathcal{I}_{2\epsilon_j}$. By the triangle inequality, $\forall y \in \partial G^*_\gamma$ and $\forall z \in \mathcal{I}_{2\epsilon_j}$,

$$\begin{aligned}
\rho(x, \widehat{G}_j \cap G^*_\gamma) &\leq \rho(x, y) + \rho(y, z) + \rho(z, \widehat{G}_j \cap G^*_\gamma) \\
&\leq \rho(x, y) + \rho(y, z) + \sup_{z' \in \mathcal{I}_{2\epsilon_j}} \rho(z', \widehat{G}_j \cap G^*_\gamma) \\
&\leq \rho(x, y) + \rho(y, z) + \epsilon_j,
\end{aligned}$$

where the last step follows using Corollary 2. This implies that $\forall y \in \partial G^*_\gamma$,

$$\begin{aligned}
\rho(x, \widehat{G}_j \cap G^*_\gamma) &\leq \rho(x,y) + \inf_{z \in \mathcal{I}_{2\epsilon_j}} \rho(y,z) + \epsilon_j \\
&= \rho(x,y) + \rho(y, \mathcal{I}_{2\epsilon_j}) + \epsilon_j \\
&\leq \rho(x,y) + \sup_{y' \in \partial G^*_\gamma} \rho(y', \mathcal{I}_{2\epsilon_j}) + \epsilon_j \\
&\leq \rho(x,y) + 2C_3\epsilon_j + \epsilon_j,
\end{aligned}$$

where the last step invokes assumption [**B**]. This in turn implies that

$$\rho(x, \widehat{G}_j \cap G^*_\gamma) \leq \inf_{y \in \partial G^*_\gamma} \rho(x,y) + (2C_3+1)\epsilon_j \leq 2\epsilon_j + (2C_3+1)\epsilon_j.$$

The second step is true for $x \in G^*_\gamma \setminus \mathcal{I}_{2\epsilon_j}$, because if it was not true then $\forall y \in \partial G^*_\gamma$, $\rho(x,y) > 2\epsilon_j$ and hence there exists a closed $2\epsilon_j$-ball around x that is in G^*_γ. This contradicts the fact that $x \notin \mathcal{I}_{2\epsilon_j}$. Therefore, we have:

$$\sup_{x \in G^*_\gamma \setminus \mathcal{I}_{2\epsilon_j}} \rho(x, \widehat{G}_j \cap G^*_\gamma) \leq (2C_3+3)\epsilon_j$$

And going back to (2.19), we get:

$$\sup_{x \in G^*_\gamma} \rho(x, \widehat{G}_j) \leq (2C_3+3)\epsilon_j. \tag{2.19}$$

From Eqs. (2.18) and (2.19), we have that for all densities satisfying assumptions [**A1, B**], if $j \equiv j(n)$ is such that $2^j = O(s_n^{-1}(n/\log n)^{1/d})$, $j \geq J_0$, and $n \geq n_0 \equiv n_0(f_{\max}, d, \delta_1, \epsilon_o, C_1, \alpha)$, then with probability $> 1 - 3/n$,

$$d_\infty(\widehat{G}_j, G^*_\gamma) = \max\{\sup_{x \in G^*_\gamma} \rho(x, \widehat{G}_j), \sup_{x \in \widehat{G}_j} \rho(x, G^*_\gamma)\} \leq (2C_3+3)\epsilon_j.$$

And addressing both Case I ($j < J_0$) and Case II ($j \geq J_0$), we finally have that for all densities satisfying assumptions [**A1, B**], if $j \equiv j(n)$ is such that $2^j = O(s_n^{-1}(n/\log n)^{1/d})$, and $n \geq n_0 \equiv n_0(f_{\max}, d, \delta_1, \epsilon_o, C_1, \alpha)$, then with probability $> 1 - 3/n$,

$$d_\infty(\widehat{G}_j, G^*_\gamma) \leq \max(2C_3+3, 8\sqrt{d}\epsilon_o^{-1})\epsilon_j.$$

□

We now establish the result of Theorem 1. Since the regularity parameter α is known, the appropriate resolution can be chosen as $2^{-j} \asymp s_n(n/\log n)^{-\frac{1}{(d+2\alpha)}}$. Let Ω denote the event such that the bounds of Lemma 3 (with $\delta = 1/n$) and Lemma 4 hold. Then for $n \geq n_0$, $P(\bar{\Omega}) \leq 4/n$ where $\bar{\Omega}$ denotes the complement of Ω. For $n < n_0$, we can use the trivial inequality $P(\bar{\Omega}) \leq 1$. So we have, for all n

$$P(\bar{\Omega}) \leq \max(4, n_0)\frac{1}{n} =: C'\frac{1}{n}$$

Here $C' \equiv C'(f_{\max}, d, \delta_1, \epsilon_o, C_1, \alpha)$.

So $\forall f \in \mathcal{F}_1^*(\alpha)$, we have: (Explanation for each step is provided after the equations.)

$$\begin{aligned}
\mathbb{E}[d_\infty(\widehat{G}_j, G_\gamma^*)] &= P(\Omega)\mathbb{E}[d_\infty(\widehat{G}_j, G_\gamma^*)|\Omega] + P(\bar{\Omega})\mathbb{E}[d_\infty(\widehat{G}_j, G_\gamma^*)|\bar{\Omega}] \\
&\leq \mathbb{E}[d_\infty(\widehat{G}_j, G_\gamma^*)|\Omega] + P(\bar{\Omega})\sqrt{d} \\
&\leq \max(2C_3 + 3, 8\sqrt{d}\epsilon_o^{-1}) \left[\left(\frac{\Psi_j}{C_1} \right)^{1/\alpha} + \sqrt{d}2^{-j} \right] + C'\frac{\sqrt{d}}{n} \\
&\leq C \max \left\{ \left(2^{jd}\frac{\log n}{n} \right)^{\frac{1}{2\alpha}}, 2^{-j}, \frac{1}{n} \right\} \\
&\leq C \max \left\{ s_n^{-d/2\alpha} \left(\frac{n}{\log n} \right)^{-\frac{1}{d+2\alpha}}, s_n \left(\frac{n}{\log n} \right)^{-\frac{1}{d+2\alpha}}, \frac{1}{n} \right\} \\
&\leq C s_n \left(\frac{n}{\log n} \right)^{-\frac{1}{d+2\alpha}}.
\end{aligned}$$

Here $C \equiv C(C_1, C_3, \epsilon_o, f_{\max}, \delta_1, d, \alpha)$. The second step follows by observing the trivial bounds $P(\Omega) \leq 1$ and since the domain $\mathcal{X} = [0,1]^d$, $\mathbb{E}[d_\infty(\widehat{G}_j, G_\gamma^*)|\bar{\Omega}] \leq \sqrt{d}$. The third step follows from Lemma 4 and the fourth one using Lemma 3. The fifth step follows since the chosen resolution $2^{-j} \asymp s_n(n/\log n)^{-\frac{1}{(d+2\alpha)}}$.

■

2.6.5 Proof of Theorem 2

To analyze the resolution chosen by the complexity penalized procedure of (2.9) based on the vernier, we first establish two results regarding the vernier. Using Lemma 2, we have the following corollary that bounds the deviation of true and empirical vernier.

Corollary 3. *Consider* $0 < \delta < 1$. *With probability at least* $1 - \delta$ *with respect to the draw of the data, the following is true for all* $j \geq 0$*:*

$$|\mathcal{V}_{\gamma,j} - \widehat{\mathcal{V}}_{\gamma,j}| \leq \Psi_{j'}.$$

Proof. Let $A_0 \in \mathcal{A}_j$ denote the cell achieving the min defining $\mathcal{V}_{\gamma,j}$ and $A_1 \in \mathcal{A}_j$ denote the cell achieving the min defining $\widehat{\mathcal{V}}_{\gamma,j}$. Also let A'_{00} and A'_{10} denote the subcells at resolution j' within A_0 and A_1, respectively, that have maximum average density deviation from γ. Similarly, let A'_{01} and A'_{11} denote the subcells at resolution j' within A_0 and A_1, respectively, that have maximum empirical density deviation from γ. Then we have: (Explanation for the steps are given after the equations.)

$$\begin{aligned}
\mathcal{V}_{\gamma,j} - \widehat{\mathcal{V}}_{\gamma,j} &= |\gamma - \bar{f}(A'_{00})| - |\gamma - \widehat{f}(A'_{11})| \\
&\leq |\gamma - \bar{f}(A'_{10})| - |\gamma - \widehat{f}(A'_{11})| \\
&\leq |\bar{f}(A'_{10}) - \widehat{f}(A'_{11})| \\
&= \max\{\bar{f}(A'_{10}) - \widehat{f}(A'_{11}), \widehat{f}(A'_{11}) - \bar{f}(A'_{10})\} \\
&\leq \max\{\bar{f}(A'_{10}) - \widehat{f}(A'_{10}), \widehat{f}(A'_{11}) - \bar{f}(A'_{11})\} \\
&\leq \max_{A \in \mathcal{A}_{j'}} |\bar{f}(A) - \widehat{f}(A)| \\
&\leq \Psi_{j'}
\end{aligned}$$

The first inequality invokes definition of A_0, the third inequality invokes definitions of the subcells A'_{10}, A'_{11}, and the last one follows from Lemma 2. Similarly,

$$\begin{aligned}
\widehat{\mathcal{V}}_{\gamma,j} - \mathcal{V}_{\gamma,j} &= |\gamma - \widehat{f}(A'_{11})| - |\gamma - \bar{f}(A'_{00})| \\
&\leq |\gamma - \widehat{f}(A'_{01})| - |\gamma - \bar{f}(A'_{00})| \\
&\leq |\bar{f}(A'_{00}) - \widehat{f}(A'_{01})|
\end{aligned}$$

Here the first inequality invokes definition of A_1. The rest follows as above, considering cell A_0 instead of A_1. □

The second result establishes that the vernier is sensitive to the resolution and density regularity.

Lemma 5. *Consider densities satisfying assumptions **[A]** and **[B]**. Recall that* $j' = \lfloor j + \log_2 s_n \rfloor$*, where* s_n *is a monotone diverging sequence. There exists* $C \equiv C(C_2, f_{\max}, \delta_2, \alpha) > 0$ *such that if* n *is large enough so that* $s_n > 8\max(3\epsilon_o^{-1}, 28, 12C_3)\sqrt{d}$*, then for all* $j \geq 0$*,*

$$\min(\delta_1, C_1)2^{-j'\alpha} \leq \mathcal{V}_{\gamma,j} \leq C(\sqrt{d}2^{-j})^{\alpha}.$$

Proof. We first establish the upper bound. Recall assumption [**A**] and consider the cell $A_0 \in \mathcal{A}_j$ that contains the point x_0. Then $A_0 \cap \partial G_\gamma^* \neq \emptyset$. Let A_0' denote the subcell at resolution j' within A_0 that has maximum average density deviation from γ. Consider two cases:

(i) If the resolution is high enough so that $\sqrt{d}2^{-j} \leq \delta_2$, then the density regularity assumption [**A2**] holds $\forall x \in A_0$ since $A_0 \subset B(x_0, \delta_2)$, the δ_2-ball around x_0. The same holds also for the subcell A_0'. Hence

$$|\gamma - \bar{f}(A_0')| \leq C_2(\sqrt{d}2^{-j})^{\alpha}$$

(ii) If the resolution is not high enough and $\sqrt{d}2^{-j} > \delta_2$, the following trivial bound holds:

$$|\gamma - \bar{f}(A_0')| \leq f_{\max} \leq \frac{f_{\max}}{\delta_2^{\alpha}}(\sqrt{d}2^{-j})^{\alpha}$$

The last step holds since $\sqrt{d}2^{-j} > \delta_2$.

Hence we can say for all $j \geq 0$ there exists $A_0 \in \mathcal{A}_j$ such that

$$\max_{A' \in \mathcal{A}_{j'} \cap A_0} |\gamma - \bar{f}(A')| = |\gamma - \bar{f}(A_0')| \leq \max\left(C_2, \frac{f_{\max}}{\delta_2^{\alpha}}\right)(\sqrt{d}2^{-j})^{\alpha}$$

This yields the upper bound on the vernier:

$$\mathcal{V}_{\gamma,j} \leq \max\left(C_2, \frac{f_{\max}}{\delta_2^{\alpha}}\right)(\sqrt{d}2^{-j})^{\alpha} := C(\sqrt{d}2^{-j})^{\alpha}$$

where $C \equiv C(C_2, f_{\max}, \delta_2, \alpha)$.

For the lower bound, consider any cell $A \in \mathcal{A}_j$. We will show that the level set regularity assumption [**B**] implies that for large enough n (so that the sidelength $2^{-j'}$ is small enough),

the boundary does not intersect all subcells at resolution j' within the cell A at resolution j. And in fact, there exists at least one subcell $A'_1 \in A \cap \mathcal{A}_{j'}$ such that $\forall x \in A'_1$,

$$\rho(x, \partial G^*_\gamma) \geq 2^{-j'}.$$

We establish this statement formally later on, but for now assume that it holds. The local density regularity condition **[A]** now gives that for all $x \in A'_1$, $|\gamma - f(x)| \geq \min(\delta_1, C_1 2^{-j'\alpha})$ $\geq \min(\delta_1, C_1) 2^{-j'\alpha}$. So we have

$$\max_{A' \in A \cap \mathcal{A}_{j'}} |\gamma - \bar{f}(A')| \geq |\gamma - \bar{f}(A'_1)| \geq \min(\delta_1, C_1) 2^{-j'\alpha}.$$

Since this is true for any $A \in \mathcal{A}_j$, in particular, this is true for the cell achieving the min defining $\mathcal{V}_{\gamma,j}$. Hence, the lower bound on the vernier $\mathcal{V}_{\gamma,j}$ follows.

We now formally prove that the level set regularity assumption **[B]** implies that for large enough n (so that $s_n > 8 \max(3\epsilon_o^{-1}, 28, 12C_3)\sqrt{d}$), $\exists A'_1 \in A \cap \mathcal{A}_{j'}$ such that $\forall x \in A'_1$,

$$\rho(x, \partial G^*_\gamma) \geq 2^{-j'}.$$

Observe that if we consider any cell at resolution $j'' := j' - 2$ that does not intersect the boundary ∂G^*_γ, then it contains a cell at resolution j' that is greater than $2^{-j'}$ away from the boundary. Thus, it suffices to show that for large enough n (so that $s_n > 8 \max(3\epsilon_o^{-1}, 28, 12C_3)\sqrt{d}$), $\exists A'' \in A \cap \mathcal{A}_{j''}$ such that $A'' \cap \partial G^*_\gamma = \emptyset$. We prove the last statement by contradiction. Suppose that for $s_n > 8 \max(3\epsilon_o^{-1}, 28, 12C_3)\sqrt{d}$, all subcells in A at resolution j'' intersect the boundary ∂G^*_γ. Let $\epsilon = 3\sqrt{d}2^{-j''}$. Then,

$$\epsilon = 3\sqrt{d}2^{-j''} = 12\sqrt{d}2^{-j'} < \frac{24\sqrt{d}}{s_n} 2^{-j} \leq \frac{24\sqrt{d}}{s_n} \leq \epsilon_o,$$

where the last step follows since $s_n \geq 24\sqrt{d}\epsilon_o^{-1}$. By choice of ϵ, every closed ϵ-ball in A must contain an entire subcell at resolution j'' and in fact must contain an open neighborhood around that subcell. Since the boundary intersects all subcells at resolution j'', this implies that every closed ϵ-ball in A contains a boundary point and in fact contains an open neighborhood around that boundary point. Thus, (i) every closed ϵ-ball in A contains points not

in G^*_γ, and hence cannot lie in $\mathcal{I}_\epsilon(G^*_\gamma)$. Also, observe that since all subcells in A at resolution j'' intersect the boundary of G^*_γ, (ii) there exists a boundary point x_1 that is within $\sqrt{d}2^{-j''}$ of the center of cell A. From (i) and (ii) it follows that,

$$\begin{aligned}\rho(x_1, \mathcal{I}_\epsilon(G^*_\gamma)) \geq \frac{2^{-j}}{2} - \sqrt{d}2^{-j''} - 2\epsilon &= \frac{2^{-j}}{2} - 28\sqrt{d}2^{-j'} \\ &> 2^{-j}\left(\frac{1}{2} - \frac{56\sqrt{d}}{s_n}\right) > \frac{2^{-j}}{4},\end{aligned}$$

where the last step follows since $s_n > 224\sqrt{d}$. However, assumption **[B]** implies that for $\epsilon \leq \epsilon_o$,

$$\rho(x_1, \mathcal{I}_\epsilon(G^*_\gamma)) \leq C_3\epsilon = 3C_3\sqrt{d}2^{-j''} = 12C_3\sqrt{d}2^{-j'} \leq \frac{24C_3\sqrt{d}2^{-j}}{s_n} \leq \frac{2^{-j}}{4},$$

where the last step follows since $s_n > 96C_3\sqrt{d}$, and we have a contradiction.

This completes the proof of Lemma 5. □

We are now ready to prove Theorem 2. To analyze the resolution $\widehat{j}$ chosen by (2.9), we first derive upper bounds on $\mathcal{V}_{\gamma,\widehat{j}}$ and $\Psi_{\widehat{j}'}$, that effectively characterize the approximation error and estimation error, respectively. Thus, a bound on the vernier $\mathcal{V}_{\gamma,\widehat{j}}$ will imply that the chosen resolution $\widehat{j}$ cannot be too coarse and a bound on the penalty will imply that the chosen resolution is not too fine. Using Corollary 3 and (2.9), we have the following oracle inequality that holds with probability at least $1 - \delta$:

$$\mathcal{V}_{\gamma,\widehat{j}} \leq \widehat{\mathcal{V}}_{\gamma,\widehat{j}} + \Psi_{\widehat{j}'} = \min_{0\leq j\leq J}\left\{\widehat{\mathcal{V}}_{\gamma,j} + \Psi_{j'}\right\} \leq \min_{0\leq j\leq J}\left\{\mathcal{V}_{\gamma,j} + 2\Psi_{j'}\right\}.$$

Lemma 5 provides an upper bound on the vernier $\mathcal{V}_{\gamma,j}$, and Lemma 3 provides an upper bound on the penalty $\Psi_{j'}$. We now plug these bounds into the oracle inequality. Here C may denote a different constant from line to line.

$$\begin{aligned}\mathcal{V}_{\gamma,\widehat{j}} \leq \widehat{\mathcal{V}}_{\gamma,\widehat{j}} + \Psi_{\widehat{j}'} &\leq C\min_{0\leq j\leq J}\left\{2^{-j\alpha} + \sqrt{2^{j'd}\frac{\log n}{n}}\right\} \\ &\leq C\min_{0\leq j\leq J}\left\{\max\left(2^{-j\alpha}, \sqrt{2^{jd}s_n^d\frac{\log n}{n}}\right)\right\} \\ &\leq Cs_n^{\frac{d\alpha}{d+2\alpha}}\left(\frac{n}{\log n}\right)^{-\frac{\alpha}{d+2\alpha}}.\end{aligned}$$

Here $C \equiv C(C_2, f_{\max}, \delta_2, d, \alpha)$. The second step uses the definition of j', and the last step follows by balancing the two terms for optimal resolution j^* given by $2^{-j^*} \asymp s_n^{\frac{d}{d+2\alpha}} (n/\log n)^{-\frac{1}{d+2\alpha}}$. This establishes the desired bounds on $\mathcal{V}_{\gamma,\widehat{j}}$ and $\Psi_{\widehat{j}'}$.

Now, using Lemma 5 and the definition of j', we have the following upper bound on the sidelength: For $s_n > 8\max(3\epsilon_o^{-1}, 28, 12C_3)\sqrt{d}$,

$$2^{-\widehat{j}} \leq s_n 2^{-\widehat{j}'} \quad \leq \quad s_n \left(\frac{\mathcal{V}_{\gamma,\widehat{j}}}{\min(\delta_1, C_1)} \right)^{\frac{1}{\alpha}} \leq c_2 s_n s_n^{\frac{d}{d+2\alpha}} \left(\frac{n}{\log n} \right)^{-\frac{1}{d+2\alpha}},$$

where $c_2 \equiv c_2(C_1, C_2, f_{\max}, \delta_1, \delta_2, d, \alpha) > 0$. Also notice that since $2^J \asymp s_n^{-1} (n/\log n)^{1/d}$, we have $2^{j'} \leq 2^{J'} \leq s_n 2^J \asymp (n/\log n)^{1/d}$, and thus j' satisfies the condition of Lemma 3. Therefore, using Lemma 3, we get a lower bound on the sidelength: With probability at least $1 - 2/n$,

$$\begin{aligned} 2^{-\widehat{j}} > \frac{s_n}{2} 2^{-\widehat{j}'} \geq \frac{s_n}{2} \left(\frac{\Psi^2_{\widehat{j}'}}{c_3^2} \frac{n}{\log n} \right)^{-\frac{1}{d}} &\geq c_1 s_n \left(s_n^{\frac{2d\alpha}{d+2\alpha}} \left(\frac{n}{\log n} \right)^{-\frac{2\alpha}{d+2\alpha}} \frac{n}{\log n} \right)^{-1/d} \\ &= c_1 s_n s_n^{\frac{-2\alpha}{d+2\alpha}} \left(\frac{n}{\log n} \right)^{\frac{-1}{d+2\alpha}} \\ &= c_1 s_n^{\frac{d}{d+2\alpha}} \left(\frac{n}{\log n} \right)^{\frac{-1}{d+2\alpha}}, \end{aligned}$$

where $c_1 \equiv c_1(C_2, f_{\max}, \delta_2, d, \alpha) > 0$. So we have for $s_n > 8\max(3\epsilon_o^{-1}, 28, 12C_3)\sqrt{d}$, with probability at least $1 - 2/n$,

$$c_1 s_n^{\frac{d}{d+2\alpha}} \left(\frac{n}{\log n} \right)^{-\frac{1}{d+2\alpha}} \leq 2^{-\widehat{j}} \leq c_2 s_n s_n^{\frac{d}{d+2\alpha}} \left(\frac{n}{\log n} \right)^{-\frac{1}{d+2\alpha}}, \tag{2.20}$$

where $c_1 \equiv c_1(C_2, f_{\max}, \delta_2, d, \alpha) > 0$ and $c_2 \equiv c_2(C_1, C_2, f_{\max}, \delta_1, \delta_2, d, \alpha) > 0$. Hence the automatically chosen resolution behaves as desired.

Now we can invoke Lemma 4 to derive the rate of convergence for the Hausdorff error. Consider large enough $n \geq n_1(C_3, \epsilon_o, d)$ so that $s_n > 8\max(3\epsilon_o^{-1}, 28, 12C_3)\sqrt{d}$. Also, recall that the condition of Lemma 4 requires that $n \geq n_0(f_{\max}, d, \delta_1, \epsilon_o, C_1, \alpha)$. Pick $n \geq \max(n_0, n_1)$ and let Ω denote the event such that the bounds of Lemma 3, Lemma

4, and the upper and lower bounds on the chosen sidelength in (2.20) hold with $\delta = 1/n$. Then, we have $P(\bar{\Omega}) \leq 6/n$. For $n < \max(n_0, n_1)$, we can use the trivial inequality $P(\bar{\Omega}) \leq 1$. So we have, for all n

$$P(\bar{\Omega}) \leq \max(6, \max(n_0, n_1))\frac{1}{n} =: C\frac{1}{n},$$

where $C \equiv C(C_1, C_3, \epsilon_o, f_{\max}, \delta_1, d, \alpha)$.

So $\forall f \in \mathcal{F}_2^*(\alpha)$, we have: (Here C may denote a different constant from line to line. Explanation for each step is provided after the equations.)

$$\begin{aligned}
\mathbb{E}[d_\infty(\widehat{G}, G_\gamma^*)] &= P(\Omega)\mathbb{E}[d_\infty(\widehat{G}, G_\gamma^*)|\Omega] + P(\bar{\Omega})\mathbb{E}[d_\infty(\widehat{G}, G_\gamma^*)|\bar{\Omega}] \\
&\leq \mathbb{E}[d_\infty(\widehat{G}, G_\gamma^*)|\Omega] + P(\bar{\Omega})\sqrt{d} \\
&\leq C\left[\left(\frac{\Psi_{\widehat{j}}}{C_1}\right)^{1/\alpha} + \sqrt{d}2^{-\widehat{j}} + \frac{\sqrt{d}}{n}\right] \\
&\leq C\max\left\{\left(2^{\widehat{j}d}\frac{\log n}{n}\right)^{\frac{1}{2\alpha}}, 2^{-\widehat{j}}, \frac{1}{n}\right\} \\
&\leq C\max\left\{s_n^{\frac{-d^2/2\alpha}{d+2\alpha}}\left(\frac{n}{\log n}\right)^{-\frac{1}{d+2\alpha}}, s_n s_n^{\frac{d}{d+2\alpha}}\left(\frac{n}{\log n}\right)^{-\frac{1}{d+2\alpha}}, \frac{1}{n}\right\} \\
&\leq Cs_n s_n^{\frac{d}{d+2\alpha}}\left(\frac{n}{\log n}\right)^{-\frac{1}{d+2\alpha}} \\
&\leq Cs_n^2\left(\frac{n}{\log n}\right)^{-\frac{1}{d+2\alpha}}.
\end{aligned}$$

Here $C \equiv C(C_1, C_2, C_3, \epsilon_o, f_{\max}, \delta_1, \delta_2, d, \alpha)$. The second step follows by observing the trivial bounds $P(\Omega) \leq 1$ and since the domain $\mathcal{X} = [0,1]^d$, $\mathbb{E}[d_\infty(\widehat{G}, G_\gamma^*)|\bar{\Omega}] \leq \sqrt{d}$. The third step follows from Lemma 4 and the fourth one from Lemma 3. The fifth step follows using the upper and lower bounds established on $2^{-\widehat{j}}$ in (2.20).

■

2.6.6 Proof of Proposition 3

We proceed by formally defining the class $\mathcal{F}_{BF}(\alpha, \zeta)$. The class corresponds to densities bounded above by $f_{\max}$, satisfying the local density regularity assumptions **[A1,A2]** for

points within the support set, and the densities have support sets that are Hölder-ζ boundary fragments. That is,

$$G_0^* = \{(\tilde{x}, x_d); \tilde{x} \in [0,1]^{d-1}, 0 \leq x_d \leq g(\tilde{x})\},$$

where the function g satisfies $h \leq g(\tilde{x}) \leq 1 - h$, where $0 < h < 1/2$ is a constant, and g is Hölder-ζ smooth. That is, g has continuous partial derivatives of up to order $[\zeta]$, where $[\zeta]$ denotes the maximal integer that is $< \zeta$, and $\exists \delta > 0$ such that

$$\forall \tilde{z}, \tilde{x} \in [0,1]^{d-1} : ||\tilde{z} - \tilde{x}|| \leq \delta \Rightarrow |g(\tilde{z}) - TP_{\tilde{x}}(\tilde{z}, [\zeta])| \leq L\|z - x\|^{\alpha}$$

where $L, \zeta > 0$, $TP_{\tilde{x}}(\cdot, [\zeta])$ denotes the degree $[\zeta]$ Taylor polynomial approximation of g expanded around $\tilde{x}$, and $||\cdot||$ denotes Euclidean norm.

The proof is motivated by the minimax lower bound proof of Theorem 1 in [35], however the construction is slightly different for support set estimation. For the sake of completeness, we present the entire proof here. We will use the following theorem from [57].

Theorem 4 (Main Theorem of Risk Minimization (Kullback divergence version))**.** *Let Θ be a class of models. Associated with each model $\theta \in \Theta$ we have a probability measure P_θ. Let $M \geq 2$ be an integer and let $d(\cdot, \cdot) : \Theta \times \Theta \to \mathbb{R}$ be a semidistance. Suppose we have $\{\theta_0, \ldots, \theta_M\} \in \Theta$ such that*

1. $d(\theta_j, \theta_k) \geq 2s > 0, \quad \forall_{0 \leq j,k \leq M}$,
2. $P_{\theta_j} \ll P_{\theta_0}, \quad \forall_{j=1,\ldots,M}$,
3. $\frac{1}{M} \sum_{j=1}^{M} KL(P_{\theta_j} \| P_{\theta_0}) \leq \kappa \log M$,

where $0 < \kappa < 1/8$. The following bound holds.

$$\inf_{\hat{\theta}_n} \sup_{\theta \in \Theta} P_\theta \left(d(\hat{\theta}, \theta) \geq s \right) \geq \frac{\sqrt{M}}{1 + \sqrt{M}} \left(1 - 2\kappa - 2\sqrt{\frac{\kappa}{\log M}} \right) > 0,$$

where the infimum is taken with respect to the collection of all possible estimators of θ, and KL denotes the Kullback-Leibler divergence.

The following corollary follows immediately from the theorem using Markov's inequality.

Corollary 4. *Under the assumptions of Theorem 4 we have*

$$\inf_{\hat{\theta}_n} \sup_{\theta\in\Theta} \mathbb{E}[d(\hat{\theta},\theta)] \geq s\frac{\sqrt{M}}{1+\sqrt{M}}\left(1-2\kappa-2\sqrt{\frac{\kappa}{\log M}}\right) > cs,$$

for some $c \equiv c(\kappa, M) > 0$.

We now construct the model class $\Theta \equiv \mathcal{F}$ of densities that is a subset of densities from the class $\mathcal{F}_{BF}(\alpha,\zeta)$. Thus, Corollary 4 would give a minimax lower bound for the class $\mathcal{F}_{BF}(\alpha,\zeta)$. Consider $\{f_0,\ldots,f_M\} \in \mathcal{F}$ as follows. Let $x = (\tilde{x}, x_d) \in [0,1]^d$, where $\tilde{x} \in [0,1]^{d-1}$ and $x_d \in [0,1]$. Also let

$$m = \left\lceil c_0 \left(\frac{n}{\log n}\right)^{\frac{1}{\zeta(\alpha+1)+d-1}} \right\rceil,$$

where $c_0 > 0$ is a constant to be specified later. Define

$$\tilde{x}_{\tilde{j}} = \frac{\tilde{j}-1/2}{m}, \quad B_{\tilde{j}} = \left\{ x : \tilde{x} \in \left(\tilde{x}_{\tilde{j}} - \frac{1}{2m}, \tilde{x}_{\tilde{j}} + \frac{1}{2m}\right)\right\}$$

and

$$\eta_{\tilde{j}}(\tilde{x}) = \frac{L}{m^\zeta} K(m(\tilde{x}-\tilde{x}_{\tilde{j}})),$$

where $\tilde{j} \in \{1,\ldots,m\}^{d-1}$ and $K > 0$ is a Hölder-ζ function with constant 1, and $\mathrm{supp}(K) = (-1/2,1/2)^{d-1}$. Now define

$$f_0(x) = g_0(x) \quad \text{and} \quad f_{\tilde{j}}(x) = g_0(x) + g_{1,\tilde{j}}(x) + g_2(x),$$

where

$$g_0(x) = \begin{cases} 0 & x_d > 1/2, \tilde{x} \in [0,1]^{d-1} \\ \frac{C_1+C_2}{2}\left(\frac{1}{2}-x_d\right)^\alpha & 1/2-\delta_2 < x_d \leq 1/2, \tilde{x} \in [0,1]^{d-1} \\ \frac{1-\frac{C_1+C_2}{2}\frac{\delta_2^{\alpha+1}}{\alpha+1}}{1/2-\delta_2} & x_d \leq 1/2-\delta_2, \tilde{x} \in [0,1]^{d-1} \end{cases}$$

$$g_{1,\tilde{j}}(x) = \begin{cases} -\frac{C_1+C_2}{2}\left(\frac{1}{2}-x_d\right)^\alpha & \frac{1}{2}-\eta_{\tilde{j}}(\tilde{x}) < x_d \leq \frac{1}{2}, \tilde{x} \in B_{\tilde{j}} \\ -\frac{C_1+C_2}{2}\left(\frac{1}{2}-x_d\right)^\alpha + & \frac{1}{2}-\frac{3}{2}\eta_{\tilde{j}}(\tilde{x}) < x_d \leq \frac{1}{2}-\eta_{\tilde{j}}(\tilde{x}), \\ \quad \frac{C_1+C_2}{2}\left(\frac{1}{2}-\eta_{\tilde{j}}(\tilde{x})-x_d\right)^\alpha & \tilde{x} \in B_{\tilde{j}} \\ 0 & \text{elsewhere} \end{cases}$$

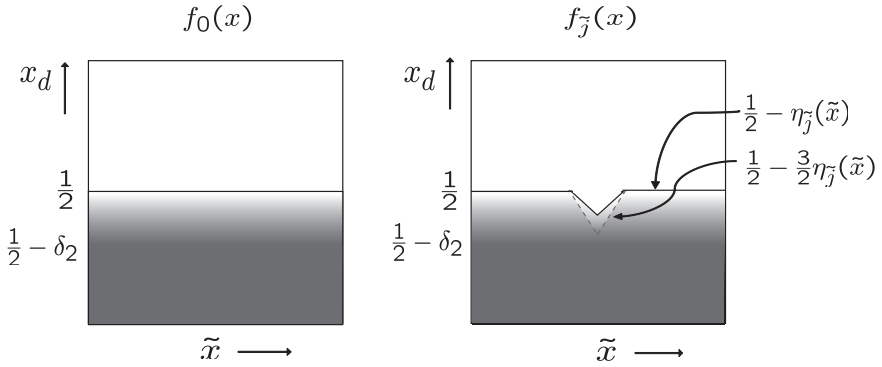


Figure 2.2 Densities used in the lower bound construction for Hausdorff accurate support set estimation.

and

$$g_2(x) = \begin{cases} \frac{C'(\alpha,L,K,C_1,C_2)}{1/2-\delta_2} m^{-(\zeta(\alpha+1)+d-1)} & x_d \le 1/2-\delta_2, \tilde{x} \in [0,1]^{d-1} \\ 0 & \text{elsewhere} \end{cases}$$

where $C'(\alpha, L, K, C_1, C_2) = \frac{C_1+C_2}{2}\left(1-\frac{1}{2^{\alpha+1}}\right)\frac{L^{\alpha+1}}{\alpha+1}\|K\|_{\alpha+1}^{\alpha+1}$. See Figure 2.2.

Thus, $M = m^{d-1}$. Observe that $f_0, \ldots, f_M$ are valid densities since $\int g_0 = 1$, $\int g_{1,\tilde{j}} + \int g_2 = 0$ and $f_0, \ldots, f_M \ge 0$ provided δ_2 is small enough but fixed and n is large enough but fixed. Moreover, observe that provided δ_1 is small enough but fixed, the densities satisfy assumptions **[A1, A2]** for all points within the support. The exact requirements on δ_1, δ_2 and n can be specified but are cumbersome and of no interest to the results. The corresponding support sets are given as:

$$\begin{aligned} G_0^* &= \{x : 0 \le x_d < 1/2\} \\ G_{\tilde{j}}^* &= \{x : 0 \le x_d < 1/2 - \eta_{\tilde{j}}(\tilde{x})\} \end{aligned}$$

Observe that the support sets are Hölder-ζ boundary fragments. Thus, $\mathcal{F} \subset \mathcal{F}_{BF}(\alpha, \zeta)$.

Now, we show that $\mathcal{F}$ satisfies the assumptions of Theorem 4 for $d \equiv d_\infty$.

1. For all $\tilde{j} \ne \tilde{k}$,

$$d_\infty(G_{\tilde{j}}^*, G_{\tilde{k}}^*) = \max(\max_{\tilde{x}} \eta_{\tilde{j}}(\tilde{x}), \max_{\tilde{x}} \eta_{\tilde{k}}(\tilde{x})) = L \max_{\tilde{x}} K(\tilde{x}) m^{-\zeta} =: 2s > 0,$$

and also for all $\tilde{j}$

$$d_\infty(G_{\tilde{j}}^*, G_0^*) = \max_{\tilde{x}} \eta_{\tilde{j}}(\tilde{x}) = L \max_{\tilde{x}} K(\tilde{x}) m^{-\zeta} =: 2s > 0.$$

2. Clearly, $P_{\tilde{j}} \ll P_0, \quad \forall_{\tilde{j}}$ by construction.

3. We now evaluate the KL divergence.

$$\mathrm{KL}(P_{\tilde{j}} \| P_0) = \mathbb{E}_{\tilde{j}} \left[\sum_{i=1}^{n} \log \frac{f_{\tilde{j}}(X_i)}{f_0(X_i)} \right] = n \int_{[0,1]^d} \log \frac{f_{\tilde{j}}(x)}{f_0(x)} f_{\tilde{j}}(x) dx$$

The last integral consists of three terms considering where $f_{\tilde{j}}(x) > 0$ that we evaluate next.

$$\begin{aligned}
\mathrm{I} &= n \int_{[0,1]^{d-1}} \int_0^{\frac{1}{2}-\delta_2} \log \frac{g_0(x) + g_2(x)}{g_0(x)} \left(g_0(x) + g_2(x)\right) dx_d d\tilde{x} \\
&= n \log \left(1 + \frac{C' m^{-(\zeta(\alpha+1)+d-1)}}{1 - \frac{C_1+C_2}{2} \frac{h^{\alpha+1}}{\alpha+1}} \right) C' m^{-(\zeta(\alpha+1)+d-1)} \\
&\leq n C' C'' m^{-2(\zeta(\alpha+1)+d-1)} \\
&= C' C'' (2c_0)^{-2(\zeta(\alpha+1)+d-1)} \frac{(\log n)^2}{n} \leq \kappa \log M
\end{aligned}$$

where the inequality follows from $\log(1+x) \leq x$ and defining $C'' = \frac{C'}{1 - \frac{C_1+C_2}{2} \frac{h^{\alpha+1}}{\alpha+1}}$. In the last step, $0 < \kappa < 1/8$ by appropriate choice of c_0.

$$\mathrm{II} = n \int_{[0,1]^{d-1} \setminus B_{\tilde{j}}} \int_{1/2-\delta_2}^{1/2} \log \frac{g_0(x)}{g_0(x)} g_0(x) dx_d d\tilde{x} = 0$$

$$\begin{aligned}
\mathrm{III} &= n \int_{B_{\tilde{j}}} \int_{\frac{1}{2} - \frac{3}{2}\eta_{\tilde{j}}(\tilde{x})}^{\frac{1}{2} - \eta_{\tilde{j}}(\tilde{x})} \log \frac{g_0(x) + g_1(x)}{g_0(x)} \left(g_0(x) + g_1(x)\right) dx_d d\tilde{x} \\
&= n \int_{B_{\tilde{j}}} \int_{\frac{1}{2} - \frac{3}{2}\eta_{\tilde{j}}(\tilde{x})}^{\frac{1}{2} - \eta_{\tilde{j}}(\tilde{x})} \log \left(1 - \frac{\eta_{\tilde{j}}(\tilde{x})}{\frac{1}{2} - x_d} \right)^{\alpha} \frac{C_1 + C_2}{2} \left(\frac{1}{2} - \eta_{\tilde{j}}(\tilde{x}) - x_d \right)^{\alpha} dx_d d\tilde{x} \\
&\leq 0
\end{aligned}$$

Finally, we get

$$\frac{1}{M} \sum_{j=1}^{M} \mathrm{KL}(P_{\tilde{j}} \| P_0) \leq \kappa \log M.$$

Thus, all the conditions of Theorem 4 are satisfied and Corollary 4 implies the desired lower bound since $s := L \max_{\tilde{x}} K(\tilde{x}) m^{-\zeta}/2$.

■

2.6.7 Proof sketch of Theorem 3

We derive an upper bound on the Hausdorff error of the estimator proposed in (2.10) for support set estimation ($\gamma = 0$). We follow the proof of Theorem 1, except that instead of Lemma 2 based on the VC inequalities, we will use the following lemma that is based on the Craig-Bernstein inequality [54].

Lemma 6. *With probability at least $1 - 1/n$, the following is true for all $j \geq 0$ and all $A \in \mathcal{A}_j$*

$$\bar{f}(A) \leq 2\widehat{f}(A) + \Psi_j^0$$

Similarly, with probability at least $1 - 1/n$, the following is true for all $j \geq 0$ and all $A \in \mathcal{A}_j$

$$\widehat{f}(A) \leq 2\bar{f}(A) + \Psi_j^0$$

Proof. The proof hinges on the following concentration inequality due to Craig [54]:

Proposition 5 (Craig93)**.** *Let $\{U_i\}_{i=1}^n$ be independent random variables satisfying the Bernstein moment condition*

$$\mathbb{E}[|U_i - \mathbb{E}[U_i]|^k] = var(U_i)\frac{k!}{2}h^{k-2},$$

for some $h > 0$ and all $k \geq 2$. Then

$$P\left(\frac{1}{n}(U_i - \mathbb{E}[U_i]) \geq \frac{\tau}{n\epsilon} + \frac{n\epsilon\ var(\frac{1}{n}U_i)}{2(1-c)}\right) \leq e^{-\tau}$$

for $0 < \epsilon h \leq c < 1$ and $\tau > 0$.

First let $U_i = -\mathbf{1}_{X_i \in A}$. Then $\mathbb{E}[U_i] = -P(A)$. Since $|U_i - \mathbb{E}[U_i]| \leq 1$, the Bernstein moment condition is satisfied as follows.

$$\begin{aligned}\mathbb{E}[|U_i - \mathbb{E}[U_i]|^k] &= \mathbb{E}[|U_i - \mathbb{E}[U_i]|^{k-2}|U_i - \mathbb{E}[U_i]|^2] \leq \mathbb{E}[|U_i - \mathbb{E}[U_i]|^2] \\ &= var(U_i) \leq var(U_i)\frac{k!}{2}h^{k-2}\end{aligned}$$

for $h = 1$ and all $k \geq 2$. Therefore, we have with probability $> 1 - e^{-\tau}$,

$$\begin{aligned}-\widehat{P}(A) + P(A) &\leq \frac{\tau}{n\epsilon} + \frac{n\epsilon\ var(\frac{1}{n}U_i)}{2(1-c)} \leq \frac{\tau}{n\epsilon} + \frac{\epsilon\ var(U_i)}{2(1-c)} \\ &\leq \frac{\tau}{n\epsilon} + \frac{\epsilon P(A)}{2(1-c)}\end{aligned}$$

The last step follows since $var(U_i) \leq \mathbb{E}[|U_i|^2] \leq \mathbb{E}[|U_i|] = P(A)$. Setting $\epsilon = c = 1/2$, we have with probability $> 1 - 2^{jd}e^{-\tau}$, for all $A \in \mathcal{A}_j$

$$P(A) \leq 2\widehat{P}(A) + \frac{4\tau}{n}$$

Now let $\tau = \log \frac{2^{jd}}{\delta_j}$, $\delta_j = \delta 2^{-(j+1)}$ and apply union bound to get with probability $> 1 - \delta$, for all resolutions $j \geq 0$ and all $A \in \mathcal{A}_j$

$$P(A) \leq 2\widehat{P}(A) + \frac{4 \log \frac{2^{j(d+1)}2}{\delta}}{n}.$$

The first result follows by dividing by $\mu(A) = 2^{-jd}$ and setting $\delta = 1/n$.

To get the second result, let $U_i = \mathbf{1}_{X_i \in A}$ and proceed as before. We get with probability $> 1 - \delta$, for all resolutions $j \geq 0$ and all $A \in \mathcal{A}_j$

$$\widehat{P}(A) \leq \frac{3}{2}P(A) + \frac{2 \log \frac{2^{j(d+1)}2}{\delta}}{n} \leq 2P(A) + \frac{4 \log \frac{2^{j(d+1)}2}{\delta}}{n}.$$

□

Analogous to Lemma 3, there exist constants $c_5, c_6 > 0$, such that for resolutions satisfying $2^j = O((n/\log n)^{1/d})$,

$$c_5 \frac{2^{jd} \log n}{n} \leq \Psi_j^0 \leq c_6 \frac{2^{jd} \log n}{n}. \tag{2.21}$$

Also, the following analogue of Proposition 4 holds.

Proposition 6. *For any $n \geq n_1(d, C_1)$, if $j \equiv j(n)$ is such that $2^j = O(s_n^{-1}(n/\log n)^{1/d})$, then, with probability at least $1 - 1/n$*

$$\sup_{x \in \widehat{G}_{0,j} \Delta G_0^*} \rho(x, \partial G_0^*) \leq \left(\frac{\Psi_j^0}{C_1}\right)^{1/\alpha} + \sqrt{d}2^{-j} = \epsilon_j.$$

Proof. Proof follows along the lines of the proof of Proposition 4. If $\widehat{G}_{0,j} \Delta G_0^* = \emptyset$, then $\sup_{x \in \widehat{G}_{0,j} \Delta G_0^*} \rho(x, \partial G_0^*) = 0 < \epsilon_j$ by definition. If $\widehat{G}_{0,j} \Delta G_0^* \neq \emptyset$, consider $x \in \widehat{G}_{0,j} \Delta G_0^*$. Let $A_x \in \mathcal{A}_j$ denote the cell containing x at resolution j. Consider two cases:

(i) $A_x \cap \partial G_0^* \neq \emptyset$. This implies that

$$\rho(x, \partial G_0^*) \leq \sqrt{d} 2^{-j}.$$

(ii) $A_x \cap \partial G_0^* = \emptyset$. Since $x \in \widehat{G}_{0,j} \Delta G_0^*$, it is erroneously excluded from the support set estimate $\widehat{G}_{0,j}$. Therefore, $\bar{f}(A_x) > 0$ and $\widehat{f}(A_x) = 0$. (Notice that if $\bar{f}(A_x) = 0$, then $\widehat{f}(A_x) = 0$ as no data points lie in A_x, hence a cell cannot be erroneously included in the support set estimate.) Since $\bar{f}(A_x) > 0$ and $A_x \cap \partial G_0^* = \emptyset$, $A_x \subset G_0^*$. Using Lemma 6, since $\widehat{f}(A_x) = 0$, we get $\bar{f}(A_x) \leq \Psi_j^0$ with probability at least $1 - 1/n$.

Now let x_1 be any point in A_x such that $0 < f(x_1) \leq \bar{f}(A_x)$ (Notice that at least one such point must exist in A_x since this cell does not intersect the boundary). As argued above, $\bar{f}(A_x) \leq \Psi_j^0$ with probability at least $1 - 1/n$. From (2.21), for resolutions satisfying $2^j = O(s_n^{-1}(n/\log n)^{1/d})$, and for large enough $n \geq n_1(d, \delta_1)$, $\Psi_j^0 \leq \delta_1$ and hence $f(x_1) \leq \delta_1$, with probability at least $1 - 1/n$. Also, $x_1 \in A_x \subset G_0^*$. Thus, the density regularity assumption **[A1]** holds at x_1 with probability $> 1 - 1/n$ and we have

$$\rho(x_1, \partial G_0^*) \leq \left(\frac{f(x_1)}{C_1}\right)^{1/\alpha} \leq \left(\frac{\bar{f}(A_x)}{C_1}\right)^{1/\alpha} \leq \left(\frac{\Psi_j^0}{C_1}\right)^{1/\alpha}.$$

Since $x, x_1 \in A_x$,

$$\rho(x, \partial G_0^*) \leq \rho(x_1, \partial G_0^*) + \sqrt{d} 2^{-j} \leq \left(\frac{\Psi_j^0}{C_1}\right)^{1/\alpha} + \sqrt{d} 2^{-j}.$$

□

Rest of the proof of Theorem 3 follows as for Theorem 1. Since Ψ_j^0 behaves essentially as the square of Ψ_j, we get a bound that scales as $s_n(n/\log n)^{-1/(d+\alpha)}$.

■

2.6.8 Proof sketch for $\alpha \geq 0$

First consider the non-adaptive setting when α is known to be zero. In this case the plug-in histogram estimator of (2.7), along with a choice of resolution j such that $2^{-j} \asymp$

$s_n(n/\log n)^{-1/d}$, achieves minimax optimal performance for the class of densities given by $\mathcal{F}_1^*(0)$. This follows along the lines of the proof of Theorem 1 except that for the case $\alpha = 0$, the following result analogous to Proposition 4 holds.

Proposition 7. *For any $n \geq n_1(f_{\max}, d, C_1)$, if $j \equiv j(n)$ is such that $2^j = O(s_n^{-1}\,(n/\log n)^{1/d})$, then, with probability at least $1 - 2/n$,*

$$\sup_{x \in \widehat{G}_j \Delta G_\gamma^*} \rho(x, \partial G_\gamma^*) \leq \sqrt{d} 2^{-j} =: \epsilon_j.$$

Proof. If $\alpha = 0$, then $\forall x \in \mathcal{X}$, $|\gamma - f(x)| \geq \min(C_1, \delta_1)$. Consider any cell A that does not intersect the boundary. Then $|\gamma - \bar{f}(A)| \geq \min(C_1, \delta_1) \geq \Psi_j \geq |\bar{f}(A) - \widehat{f}(A)|$. The second step holds, with probability at least $1 - 1/n$ for $n \geq n_1(f_{\max}, d, C_1, \delta_1)$ and resolutions satisfying $2^j = O(s_n^{-1}(n/\log n)^{1/d})$, using Lemma 3. And the third step follows with probability at least $1 - 1/n$ using Lemma 2 (with $\delta = 1/n$). Since $|\gamma - \bar{f}(A)| \geq |\bar{f}(A) - \widehat{f}(A)|$, for resolutions satisfying $2^j = O(s_n^{-1}(n/\log n)^{1/d})$ and $n \geq n_1\ (f_{\max}, d, C_1, \delta_1)$, with probability at least $1 - 2/n$, all cells A that do not intersect the boundary are correctly included or excluded from the level set estimate. Hence, $\sup_{x \in G_\gamma^* \Delta \widehat{G}_j} \rho(x, \partial G_\gamma^*) \leq \sqrt{d} 2^{-j}$. □

This yields a corresponding Hausdorff error bound (analogous to Lemma 4) of

$$d_\infty(\widehat{G}_j, G_\gamma^*) \leq \max(2C_3 + 3, 8\sqrt{d}\epsilon_o^{-1}) \left[2\sqrt{d} 2^{-j}\right]. \tag{2.22}$$

Thus, the result follows as $2^{-j} \asymp s_n(n/\log n)^{-1/d}$.

Next, we prove that adaptivity can be achieved, and hence Theorem 2 holds, for the whole range $\alpha \geq 0$ using the modified vernier and penalty proposed in Section 2.4.4. First, notice that Corollary 3 still holds for the modified vernier and modified penalty since $\mathcal{V}_{\gamma,j}, \widehat{\mathcal{V}}_{\gamma,j}$ as well as $\Psi_{j'}$ are all scaled by the same factor of $2^{-j'/2}$. And we have the following analogue of Lemma 5 using the modified vernier:

Lemma 7. *Consider densities satisfying assumption **[A]** for $\alpha \geq 0$ and assumption **[B]**. Recall that $j' = \lfloor j + \log_2 s_n \rfloor$, where s_n is a diverging sequence. There exists $C \equiv C(C_2, f_{\max}, \delta_1) > 0$ such that for n large enough (so that $s_n > 8\max(3\epsilon_o^{-1}, 28, 12C_3)\sqrt{d}$), then for all $j \geq 0$*

$$\min(\delta_1, C_1) 2^{-j'\alpha} 2^{-j'/2} \leq \mathcal{V}_{\gamma,j} \leq C(\sqrt{d} 2^{-j})^\alpha 2^{-j'/2}.$$

Following the proof of Theorem 2, we derive upper bounds on $\mathcal{V}_{\gamma,\widehat{j}}$ and $\Psi_{\widehat{j}'}$ using the oracle inequality. Since both the modified vernier and penalty are scaled by the same factor, the two terms in the oracle inequality are still balanced for the same optimal resolution j^* given by $2^{-j^*} \asymp s_n^{\frac{d}{d+2\alpha}} (n/\log n)^{-\frac{1}{d+2\alpha}}$. Hence we get:

$$\mathcal{V}_{\gamma,\widehat{j}} \leq \widehat{\mathcal{V}}_{\gamma,\widehat{j}} + \Psi_{\widehat{j}'} \;\; \leq \;\; C 2^{-j^{*'}/2} 2^{-j^*\alpha} \leq C s_n^{-1/2} s_n^{\frac{d(\alpha+1/2)}{d+2\alpha}} \left(\frac{n}{\log n}\right)^{-\frac{(\alpha+1/2)}{d+2\alpha}}.$$

Using this upper bound on $\mathcal{V}_{\gamma,\widehat{j}}$ and $\Psi_{\widehat{j}'}$, we derive upper and lower bounds on the chosen resolution $\widehat{j}$ as in the proof of Theorem 2. Using Lemma 7, we have the following upper bound on the sidelength: For $s_n > 8\max(3\epsilon_o^{-1}, 28, 12C_3)\sqrt{d}$,

$$\begin{aligned} 2^{-\widehat{j}} \leq s_n \left(\frac{\mathcal{V}_{\gamma,\widehat{j}}}{\min(\delta_1, C_1)}\right)^{1/(\alpha+1/2)} &\leq c_2 s_n^{\frac{2\alpha}{2\alpha+1}} s_n^{\frac{d}{d+2\alpha}} \left(\frac{n}{\log n}\right)^{-\frac{1}{d+2\alpha}} \\ &\leq c_2 s_n s_n^{\frac{d}{d+2\alpha}} \left(\frac{n}{\log n}\right)^{-\frac{1}{d+2\alpha}}. \end{aligned}$$

And using Lemma 3 for the modified penalty, we have:

$$c_3 2^{-j'/2} \sqrt{2^{j'd} \frac{\log n}{n}} \leq \Psi_{j'}.$$

This provides a lower bound on the sidelength:

$$\begin{aligned} 2^{-\widehat{j}} > \frac{s_n}{2} \left(\frac{\Psi_{\widehat{j}'}^2}{4c_3^2} \frac{n}{\log n}\right)^{-\frac{1}{(d-1)}} &\geq c_1 s_n \left(s_n^{-1} s_n^{\frac{2d(\alpha+1/2)}{d+2\alpha}} \left(\frac{n}{\log n}\right)^{-\frac{2(\alpha+1/2)}{d+2\alpha}} \frac{n}{\log n}\right)^{-\frac{1}{(d-1)}} \\ &= c_1 s_n s_n^{\frac{1}{(d-1)}} s_n^{\frac{-2d(\alpha+1/2)}{(d-1)(d+2\alpha)}} \left(\frac{n}{\log n}\right)^{\frac{-1}{d+2\alpha}} \\ &= c_1 s_n^{\frac{d}{d+2\alpha}} \left(\frac{n}{\log n}\right)^{\frac{-1}{d+2\alpha}}. \end{aligned}$$

So as before we have for $s_n > 8\max(3\epsilon_o^{-1}, 28, 12C_3)\sqrt{d}$, with probability at least $1 - 2/n$,

$$c_1 s_n^{\frac{d}{d+2\alpha}} \left(\frac{n}{\log n}\right)^{-\frac{1}{d+2\alpha}} \leq 2^{-\widehat{j}} \leq c_2 s_n s_n^{\frac{d}{d+2\alpha}} \left(\frac{n}{\log n}\right)^{-\frac{1}{d+2\alpha}},$$

where $c_1 \equiv c_1(C_2, f_{\max}, \delta_1, d, \alpha) > 0$ and $c_2 \equiv c_2(C_1, C_2, f_{\max}, \delta_1, d, \alpha) > 0$. Hence the automatically chosen resolution behaves as desired for $\alpha \geq 0$.

To arrive at the result of Theorem 2 for $\alpha \geq 0$, follow the same arguments as before but using Lemma 4 to bound the Hausdorff error for $\alpha > 0$, and (2.22) to bound the Hausdorff error for $\alpha = 0$. Thus, Theorem 2 holds and the proposed method is adaptive for all $\alpha \geq 0$ (including the jump case), using the modified vernier and penalty.

■

Chapter 3

Quantitative Analysis of Semi-Supervised Learning

Empirical evidence shows that in favorable situations semi-supervised learning (SSL) algorithms can capitalize on the abundance of *unlabeled* training data to improve the performance of a learning task, in the sense that fewer *labeled* training data are needed to achieve a target error bound. However, in other situations unlabeled data do not seem to help. Recent attempts at theoretically characterizing the situations in which unlabeled data can help have met with little success, and sometimes appear to conflict with each other and intuition. In this chapter, we attempt to bridge the gap between practice and theory of semi-supervised learning. We develop a rigorous framework for analyzing the situations in which unlabeled data can help and quantify the improvement possible using finite sample error bounds. We show that there are large classes of problems for which SSL can significantly outperform supervised learning, in finite sample regimes and sometimes also in terms of error convergence rates. Moreover, we also provide a characterization of the relative value of unlabeled and labeled data.

3.1 Introduction

Supervised learning involves learning a mapping from an input or feature space $\mathcal{X}$ to an output or label space $\mathcal{Y}$ given n labeled training examples $\{X_i, Y_i\}_{i=1}^n$, that are independent and identically distributed according to a joint probability law P_{XY}. Labeled training data can be expensive, time-consuming and difficult to obtain in many applications. For example, hand-written character or speech recognition and document classification require

an experienced human annotator, or in some applications each label might be the outcome of a specially designed experiment. Semi-supervised learning (SSL) aims to capitalize on the abundance of unlabeled training data to improve learning performance. A thorough survey of semi-supervised learning literature is available in [58]. Empirical evidence suggests that in certain favorable situations unlabeled data can help, while in other situations it does not. Recent attempts at developing a theoretical basis for semi-supervised learning have been mostly pessimistic [26, 27, 40], and only provide a partial and sometimes apparently conflicting ([27] vs. [41]) explanations of whether or not, and to what extent, unlabeled data can help in learning. In this chapter, we develop a minimax framework to identify situations in which unlabeled data help to improve learning and quantify the value of unlabeled data using finite sample error bounds.

Two common supervised learning tasks are classification and regression. In binary classification, the target function or optimal mapping that minimizes the probability of error is given as $f^*(x) = \mathbf{1}_{\{P_{Y|X}(Y=1|X=x) \geq P_{Y|X}(Y=0|X=x)\}}$, and in regression, the target function or optimal mapping that minimizes the mean square error corresponds to the conditional mean and is given as $f^*(x) = \mathbb{E}_{Y|X}[Y|X = x]$. Notice that the target function or optimal mapping f^* for both classification and regression only depends on the conditional distribution of the label Y given the input features X. Since unlabeled data can only provide information about the marginal distribution, unlabeled data can only be expected to help in supervised learning situations where the marginal distribution of the features P_X provides some information about the conditional distribution $P_{Y|X}$. In other words, there exists a *link* between the marginal and conditional distributions. Two common collections of linked distributions considered in the literature for which semi-supervised learning is expected to yield promising results are the high density smoothness or cluster assumption [26–28] and the manifold assumption [27, 41]. The former assumes the link that the target function is locally smooth/regular over high (marginal) density regions of the feature space (but may not be globally smooth), whereas the latter assumes that the target function lies on a low-dimensional manifold and is smooth with respect to the geodesic distance on the manifold.

In the cluster case, knowledge of the high density regions or clusters reduces the problem of estimating an inhomogeneous function to a homogeneous function, and in the manifold case, knowledge of the manifold reduces a high-dimensional problem to a low-dimensional problem. Thus, knowledge of these high density regions or the manifold (henceforth called *decision regions*) which can be gleaned from unlabeled data, can greatly simplify the learning task.

In this work, we focus on learning under the high density smoothness or cluster assumption. We formalize this assumption in the next section and go on to establish that there exist nonparametric classes of distributions, denoted $\mathcal{P}_{XY}$, for which the high density regions are discernable from unlabeled data. Moreover, we show that there exist *clairvoyant* supervised learners that, given perfect knowledge of the decision regions denoted by $\mathcal{D}$, can significantly outperform any generic supervised learner f_n based on the n labeled samples in these classes. That is, if $\mathcal{R}$ denotes a risk of interest, $\widehat{f}_{\mathcal{D},n}$ denotes the clairvoyant supervised learner, and $\mathbb{E}$ denotes expectation with respect to training data, then $\sup_{\mathcal{P}_{XY}} \mathbb{E}[\mathcal{R}(\widehat{f}_{\mathcal{D},n})] < \inf_{f_n} \sup_{\mathcal{P}_{XY}} \mathbb{E}[\mathcal{R}(f_n)]$. This would imply that knowledge of the decision regions simplifies the supervised learning task. Based on this, we establish that there also exist semi-supervised learners, denoted $\widehat{f}_{m,n}$, that use m unlabeled examples in addition to the n labeled examples in order to estimate the decision regions, which perform as well as $\widehat{f}_{\mathcal{D},n}$, provided that m grows appropriately relative to n. Specifically, if the error bound for $\widehat{f}_{\mathcal{D},n}$ decays polynomially (exponentially) in n, then the number of unlabeled data m needs to grow polynomially (exponentially) with the number of labeled data n. We provide general results for a broad range of learning problems using finite sample error bounds. Then we consider regression problems in detail, and examine a concrete instantiation of these general results by deriving a minimax lower bound on the performance of any supervised learner and compare that to upper bounds on the errors of $\widehat{f}_{\mathcal{D},n}$ and $\widehat{f}_{m,n}$.

In their seminal papers, Castelli and Cover [59, 60] had suggested that, in the binary classification setting, the marginal distribution can be viewed as a mixture of class conditional

distributions:

$$P_X(x) = aP(x|Y=1) + (1-a)P(x|Y=0),$$

where $a = P(Y=1)$. If this mixture is identifiable, that is, learning P_X is sufficient to resolve the component distributions, then the classification problem reduces to a simple hypothesis testing problem of deciding the label (0/1) for each component. For hypothesis testing problems, the error converges exponentially fast in the number of labeled examples, whereas the error convergence is typically polynomial for classification. The ideas in this chapter are similar, except that we do not require identifiability of the mixture component densities, and show that it suffices to only approximately learn the decision regions over which the label is smooth. More recent attempts at theoretically characterizing SSL have been relatively pessimistic. Rigollet [26] establishes that for a fixed collection of distributions satisfying a cluster assumption, unlabeled data do not provide an improvement in convergence rate. A similar argument was made by Lafferty and Wasserman [27], based on the work of Bickel and Li [61], for the manifold case. However, in a recent paper, Niyogi [41] gives a constructive example of a class of distributions supported on a manifold whose complexity increases with the number of labeled examples, and he shows a lower bound of $\Omega(1)$ for any supervised learner (that is, the error of any supervised learner is bounded from below by a constant), whereas there exists a semi-supervised learner that can provide an error bound of $O(n^{-1/2})$, assuming infinite unlabeled data. We bridge the gap between these seemingly conflicting views. Our arguments can be understood by the simple example shown in Fig. 3.1, where the distribution is supported on two components separated by a margin ξ and the target function is smooth over each component. Given a finite sample of data, the high density regions may or may not be discernable depending on the sampling density (see Fig. 3.1(b), (c)). If ξ is fixed (this is similar to fixing the class of cluster-based distributions in [26] or the manifold in [27, 61]), then given enough labeled data a supervised learner can achieve optimal performance (since, eventually, it operates in regime (c) of Fig. 3.1) and unlabeled data may not help. Thus, in this example, there is no improvement due to unlabeled data in terms of the rate of error convergence for a fixed collection of distributions. However,

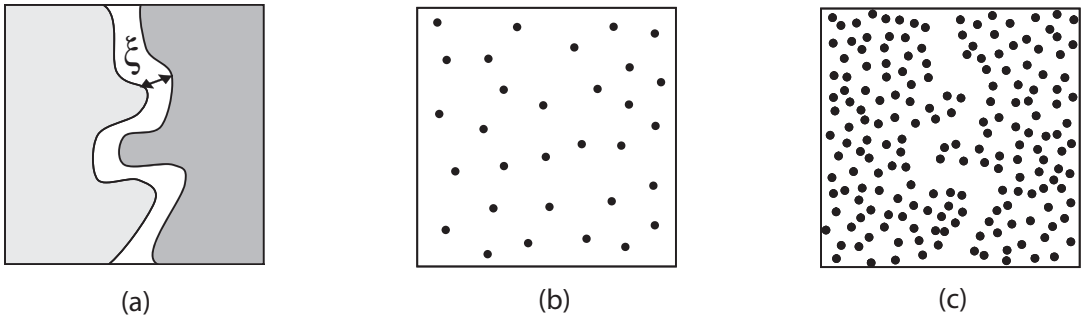


Figure 3.1 (a) Two separated high density regions with different labels that (b) cannot be discerned if the sample size is too small, but (c) can be estimated if sample density is high enough.

since the underlying true separation between the components is unknown, given a finite sample of data, there always exists a distribution for which the high density regions are indiscernible (e.g., $\xi \to 0$). This perspective is similar in spirit to the argument in [41]. We claim that meaningful characterizations of SSL performance and quantifications of the value of unlabeled data require finite sample error bounds, and that rates of convergence and asymptotic analysis may not capture the distinctions between SSL and supervised learning. Simply stated, if the high density regions are discernable from a finite sample size m of unlabeled data but not from a finite sample size $n < m$ of labeled data, then SSL can provide better performance than supervised learning. Further, we also show that there are certain plausible situations in which SSL yields rates of convergence that cannot be achieved by any supervised learner.

The rest of this chapter is organized as follows. In the next section, we describe a mathematical model for the cluster assumption. Section 3.3 describes a procedure for learning the high density regions using unlabeled data. Our main result characterizing the relative performance of supervised and semi-supervised learning is presented in Section 3.4, and Section 3.5 applies the result to the regression problem. Conclusions are discussed in Section 3.6, and proofs are deferred to Section 3.7.

3.2 Characterization of model distributions under the cluster assumption

In this section, we describe a mathematical model for the high density smoothness or cluster assumption. We define the collection of joint distributions $\mathcal{P}_{XY}(\xi) = \mathcal{P}_X \times \mathcal{P}_{Y|X}$ indexed by a margin parameter ξ as follows. Let X, Y be bounded random variables with marginal density $p(x) \in \mathcal{P}_X$ and conditional label distribution $p(Y|X = x) \in \mathcal{P}_{Y|X}$, supported on the domain $\mathcal{X} = [0, 1]^d$.

The marginal density $p(x) = \sum_{i=1}^{I} a_i p_i(x)$ is the mixture of a finite, but unknown, number of component densities $\{p_i\}_{i=1}^{I}$, where $I < \infty$. Here the unknown mixing proportions $a_i \geq a > 0$ and $\sum_{i=1}^{I} a_i = 1$. In addition, we place the following assumptions on the mixture component densities $\{p_i\}_{i=1}^{I}$:

1. p_i is supported on a unique compact, connected set $C_i \subseteq \mathcal{X}$ with Lipschitz boundaries. Specifically, we assume the following form for the component support sets:

$$C_i = \{x \equiv (x_1, \ldots, x_d) \in \mathcal{X} : g_i^{(1)}(x_1, \ldots, x_{d-1}) \leq x_d \leq g_i^{(2)}(x_1, \ldots, x_{d-1})\},$$

   where $g_i^{(1)}(\cdot), g_i^{(2)}(\cdot)$ are $d-1$ dimensional Lipschitz boundary functions with Lipschitz constant L.

   This form is a slight generalization of the boundary fragment class of sets considered in literature on set estimation and image reconstruction [34], and is used only for ease of analysis. The insights developed in this chapter apply to more generic formulations of the cluster assumption, however that would require a more sophisticated and detailed analysis that would not add much in terms of understanding the distinctions between supervised and semi-supervised learning.

2. p_i is bounded from above and below, $0 < b \leq p_i \leq B$.

3. p_i is Hölder-α_1 smooth on C_i with Hölder constant κ_1. Formally, p_i has continuous partial derivatives of up to order $[\alpha_1]$, where $[\alpha_1]$ denotes the maximal integer that is

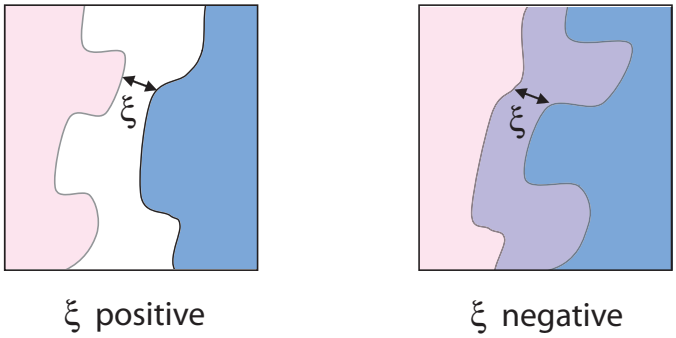


Figure 3.2 The margin ξ measures the minimal width of a decision region, or separation between support sets of the marginal mixture component densities. The margin is positive if there is no overlap between the component support sets, and negative otherwise.

$< \alpha_1$, and $\exists \delta > 0$ such that

$$\forall z, x \in C_i : ||z - x|| \leq \delta \Rightarrow |p_i(z) - TP_x(z, [\alpha_1])| \leq \kappa_1 \|z - x\|^{\alpha_1}$$

where $\kappa_1, \alpha_1 > 0$, $TP_x(\cdot, [\alpha_1])$ denotes the degree $[\alpha_1]$ Taylor polynomial approximation of p_i expanded around x, and $|| \cdot ||$ denotes Euclidean norm.

Let the conditional label density on each component C_i be denoted by $p_i(Y|X = x)$. Thus, a labeled training point (X, Y) is obtained as follows. With probability a_i, X is drawn from p_i and Y is drawn from $p_i(Y|X = x)$. In the supervised setting, we assume access to n labeled training data $\mathcal{L} = \{X_i, Y_i\}_{i=1}^n$ drawn i.i.d according to $P_{XY} \in \mathcal{P}_{XY}(\xi)$, and in the semi-supervised setting, we assume access to m additional unlabeled training data $\mathcal{U} = \{X_i\}_{i=1}^m$ drawn i.i.d according to $P_X \in \mathcal{P}_X$.

The characterization presented above implies that the overall marginal density $p(x)$ jumps at the cluster boundaries, since the component densities are bounded from below, and is smooth over the regions that are obtained as intersections of $\{C_i\}_{i=1}^I$ or their complements $\{C_i^c\}_{i=1}^I$. Let this collection of sets be denoted by $\mathcal{D}$, excluding the set $\cap_{i=1}^I C_i^c$ that does not lie in the support of the marginal density. That is, a set $D \in \mathcal{D}$ is of the form $d_1 \cap d_2 \cap \cdots \cap d_I$ where $d_i \in \{C_i, C_i^c\}$, but not including $\cap_{i=1}^I C_i^c$. Observe that $|\mathcal{D}| \leq 2^I$, and in practical situations the cardinality of $\mathcal{D}$ is much smaller as only a few of the sets are non-empty.

The cluster assumption is that the overall target function is smooth or constant on each of the regions $D \in \mathcal{D}$, except possibly the decision boundaries, hence the sets in $\mathcal{D}$ are called *decision sets*. At this point, we do not consider a specific target function; in Section 3.5, we will specify the smoothness assumptions on the target function in the regression setting.

The collection $\mathcal{P}_{XY}$ is indexed by a margin parameter ξ, which denotes the minimum width of a decision region or separation between the components C_i. The margin ξ is assigned a positive sign if there is no overlap between components, otherwise it is assigned a negative sign as illustrated in Figure 3.2. Formally, for $i, j \in \{1, \ldots, I\}$, let

$$\begin{aligned} d_{ij} &:= \min_{p,q \in \{1,2\}} \|g_i^{(p)} - g_j^{(q)}\|_\infty \qquad\qquad i \neq j, \\ d_{ii} &:= \|g_i^{(1)} - g_i^{(2)}\|_\infty, \end{aligned}$$

where $\|\cdot\|_\infty$ denotes the sup-norm, and

$$\text{sgn} = \begin{cases} 1 & \text{if } C_i \cap C_j = \emptyset \ \forall i \neq j, \text{ where } i, j \in \{1, \ldots, I\} \\ -1 & \text{otherwise} \end{cases}$$

Then the margin is defined as

$$\xi = \text{sgn} \cdot \min_{i,j \in \{1,\ldots,I\}} d_{ij}.$$

3.3 Learning Decision Regions

Ideally, we would like to break a given learning task into separate subproblems on each $D \in \mathcal{D}$, since the cluster assumption is that the target function is smooth on each decision region. In the section, we show that the decision regions are learnable using unlabeled data. Since the marginal density p is smooth within each decision region $D \in \mathcal{D}$, but exhibits jumps at the decision boundaries, the collection $\mathcal{D}$ can be learnt by estimating the marginal density from unlabeled data as follows:

1) *Marginal density estimation* — The procedure is based on the sup-norm kernel density estimator proposed in [62]. Consider a uniform square grid over the domain $\mathcal{X} = [0, 1]^d$ with spacing $2h_m$, where $h_m = \kappa_0 \left((\log m)^2/m\right)^{1/d}$ and $\kappa_0 > 0$ is a constant. For any point $x \in \mathcal{X}$, let $\bar{x}$ denote the closest point on the grid. Let K denote the kernel and $H_m = h_m \mathbf{I}$, then the

estimator of $p(x)$ is

$$\widehat{p}(x) = \frac{1}{mh_m^d} \sum_{i=1}^{m} K(H_m^{-1}(X_i - \bar{x})).$$

2) *Decision region estimation* — Two points $x_1, x_2 \in \mathcal{X}$ are said to be *connected*, denoted by $x_1 \leftrightarrow x_2$, if there exists a sequence of points $x_1 = z_1, z_2, \ldots, z_{k-1}, z_k = x_2$ such that $z_2, \ldots, z_{k-1} \in \mathcal{U}$, $\|z_j - z_{j+1}\| \leq 2\sqrt{d}h_m$. That is, there exists a sequence of $2\sqrt{d}h_m$-dense unlabeled data points between x_1 and x_2. Two points $x_1, x_2 \in \mathcal{X}$ are said to be *p-connected* if in addition to being connected, the sequence is such that for all points that satisfy $\|z_i - z_j\| \leq h_m \log m$, $|\widehat{p}(z_i) - \widehat{p}(z_j)| \leq \delta_m := (\log m)^{-1/3}$. That is, there exists a sequence of $2\sqrt{d}h_m$-dense unlabeled data points between x_1 and x_2 such that the marginal density varies smoothly along the sequence. All points that are pairwise p-connected specify an empirical decision region. This region estimation procedure is similar in spirit to the semi-supervised learning algorithm proposed in [63]. In practice, p-connectedness only need to be evaluated for the test point X and the training points with labels, that is $\{X_i\}_{i=1}^n \in \mathcal{L}$.

The following lemma shows that if the margin $|\xi|$ is large relative to the average spacing between unlabeled data points ($m^{-1/d}$), then with high probability, two points are p-connected (lie in the same empirical decision region) if and only if they lie in the same decision region $D \in \mathcal{D}$, provided the points are not too close to the decision boundaries.

Lemma 8. *Denote the set of boundary points as*

$$\mathcal{B} := \{z : z_d = g_i^{(p)}(z_1, \ldots, z_{d-1}), i \in \{1, \ldots, I\}, p \in \{1, 2\}\}$$

and define the boundary region as

$$\mathcal{R}_{\mathcal{B}} := \{x : \inf_{z \in \mathcal{B}} \|x - z\| \leq 2\sqrt{d}h_m\}.$$

If $|\xi| > C_o(m/(\log m)^2)^{-1/d}$, where $C_o = 6\sqrt{d}\kappa_0$, then for all $p \in \mathcal{P}_X$, all pairs of points $x_1, x_2 \in supp(p) \setminus \mathcal{R}_{\mathcal{B}}$ and all $D \in \mathcal{D}$, with probability $> 1 - 1/m$,

$$x_1 \overset{p}{\leftrightarrow} x_2 \quad \textit{if and only if} \quad x_1, x_2 \in D,$$

for large enough $m \geq m_0 \equiv m_0(p_{\min}, I, \kappa_1, d, \alpha_1, B, K, \kappa_0)$.[1]

[1]Dependence of a constant on K implies the constant depends on a norm or moment of the kernel K.

The proof is given in Section 3.7.1.

Remark: If we are only concerned with distributions that have a positive margin, then only connectedness is needed to identify the decision regions. In fact, for the positive margin case, the decision regions correspond to connected components of the support set, and Hausdorff accurate support set estimation proposed in Chapter 2 (also see [64]) can be used to estimate the decision regions instead of identifying connecting sequences. One advantage of using Hausdorff accurate support set estimation over connecting sequences is that we can also handle densities that do not jump (are not bounded away from zero) but transition gradually to zero. However, in the negative margin case p-connectedness is needed since the supports of the mixture constituents in this case are overlapping and the decision regions are characterized by a sharp transition in the density.

3.4 SSL Performance and the Value of Unlabeled Data

We now state our main result that characterizes the performance of SSL relative to a clairvoyant supervised learner (with perfect knowledge of the decision regions), and follows as a corollary to the lemma stated above. Let $\mathcal{R}(f)$ denote a risk of interest for a learner f and the excess risk $\mathcal{E}(f) = \mathcal{R}(f) - \mathcal{R}^*$, where $\mathcal{R}^*$ is the infimum risk over all possible learners. The risk is given by the probability of error $P_{XY}(f(X) \neq Y)$ for classification and the mean square error $\mathbb{E}_{XY}[(f(X) - Y)^2]$ for regression.

Corollary 5. *Assume that the excess risk $\mathcal{E}$ is bounded. Suppose there exists a clairvoyant supervised learner $\widehat{f}_{\mathcal{D},n}$, with perfect knowledge of the decision regions $\mathcal{D}$, for which the following finite sample upper bound holds*

$$\sup_{\mathcal{P}_{XY}(\xi)} \mathbb{E}[\mathcal{E}(\widehat{f}_{\mathcal{D},n})] \;\leq\; \epsilon_2(n).$$

Then there exists a semi-supervised learner $\widehat{f}_{m,n}$ such that if $|\xi| > C_o(m/(\log m)^2)^{-1/d}$, then

$$\sup_{\mathcal{P}_{XY}(\xi)} \mathbb{E}[\mathcal{E}(\widehat{f}_{m,n})] \;\leq\; \epsilon_2(n) + O\left(\frac{1}{m} + n\left(\frac{m}{(\log m)^2}\right)^{-1/d}\right).$$

The proof is given in Section 3.7.2. This result captures the essence of the relative characterization of semi-supervised and supervised learning for the margin based model distributions. It suggests that if the regions $\mathcal{D}$ are discernable using unlabeled data (the margin is large enough compared to average spacing between unlabeled data points), then there exists a semi-supervised learner that can perform as well as a supervised learner with clairvoyant knowledge of the decision regions, provided $m \gg n$, so that $(n/\epsilon_2(n))^d = O(m/(\log m)^2)$ and the additonal term in the performance bound of the semi-supervised learner is small compared to $\epsilon_2(n)$. This implies that if $\epsilon_2(n)$ decays polynomially (exponentially) in n, then m needs to grow polynomially (exponentially) in n.

Further, suppose that the following finite sample lower bound holds for any supervised learner based on n labeled data:

$$\inf_{f_n} \sup_{\mathcal{P}_{XY}(\xi)} \mathbb{E}[\mathcal{E}(f_n)] \geq \epsilon_1(n).$$

If $\epsilon_2(n) < \epsilon_1(n)$, then there exists a clairvoyant supervised learner with perfect knowledge of the decision regions that outperforms any supervised learner that does not have this knowledge. Hence, Corollary 5 implies that SSL can provide better performance than any supervised learner provided (i) $m \gg n$ so that $(n/\epsilon_2(n))^d = O(m/(\log m)^2)$, and (ii) knowledge of the decision regions simplifies the supervised learning task, so that $\epsilon_2(n) < \epsilon_1(n)$. In the next section, we provide a concrete application of this result in the regression setting. As a simple example in the binary classification setting, if $p(x)$ is supported on two disjoint sets and if $P(Y = 1|X = x)$ is strictly greater than $1/2$ on one set and strictly less than $1/2$ on the other (that is, the label is constant on each set), then perfect knowledge of the decision regions reduces the problem to a hypothesis testing problem for which $\epsilon_2(n) = O(e^{-Cn})$, for some constant $C > 0$. However, if ξ is small relative to the average spacing $n^{-1/d}$ between labeled data points, then $\epsilon_1(n) = c\, n^{-1/d}$ where $c > 0$ is a constant. This is because in this case the decision region boundaries can only be localized to an accuracy of $n^{-1/d}$, the average spacing between labeled data points. Since the boundaries are Lipschitz, the expected volume that is incorrectly assigned to any decision region is greater than $c\, n^{-1/d}$,

where $c > 0$ is a constant. This implies that the overall expected excess risk is greater than $c\, n^{-1/d}$. A formal proof for the lower bound can be derived along the lines of the minimax lower bound proof for regression in the next section. Thus, an exponential improvement is possible using semi-supervised learning provided the number of unlabeled data examples m grows exponentially in n, the number of labeled data examples. In other words, to obtain the same performance bound as a supervised learner with n labeled examples, a semi-supervised learner only needs $n' \equiv \log n$ labeled examples in the binary classification setting, and the number m of unlabeled examples needed is exponential in n', that is, polynomial in n.

3.5 Density-adaptive Regression

Let Y denote a continuous and bounded random variable. Under squared error loss, the optimal decision rule $f^*(x) = \mathbb{E}[Y|X = x]$, and the excess risk $\mathcal{E}(f) = \mathbb{E}[(f(X) - f^*(X))^2]$. Recall that $p_i(Y|X = x)$ is the conditional density on the i-th component and let $\mathbb{E}_i$ denote expectation with respect to the corresponding conditional distribution. The optimal regression function on each component is $f_i(x) = \mathbb{E}_i[Y|X = x]$ and we assume that for $i = 1, \ldots, I$

1. f_i is uniformly bounded, $|f_i| \leq M$.

2. f_i is Hölder-α_2 smooth on C_i with Hölder constant κ_2.

This implies that the overall regression function $f^*(x)$, given as

$$f^*(x) = \sum_{i=1}^{I} \frac{a_i p_i(x)}{\sum_{j=1}^{I} a_j p_j(x)} f_i(x),$$

is piecewise Hölder-α smooth, where $\alpha = \min(\alpha_1, \alpha_2)$. That is, f^* is Hölder-α smooth on each $D \in \mathcal{D}$, except possibly at the decision boundaries. Since a Hölder-α smooth function can be locally well-approximated by a Taylor polynomial, we propose the following semi-supervised learner that performs local polynomial fits within each empirical decision region, that is, using labeled training data that are p-connected as per the definition in Section 3.3. While a spatially uniform estimator suffices to estimate a Hölder-α smooth function, we use

the following spatially adaptive estimator proposed in Section 4.1 of [5] which is shown to yield minimax optimal performance for piecewise-smooth functions. This ensures that when the decision regions are indiscernible using unlabeled data, the semi-supervised learner still achieves an error bound that is, up to logarithmic factors, no worse than the minimax lower bound for supervised learners.

$$\widehat{f}_{m,n}(X) = \widehat{f}_X(X)$$

where

$$\widehat{f}_x(\cdot) = \arg\min_{f' \in \Gamma} \sum_{i=1}^{n} (Y_i - f'(X_i))^2 \mathbf{1}_{x \overset{p}{\leftrightarrow} X_i} + \mathrm{pen}(f').$$

Here Γ denotes a collection of piecewise polynomials with quantized coefficients of degree $[\alpha]$ (the maximal integer $< \alpha$), defined over recursive dyadic partitions of the domain $\mathcal{X} = [0,1]^d$ with cells of sidelength between $2^{-\lceil \log(n/\log n)/(2\alpha+d) \rceil}$ and $2^{-\lceil \log(n/\log n)/d \rceil}$ (see [5] for details). The penalty term $\mathrm{pen}(f')$ is proportional to $\log(\sum_{i=1}^{n} \mathbf{1}_{x \overset{p}{\leftrightarrow} X_i}) \cdot \# f'$, where $\sum_{i=1}^{n} \mathbf{1}_{x \overset{p}{\leftrightarrow} X_i}$ simply denotes the number of labeled training data that are p-connected to x, that is are in the same empirical decision region as x, and $\# f'$ denotes the number of cells in the recursive dyadic partition on which f' is defined. It is shown in [5] that, under the Hölder-α assumption, this estimator obeys a finite sample error bound of $n^{-2\alpha/(2\alpha+d)}$, ignoring a logarithmic factor. Also, it is shown that for piecewise Hölder-α smooth functions, this estimator yields a finite sample error bound of $\max(n^{-2\alpha/(2\alpha+d)}, n^{-1/d})$, ignoring a logarithmic factor.

Using these results from [5] and Corollary 5, in Section 3.7.3, we derive finite sample upper bounds on the mean square excess risk of the semi-supervised learner (SSL) described above. Also, we derive finite sample minimax lower bounds on the performance of any supervised learner (SL) based on n labeled examples in Section 3.7.4. Our results are summarized in Table 3.1, for model distributions characterized by various values of the margin parameter ξ. In the table, we suppress constants and log factors in the error bounds, and assume that $m \gg n^{2d}$ so that the performance bound on the semi-supervised learner given in Corollary 5 essentially scales as $\epsilon_2(n)$. The constants c_o and C_o characterizing the margin only depend on the fixed parameters of the class $\mathcal{P}_{XY}(\xi)$. Also, ξ_0 denotes a constant, and thus the cases

Margin range ξ	SSL upper bound $\epsilon_2(n)$	SL lower bound $\epsilon_1(n)$	SSL helps
$\xi \geq \xi_0$	$n^{-2\alpha/(2\alpha+d)}$	$n^{-2\alpha/(2\alpha+d)}$	No
$\xi \geq c_o n^{-1/d}$	$n^{-2\alpha/(2\alpha+d)}$	$n^{-2\alpha/(2\alpha+d)}$	No
$c_o n^{-1/d} > \xi \geq C_o(\frac{m}{(\log m)^2})^{-1/d}$	$n^{-2\alpha/(2\alpha+d)}$	$n^{-1/d}$	Yes
$C_o(\frac{m}{(\log m)^2})^{-1/d} > \xi \geq -C_o(\frac{m}{(\log m)^2})^{-1/d}$	$n^{-1/d}$	$n^{-1/d}$	No
$-C_o(\frac{m}{(\log m)^2})^{-1/d} > \xi$	$n^{-2\alpha/(2\alpha+d)}$	$n^{-1/d}$	Yes
$-\xi_0 > \xi$	$n^{-2\alpha/(2\alpha+d)}$	$n^{-1/d}$	Yes

Table 3.1 Comparison of finite sample lower bounds on the mean square error for supervised learning, with finite sample upper bounds on the mean square error for semi-supervised learning, for the margin based model distributions. These bounds hold for $m \gg n^{2d}$ and $d \geq 2\alpha/(2\alpha - 1)$, and suppress constants and log factors.

$\xi \geq \xi_0$ and $-\xi_0 > \xi$ correspond to considering a fixed collection of distributions (whose complexity does not change with the amount of data).

Consider the case when the dimension is large or the target function is smooth enough so that $d \geq 2\alpha/(2\alpha - 1)$. If $d < 2\alpha/(2\alpha - 1)$, then the supervised learning error incurred by averaging across decision regions (which behaves like $n^{-1/d}$) is smaller than error incurred in estimating the target function away from the boundaries (which behaves like $n^{-2\alpha/(2\alpha+d)}$). Thus, when $d < 2\alpha/(2\alpha - 1)$, learning the decision regions does not simplify the supervised learning task, and there appears to be no benefit to using a semi-supervised learner. So focusing on the case when $d \geq 2\alpha/(2\alpha - 1)$, the results of Table 3.1 state that if the margin ξ is large relative to the average spacing between labeled data points $n^{-1/d}$, then a supervised learner can discern the decision regions accurately and SSL provides no gain. When $\xi \geq \xi_0$, we consider a fixed collection of distributions, and this argument is similar in spirit to the argument made by Lafferty and Wasserman [27]. However, if $\xi > 0$ is small relative to $n^{-1/d}$, but large with respect to the spacing between unlabeled data points $m^{-1/d}$,

then the proposed semi-supervised learner provides improved error bounds compared to *any* supervised learner. This is similar in spirit to the argument made by Niyogi [41] that the true underlying distribution can be more complex than can be discerned using labeled data. If $|\xi|$ is smaller than $m^{-1/d}$, the decision regions are not discernable even with unlabeled data and SSL provides no gain. However, notice that the performance of the semi-supervised learner is no worse than the minimax lower bound for supervised learners since we chose an estimator that is also optimal for piecewise smooth functions (recall that the overall target function is piecewise smooth). In the $\xi < 0$ case, when the component support sets can overlap, if the magnitude of the margin $|\xi|$ larger than $m^{-1/d}$, then the semi-supervised learner can discern the decision regions and achieves smaller error bounds ($n^{-2\alpha/(2\alpha+d)}$), whereas these regions cannot be as accurately discerned by any supervised learner. For the overlap case ($\xi < 0$), the supervised learners are always limited by the error incurred due to not resolving the decision regions ($n^{-1/d}$). In particular, for the fixed collection of distributions with $\xi < -\xi_0$, a faster rate of error convergence is attained by SSL compared to SL, provided $m \gg n^{2d}$.

3.6 Concluding Remarks

In this chapter, we develop a framework for evaluating the performance gains possible with semi-supervised learning under a cluster assumption using finite sample error bounds. The theoretical characterization we present explains why in certain situations unlabeled data can help to improve learning, while in other situations they may not. We demonstrate that there exist general situations under which semi-supervised learning can be significantly superior to supervised learning, in terms of achieving smaller finite sample error bounds than any supervised learner, and sometimes in terms of a better rate of error convergence. Moreover, our results also provide a quantification of the relative value of unlabeled to labeled data.

While we focus on the cluster assumption in this paper, we conjecture that similar techniques can be applied to quantify the performance of semi-supervised learning under the

manifold assumption as well. In the manifold case, the curvature and how close the manifold can get to itself or another manifold will play the role that the margin plays under the cluster assumption. In particular, we believe that the use of minimax lower bounding techniques is essential because many of the interesting distinctions between supervised and semi-supervised learning occur only in finite sample regimes, and rates of convergence and asymptotic analysis may not capture the complete picture.

In this work, we also show that though semi-supervised learning simplifies the learning task when the link relating the marginal and conditional distributions holds, it is possible to ensure that the performance of the semi-supervised learning does not deteriorate when the link is not discernable using unlabeled data or does not hold. For example, when the margin is small relative to the spacing between unlabeled data, the decision regions cannot be identified using unlabeled data, however by employing a more sophisticated tool (a learner that has optimal performance for piecewise smooth functions) similar to what a supervised learning algorithm would use, we ensured that the SSL performance is no worse than what a supervised learner would achieve. In this sense, the semi-supervised learner we propose is somewhat agnostic. However, if the number of decision regions can grow with n, the semi-supervised learning algorithm can perform worse because it would break the problem into a large collection of subproblems. Thus, it is of interest to develop an agnostic procedure that can identify such situations and switch from a semi-supervised learner to a supervised learner, for example using cross-validation, to determine the learner with smaller risk. We elaborate on this idea some more in Chapter 6.

3.7 Proofs

Since the component densities are bounded from below and above, define $p_{\min} := b \min_i a_i \leq p(x) \leq B =: p_{\max}$.

3.7.1 Proof of Lemma 8

We present the proof in two steps - first, we establish some results about the proposed kernel density estimator, and then using the density estimation results, we establish that the decision regions $\mathcal{D}$ can be learnt based only on unlabeled data.

1) Density estimation:

Theorem 5. ***[Sup-norm density estimation in non-boundary regions]*** *Consider the kernel density estimator proposed in Section 3.3* $\widehat{p}(x) = \frac{1}{mh_m^d}\sum_{i=1}^m K(H_m^{-1}(X_i - \bar{x}))$, *where* $H_m = h_m\mathbf{I}$, $h_m = \kappa_0((\log m)^2/m)^{1/d}$, $\kappa_0 > 0$ *is a constant, and* $\bar{x}$ *denotes the point closest to* x *on a uniform grid over the domain* $\mathcal{X} = [0,1]^d$ *with spacing* $2h_m$. *Let the kernel* K *satisfy*

$$supp(K) = [-1,1]^d, K \in (0, K_{\max}] \text{ and } \int_{[-1,1]^d} u^j K(u)du = \begin{cases} 1 & j = 0 \\ 0 & 1 \le j \le [\alpha_1] \end{cases},$$

where $supp(\cdot)$ *denotes the support of a function, then for all* $p \in \mathcal{P}_X$, *with probability at least* $1 - 1/m$,

$$\sup_{x \in supp(p)\setminus\mathcal{R}_\mathcal{B}} |p(x) - \widehat{p}(x)| \le c_3\left(h_m^{\min(1,\alpha_1)} + \sqrt{\frac{\log m}{mh_m^d}}\right) =: \epsilon_m,$$

for $m \ge m_1 \equiv m_1(K, B)$, *where* $c_3 \equiv c_3(I, \kappa_1, d, \alpha_1, B, K) > 0$ *is a constant. Notice that* ϵ_m *decreases with increasing* m.

Proof. Consider any $p \in \mathcal{P}_X$. Since $\widehat{p}(x) = \widehat{p}(\bar{x})$,

$$\sup_{x \in \text{supp}(p)\setminus\mathcal{R}_\mathcal{B}} |p(x) - \widehat{p}(x)| \le \sup_{x \in \text{supp}(p)\setminus\mathcal{R}_\mathcal{B}} |p(x) - p(\bar{x})| + \sup_{\bar{x}: x \in \text{supp}(p)\setminus\mathcal{R}_\mathcal{B}} |p(\bar{x}) - \widehat{p}(\bar{x})| \quad (3.1)$$

To bound the first term of (3.1), observe that since $x \in \text{supp}(p)\setminus\mathcal{R}_\mathcal{B}$ and $\|x - \bar{x}\| \le \sqrt{d}h_m$, by definition of $\mathcal{R}_\mathcal{B}$ if $x \in C_i$ then $\bar{x} \in C_i$ and vice versa. Thus, for all $x \in \text{supp}(p) \setminus \mathcal{R}_\mathcal{B}$,

$$\begin{aligned} |p(x) - p(\bar{x})| = \left|\sum_{i=1}^I a_i p_i(x) - a_i p_i(\bar{x})\right| &\le \sum_{i=1}^I a_i \left|p_i(x) - p_i(\bar{x})\right| \\ &= \sum_{i: x, \bar{x} \in C_i} a_i |p_i(x) - p_i(\bar{x})| \\ &\le \sum_{i: x, \bar{x} \in C_i} a_i \left(\kappa_1(\sqrt{d}h_m)^{\alpha_1} + \left|\sum_{j=1}^{[\alpha_1]} \frac{p_i^{(j)}(x)}{j!}(\bar{x} - x)^j\right|\right) \\ &\le c_1\, h_m^{\min(1,\alpha_1)}, \end{aligned}$$

where $c_1 \equiv c_1(I, \kappa_1, d, \alpha_1, B) > 0$ is a constant. The last step follows since if p_i is Hölder-α_1 smooth, then all its derivatives up to $[\alpha_1]$ are bounded and $\|x - \bar{x}\| \leq \sqrt{d} h_m$.

To bound the second term, notice that for all $\bar{x} : x \in \text{supp}(p) \setminus \mathcal{R}_{\mathcal{B}}$,

$$|p(\bar{x}) - \widehat{p}(\bar{x})| = |p(\bar{x}) - \mathbb{E}[\widehat{p}(\bar{x})]| + |\mathbb{E}[\widehat{p}(\bar{x})] - \widehat{p}(\bar{x})|$$

We now bound the two terms in the last expression.

1. For all $\bar{x} : x \in \text{supp}(p) \setminus \mathcal{R}_{\mathcal{B}}$, consider

$$|p(\bar{x}) - \mathbb{E}[\widehat{p}(\bar{x})]| = \left| p(\bar{x}) - \frac{1}{h_m^d} \int_{\bar{x}-h_m}^{\bar{x}+h_m} K(H_m^{-1}(y - \bar{x})) p(y) dy \right|$$

Notice that given the conditions on the kernel,

$$\begin{aligned} p(\bar{x}) &= \int_{[-1,1]^d} \sum_{j=0}^{[\alpha_1]} \frac{p^{(j)}(\bar{x})}{j!} (h_m u)^j K(u) du \\ &= \frac{1}{h_m^d} \int_{\bar{x}-h_m}^{\bar{x}+h_m} K(H_m^{-1}(y - \bar{x})) \sum_{j=0}^{[\alpha_1]} \frac{p^{(j)}(\bar{x})}{j!} (y - \bar{x})^j dy \end{aligned}$$

Therefore, we get

$$\begin{aligned} &|p(\bar{x}) - \mathbb{E}[\widehat{p}(\bar{x})]| \\ &= \left| \frac{1}{h_m^d} \int_{\bar{x}-h_m}^{\bar{x}+h_m} K(H_m^{-1}(y - \bar{x})) \left(\sum_{j=0}^{[\alpha_1]} \frac{p^{(j)}(\bar{x})}{j!} (y - \bar{x})^j dy - p(y) \right) dy \right| \\ &= \left| \frac{1}{h_m^d} \int_{\bar{x}-h_m}^{\bar{x}+h_m} K(H_m^{-1}(y - \bar{x})) \sum_{i=1}^{I} a_i \left(\sum_{j=0}^{[\alpha_1]} \frac{p_i^{(j)}(\bar{x})}{j!} (y - \bar{x})^j dy - p_i(y) \right) dy \right| \\ &\leq \left| \frac{1}{h_m^d} \int_{\bar{x}-h_m}^{\bar{x}+h_m} K(H_m^{-1}(y - \bar{x})) \kappa_1 \|y - \bar{x}\|^{\alpha_1} dy \right| \\ &\leq \kappa_1 \left(\int_{[-1,1]^d} \|u\|^{\alpha_1} K(u) du \right) h_m^{\alpha_1} = c_2 \, h_m^{\alpha_1}, \end{aligned}$$

where $c_2 \equiv c_2(\kappa_1, K, \alpha_1) > 0$ is a constant.

2. Now consider

$$\begin{aligned} P\left(\sup_{\bar{x}:x\in\text{supp}(p)\setminus\mathcal{R}_{\mathcal{B}}} |\mathbb{E}[\widehat{p}(\bar{x})] - \widehat{p}(\bar{x})| > \epsilon\right) &\leq \sum_{\bar{x}} P\left(|\mathbb{E}[\widehat{p}(\bar{x})] - \widehat{p}(\bar{x})| > \epsilon\right) \\ &= \sum_{\bar{x}} P\left(\left|\sum_{i=1}^{m} \mathbb{E}[Z_i] - Z_i\right| > mh_m^d\epsilon\right) \\ &\leq \sum_{\bar{x}} P\left(\sum_{i=1}^{m} |\mathbb{E}[Z_i] - Z_i| > mh_m^d\epsilon\right) \end{aligned}$$

where $Z_i = K(H_m^{-1}(X_i - \bar{x}))$. Now observe that $|\mathbb{E}[Z_i] - Z_i| \leq K_{\max}$ and

$$\begin{aligned} \text{var}(Z_i) \leq E[Z_i^2] &= \int_{\bar{x}-h_m}^{\bar{x}+h_m} K^2(H_m^{-1}(y-\bar{x}))p(y)dy \\ &= h_m^d \int_{[-1,1]^d} K^2(u)p(\bar{x} + H_m u)du \\ &= h_m^d \int_{[-1,1]^d} K^2(u)(p(\bar{x}) + o(1))du \\ &\leq 2\|K\|_2^2 p(\bar{x})h_m^d \leq 2\|K\|_2^2 B h_m^d \end{aligned}$$

Thus, using Bernstein's inequality, we get:

$$P\left(\sum_{i=1}^{m} |\mathbb{E}[Z_i] - Z_i| > mh_m^d\epsilon\right) \leq \exp\left\{-\frac{(mh_m^d\epsilon)^2/2}{2\|K\|_2^2 Bmh_m^d + K_{\max}mh_m^d\epsilon/3}\right\}$$

Setting $\epsilon = 4\|K\|_2\sqrt{B}\sqrt{\frac{\log m}{mh_m^d}}$, and observing that $K_{\max}\epsilon/3 \leq 2\|K\|_2^2 B$ for large enough $m \geq m_1 \equiv m_1(K, B)$, we get:

$$\begin{aligned} P\left(\sup_{\bar{x}:x\in\text{supp}(p)\setminus\mathcal{R}_{\mathcal{B}}} |\mathbb{E}[\widehat{p}(\bar{x})] - \widehat{p}(\bar{x})| > 4\|K\|_2\sqrt{B}\sqrt{\frac{\log m}{mh_m^d}}\right) & \\ \leq \sum_{\bar{x}} \exp\left\{-\frac{16\|K\|_2^2 Bmh_m^d \log m/2}{4\|K\|_2^2 Bmh_m^d}\right\} & \\ \leq h_m^{-d}\exp\{-2\log m\} & \\ \leq m \cdot \frac{1}{m^2} = \frac{1}{m} & \end{aligned}$$

Therefore we get, with probability at least $1 - 1/m$, for $m \geq m_1(K, B)$ we have the following bound on the second term

$$\sup_{\bar{x}:x\in\text{supp}(p)\setminus\mathcal{R}_{\mathcal{B}}} |p(\bar{x}) - \widehat{p}(\bar{x})| \leq c_2 h_m^{\alpha_1} + 4\|K\|_2\sqrt{B}\sqrt{\frac{\log m}{mh_m^d}}.$$

And putting the bounds on the two terms together: For all $p \in \mathcal{P}_X$, with probability at least $1 - 1/m$, for $m \geq m_1(K, B)$

$$\sup_{x \in \text{supp}(p) \setminus \mathcal{R}_\mathcal{B}} |p(x) - \widehat{p}(x)| \leq c_3 \left(h_m^{\min(1,\alpha_1)} + \sqrt{\frac{\log m}{m h_m^d}} \right),$$

where $c_3 \equiv c_3(I, \kappa_1, d, \alpha_1, B, K) > 0$ is a constant. □

Remark: This bound can be tightened to $O(h_m^{\alpha_1} + \sqrt{\log m / m h_m^d})$ by also estimating the density derivatives at the grid points and defining $p(x)$ as the Taylor polynomial approximation expanded around the closest grid point $\bar{x}$, see [62]. Also, the arguments of the proof hold if $h_m = \kappa_0 (\log m / m)^{-1/(d+2\alpha_1)}$. Hence, we recover the minimax rate of $O((m/\log m)^{-\alpha_1/(d+2\alpha_1)}$ for sup-norm density estimation of a Hölder-α_1 smooth density. However, we want to characterize the largest collection of distributions (smallest margin) that a semi-supervised learner can handle, and thus we seek the smallest h_m (which determines the smallest margin that can be handled) for which the bound ϵ_m decreases with increasing m.

Corollary 6. ***[Empirical density of unlabeled data]*** *Under the conditions of Theorem 5, for all $p \in \mathcal{P}_X$ and $m \geq m_3 \equiv m_3(p_{\min}, I, \kappa_1, d, \alpha_1, B, K, \kappa_0)$, with probability at least $1 - 1/m$, for all $x \in supp(p) \setminus \mathcal{R}_\mathcal{B}$, there exists an unlabeled data point $X_i \in \mathcal{U}$ such that $\|X_i - x\| \leq \sqrt{d} h_m$.*

Proof. From Theorem 5, for all $x \in \text{supp}(p) \setminus \mathcal{R}_\mathcal{B}$, for $m \geq m_1(K, B)$

$$\widehat{p}(x) \geq p(x) - \epsilon_m \geq p_{\min} - \epsilon_m > 0$$

The last step follows for large enough $m \geq m_2 \equiv m_2(p_{\min}, I, \kappa_1, d, \alpha_1, B, K, \kappa_0)$ since ϵ_m is decreasing with m. This implies that $\sum_{i=1}^m K(H_m^{-1}(X_i - x)) > 0$ for $m \geq m_3 = \max(m_1, m_2)$, and therefore there exists an unlabeled data point within $\sqrt{d} h_m$ of x. □

2) Decision region estimation - Using the density estimation results, we now show that if $|\xi| > 6\sqrt{d} h_m$, then for all $p \in \mathcal{P}_X$, all pairs of points $x_1, x_2 \in \text{supp}(p) \setminus \mathcal{R}_\mathcal{B}$ and all $D \in \mathcal{D}$, for $m \geq m_0 \equiv m_0(p_{\min}, I, \kappa_1, d, \alpha_1, B, K, \kappa_0)$ with probability $> 1 - 1/m$, we have $x_1 \overset{p}{\leftrightarrow} x_2$ if and only if $x_1, x_2 \in D$. We establish this in two steps:

1. $x_1 \in D, x_2 \notin D \Rightarrow x_1 \overset{p}{\not\leftrightarrow} x_2$:

   Since x_1 and x_2 belong to different decision regions and $x_1, x_2 \in \text{supp}(p) \setminus \mathcal{R}_{\mathcal{B}}$, all sequences connecting x_1 and x_2 through unlabeled data points pass through a region where either (i) the density is zero, or (ii) the density is positive. In case (i), there cannot exist a sequence connecting x_1 and x_2 through unlabeled data points such that for any two consecutive points z_j, z_{j+1} along the sequence $\|z_j - z_{j+1}\| \leq 2\sqrt{d}h_m$ since the region of zero density is at least $|\xi| > 6\sqrt{d}h_m$ wide. Therefore, $x_1 \not\leftrightarrow x_2$, and hence $x_1 \overset{p}{\not\leftrightarrow} x_2$. In case (ii), since x_1 and x_2 belong to different decision regions, the marginal density $p(x)$ jumps by at least $p_{\min}$ one or more times along all sequences connecting x_1 and x_2. Suppose the first jump (in the sequence) occurs where decision region D ends and another decision region $D' \neq D$ begins. Then since D, D' are at least $|\xi| > 6\sqrt{d}h_m$ wide, by Corollary 6 with probability $> 1 - 1/m$ for $m \geq m_3$, for all sequences connecting x_1 and x_2 through unlabeled data points, there exist points z, z' in the sequence that lie in $D \setminus \mathcal{R}_{\mathcal{B}}$, $D' \setminus \mathcal{R}_{\mathcal{B}}$, respectively, and $\|z - z'\| \leq h_m \log m$. We will show that $|p(z) - p(z')| \geq p_{\min} - O((h_m \log m)^{\min(1,\alpha_1)})$ which using Theorem 5 implies that $|\widehat{p}(z) - \widehat{p}(z')| \geq p_{\min} - O((h_m \log m)^{\min(1,\alpha_1)}) - 2\epsilon_m > \delta_m$ for m large enough. Hence $x_1 \overset{p}{\not\leftrightarrow} x_2$.

   To see these claims, observe that since D' and D are adjacent decision regions, if $D' = (d'_1, d'_2, \ldots, d'_I)$ and $D = (d_1, d_2, \ldots, d_I)$, then $\exists i_0$ such that $d_i = d'_i$ for all $i \neq i_0$. Thus, $\{i : z \in C_i \text{ or } z' \in C_i\} = i_0$. Since $\|z - z'\| \leq h_m \log m$, we get:

$$\begin{aligned}
|p(z) - p(z')| &= \left| \sum_{i=1}^{I} a_i p_i(z) - \sum_{i=1}^{I} a_i p_i(z') \right| \\
&= \left| \sum_{i: z \in C_i \text{ or } z' \in C_i} a_i (p_i(z) - p_i(z')) + \sum_{i: z, z' \in C_i} a_i (p_i(z) - p_i(z')) \right| \\
&\geq |a_{i_0}(p_{i_0}(z) - p_{i_0}(z'))| - \left| \sum_{i: z, z' \in C_i} a_i (p_i(z) - p_i(z')) \right| \\
&\geq ab - \left| \sum_{i: z, z' \in C_i} a_i (p_i(z) - p_i(z')) \right|
\end{aligned}$$

$$\geq p_{\min} - c_4(h_m \log m)^{\min(1,\alpha_1)},$$

where $c_4 > 0$ is a constant. The fourth step follows since $d_{i_0} \neq d'_{i_0}$ and hence either $p_{i_0}(z)$ is zero or $p_{i_0}(z')$ is zero, and since p_{i_0} is bounded from below by b and $a_i \geq a$. To see the last step, recall that the component densities p_i are Hölder-α_1 smooth and $\|z' - z\| \leq h_m \log m$. Thus, we have:

$$\begin{aligned}
\left| \sum_{i:z,z' \in C_i} a_i (p_i(z) - p_i(z')) \right| &\leq \sum_{i:z,z' \in C_i} a_i |p_i(z) - p_i(z')| \\
&\leq \sum_{i:z,z' \in C_i} a_i \left(\kappa_1 (h_m \log m)^{\alpha_1} + \left| \sum_{j=0}^{[\alpha_1]} \frac{p_i^{(j)}(z)}{j!} (z' - z)^j \right| \right) \\
&\leq c_4 (h_m \log m)^{\min(1,\alpha_1)},
\end{aligned}$$

where $c_4 \equiv c_4(I, \kappa_1, \alpha_1, B) > 0$ is a constant. Here the last step follows since if p_i is Hölder-α_1 smooth, then all its derivatives up to $[\alpha_1]$ are bounded.

Now since $z, z' \in \text{supp}(p) \setminus \mathcal{R}_{\mathcal{B}}$, using Theorem 5, we get with probability $> 1 - 1/m$, for $m \geq \max(m_1, m_3)$

$$\begin{aligned}
|\widehat{p}(z) - \widehat{p}(z')| &= |\widehat{p}(z) - p(z) + p(z) - p(z') + p(z') - \widehat{p}(z')| \\
&\geq |p(z) - p(z')| - |\widehat{p}(z) - p(z)| - |p(z') - \widehat{p}(z')| \\
&\geq p_{\min} - c_4 (h_m \log m)^{\min(1,\alpha_1)} - 2\epsilon_m \\
&> \frac{1}{(\log m)^{1/3}} = \delta_m.
\end{aligned}$$

The last step holds for large enough $m \geq m_4 \equiv m_4(p_{\min}, I, \kappa_1, d, \alpha_1, B, K, \kappa_0)$. Thus, for case (ii) we have shown that,for $m \geq \max(m_1, m_3, m_4)$ with probability $> 1 - 1/m$, for all sequences connecting x_1 and x_2 through $2\sqrt{d}h_m$-dense unlabeled data points, there exists points z, z' in the sequence such that $\|z - z'\| \leq h_m \log m$ but $|\widehat{p}(z) - \widehat{p}(z')| > \delta_m$. Thus,

$$x_1 \in D, x_2 \notin D \Rightarrow x_1 \overset{p}{\not\leftrightarrow} x_2.$$

2. $x_1, x_2 \in D \Rightarrow x_1 \overset{p}{\leftrightarrow} x_2$:

Since D has width at least $|\xi| > 6\sqrt{d}h_m$, there exists a region of width $> 2\sqrt{d}h_m$

contained in $D \setminus \mathcal{R}_{\mathcal{B}}$, and Corollary 6 implies that for $m \geq m_3$, with probability $> 1 - 1/m$, there exist sequence(s) contained in $D \setminus \mathcal{R}_{\mathcal{B}}$ connecting x_1 and x_2 through $2\sqrt{d}h_m$-dense unlabeled data points. Since the sequence is contained in $D \setminus \mathcal{R}_{\mathcal{B}}$, and the density on D is Hölder-α_1 smooth, we have for all points z, z' in the sequence such that $\|z - z'\| \leq h_m \log m$,

$$\begin{aligned}
|\widehat{p}(z) - \widehat{p}(z')| &= |\widehat{p}(z) - p(z) + p(z) - p(z') + p(z') - \widehat{p}(z')| \\
&\leq |\widehat{p}(z) - p(z)| + |p(z) - p(z')| + |p(z') - \widehat{p}(z')| \\
&\leq 2\epsilon_m + |p(z) - p(z')| \\
&\leq 2\epsilon_m + c_5(h_m \log m)^{\min(1,\alpha_1)} \\
&\leq \frac{1}{(\log m)^{1/3}} = \delta_m,
\end{aligned}$$

where $c_5 > 0$ is a constant, and the last step holds for large enough $m \geq m_5 \equiv m_5(I, \kappa_1, d, \alpha_1, B, K, \kappa_0)$. The third step follow since $z, z' \in \text{supp}(p) \setminus \mathcal{R}_{\mathcal{B}}$, and invoking Theorem 5. To see the fourth step, since $z, z' \in D$, if $z \in C_i$ then $z' \in C_i$ and vice versa. Thus,

$$\begin{aligned}
|p(z) - p(z')| &= \left| \sum_{i: z, z' \in C_i} a_i (p_i(z) - p_i(z')) \right| \leq \sum_{i: z, z' \in C_i} a_i |p_i(z) - p_i(z')| \\
&\leq \sum_{i: z, z' \in C_i} a_i \left(\kappa_1 (h_m \log m)^{\alpha_1} + \left| \sum_{j=0}^{[\alpha_1]} \frac{p_i^{(j)}(z)}{j!} (z' - z)^j \right| \right) \\
&\leq c_5 (h_m \log m)^{\min(1,\alpha_1)},
\end{aligned}$$

where $c_5 \equiv c_5(\kappa_1, I, B, \alpha_1) > 0$ is a constant. Here the third step follows since $\|z' - z\| \leq h_m \log m$, and p_i is Hölder-α_1 on C_i. The last step follows since if p_i is Hölder-α_1 smooth, then all its derivatives up to $[\alpha_1]$ are bounded. Thus, we have shown that

$$x_1, x_2 \in D \Rightarrow x_1 \overset{p}{\leftrightarrow} x_2.$$

Thus, the result of the Lemma holds for $m \geq m_0 = \max(m_1, m_3, m_4, m_5)$, where $m_0 \equiv m_0(p_{\min}, I, \kappa_1, d, \alpha_1, B, K, \kappa_0)$ is a constant.

■

3.7.2 Proof of Corollary 5

Let Ω_1 denote the event under which Lemma 8 holds. Then $P(\Omega_1^c) \leq 1/m$, where Ω^c denotes the complement of Ω. Let Ω_2 denote the event that the test point X and training data $X_1, \ldots, X_n \in \mathcal{L}$ don't lie in $\mathcal{R}_{\mathcal{B}}$. Then

$$P(\Omega_2^c) \leq (n+1)P(\mathcal{R}_{\mathcal{B}}) \leq (n+1)p_{\max}\text{vol}(\mathcal{R}_{\mathcal{B}}) = O(nh_m).$$

The last step can be explained as follows. Since the decision boundaries are Lipschitz and I is finite, the length of the decision boundaries is a finite constant, and hence $\text{vol}(\mathcal{R}_{\mathcal{B}}) = O(h_m)$.

Now observe that $\widehat{f}_{\mathcal{D},n}$ essentially uses the clairvoyant knowledge of the decision regions $\mathcal{D}$ to discern which labeled points $X_1, \ldots, X_n$ are in the same decision region as X. Conditioning on Ω_1, Ω_2, Lemma 8 implies that $X, X_i \in D$ if and only if $X \overset{p}{\leftrightarrow} X_i$ for all $i = 1, \ldots, n$. Thus, we can define a semi-supervised learner $\widehat{f}_{m,n}$ to be the same as $\widehat{f}_{\mathcal{D},n}$ except that instead of using clairvoyant knowledge of whether $X, X_i \in D$, $\widehat{f}_{m,n}$ is based on whether $X \overset{p}{\leftrightarrow} X_i$. It follows that $\sup_{\mathcal{P}_{XY}(\xi)} \mathbb{E}[\mathcal{E}(\widehat{f}_{m,n})|\Omega_1, \Omega_2] = \sup_{\mathcal{P}_{XY}(\xi)} \mathbb{E}[\mathcal{E}(\widehat{f}_{\mathcal{D},n})]$, and since the excess risk is bounded,

$$\begin{aligned}
\sup_{\mathcal{P}_{XY}(\xi)} \mathbb{E}[\mathcal{E}(\widehat{f}_{m,n})] &= \sup_{\mathcal{P}_{XY}(\xi)} \mathbb{E}[\mathcal{E}(\widehat{f}_{m,n})|\Omega_1, \Omega_2]P(\Omega_1, \Omega_2) + \mathbb{E}[\mathcal{E}(\widehat{f}_{m,n})|\Omega_1^c \cup \Omega_2^c]P(\Omega_1^c \cup \Omega_2^c) \\
&\leq \sup_{\mathcal{P}_{XY}(\xi)} \mathbb{E}[\mathcal{E}(\widehat{f}_{\mathcal{D},n})] + O\left(\frac{1}{m} + nh_m\right) \\
&\leq \epsilon_2(n) + O\left(\frac{1}{m} + n\left(\frac{m}{(\log m)^2}\right)^{-1/d}\right).
\end{aligned}$$

■

3.7.3 Semi-Supervised Learning Upper Bound

If the margin $|\xi| > C_o(m/(\log m)^2)^{-1/d}$, where $C_o = 6\sqrt{d}\kappa_0$ and $m \gg n^{2d}$, we show that the semi-supervised learner proposed in Section 3.5 achieves a finite sample error bound of $O\left((n/\log n)^{-2\alpha/(d+2\alpha)}\right)$. Observe that the clarivoyant counterpart of $\widehat{f}_{m,n}(X)$ is given as

$$\widehat{f}_{\mathcal{D},n}(X) = \widehat{f}_X(X)$$

where

$$\widehat{f}_x(\cdot) = \arg\min_{f' \in \Gamma} \sum_{i=1}^{n} (Y_i - f'(X_i))^2 \mathbf{1}_{x, X_i \in D} + \text{pen}(f').$$

Observe that $\widehat{f}_{\mathcal{D},n}$ is a standard supervised learner that performs piecewise polynomial fit on each decision region $D \in \mathcal{D}$, where the regression function is Hölder-α smooth. Let $n_D = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{X_i \in D}$ denote the number of labeled training examples that fall in a decision region $D \in \mathcal{D}$. Since the regression function on each decision region is Hölder-α smooth, it follows (for example, along the lines of Theorem 8 in [5]) that

$$\mathbb{E}[(f^*(X) - \widehat{f}_{\mathcal{D},n}(X))^2 \mathbf{1}_{X \in D} | n_D] \leq C \left(\frac{n_D}{\log n_D} \right)^{-\frac{2\alpha}{d+2\alpha}}.$$

Now consider

$$\mathbb{E}[(f^*(X) - \widehat{f}_{\mathcal{D},n}(X))^2] = \sum_{D \in \mathcal{D}} \mathbb{E}[(f^*(X) - \widehat{f}_{\mathcal{D},n}(X))^2 \mathbf{1}_{X \in D}] P(D).$$

We will establish the result by taking expectation over $n_D \sim$Binomial$(n, P(D))$ (if $P(D) = O(\log n/n)$, we simply use the fact that the excess risk is bounded), and summing over all decision regions recalling that $|\mathcal{D}|$ is finite. Consider two cases:

1. If $P(D) > \frac{28 \log n}{3n}$,

$$\begin{aligned}
&\mathbb{E}[(f^*(X) - \widehat{f}_{\mathcal{D},n}(X))^2 \mathbf{1}_{X \in D}] P(D) \\
&\quad = \mathbb{E}[\mathbb{E}[(f^*(X) - \widehat{f}_{\mathcal{D},n}(X))^2 \mathbf{1}_{X \in D} | n_D]] P(D) \\
&\quad \leq \mathbb{E}\left[C \left(\frac{n_D}{\log n_D} \right)^{-\frac{2\alpha}{d+2\alpha}} \right] P(D) \\
&\quad = \sum_{n_D=0}^{n} C \left(\frac{n_D}{\log n_D} \right)^{-\frac{2\alpha}{d+2\alpha}} P(n_D) P(D) \\
&\quad \leq C \sum_{n_D=0}^{n} \left(\frac{n_D}{\log n} \right)^{-\frac{2\alpha}{d+2\alpha}} P(n_D) P(D) \\
&\quad \leq \frac{C}{(\log n)^{-\frac{2\alpha}{d+2\alpha}}} \left[\sum_{n_D=0}^{\lceil nP(D)/2 \rceil - 1} n_D^{-\frac{2\alpha}{d+2\alpha}} P(n_D) + \sum_{n_D=\lceil nP(D)/2 \rceil}^{n} n_D^{-\frac{2\alpha}{d+2\alpha}} P(n_D) \right] P(D)
\end{aligned}$$

$$
\begin{aligned}
&\leq \frac{C}{(\log n)^{-\frac{2\alpha}{d+2\alpha}}}\left[P(n_D \leq nP(D)/2) + (nP(D)/2)^{-\frac{2\alpha}{d+2\alpha}}\right]P(D)\\
&\leq \frac{C}{(\log n)^{-\frac{2\alpha}{d+2\alpha}}}\left[e^{-\frac{3nP(D)}{28}}P(D) + 2n^{-\frac{2\alpha}{d+2\alpha}}P(D)^{\frac{d}{d+2\alpha}}\right]\\
&\leq \frac{C}{(\log n)^{-\frac{2\alpha}{d+2\alpha}}}\left[\frac{1}{n} + 2n^{-\frac{2\alpha}{d+2\alpha}}\right]\\
&= O\left(\left(\frac{n}{\log n}\right)^{-\frac{2\alpha}{d+2\alpha}}\right)
\end{aligned}
$$

The second last step follows since

$$
\begin{aligned}
P(n_D \leq nP(D)/2) &= P(nP(D) - n_D \geq nP(D)/2)\\
&= P\left(\sum_{i=1}^{n} P(D) - \mathbf{1}_{X_i \in D} \geq nP(D)/2\right)\\
&= P\left(\sum_{i=1}^{n} Z_i \geq nP(D)/2\right)\\
&\leq \exp\left\{\frac{(nP(D)/2)^2/2}{nP(D)(1-P(D)) + nP(D)/6}\right\} \leq e^{-\frac{3nP(D)}{28}}.
\end{aligned}
$$

The last step follows using Bernstein's inequality since for $Z_i = P(D) - \mathbf{1}_{X_i \in D}$, we have that $|Z_i| \leq 1$ and $\mathrm{var}(Z_i) = P(D)(1-P(D))$.

2. If $P(D) \leq \frac{28 \log n}{3n}$, we have

$$
\mathbb{E}[(f^*(X) - \widehat{f}_{\mathcal{D},n}(X))^2 \mathbf{1}_{X \in D}]P(D) \;\leq\; 4M^2 P(D) = O\left(\frac{\log n}{n}\right).
$$

Thus, it follows that since $|\mathcal{D}| \leq 2^I$

$$
\mathbb{E}[(f^*(X) - \widehat{f}_{\mathcal{D},n}(X))^2] = O\left(\left(\frac{n}{\log n}\right)^{-\frac{2\alpha}{d+2\alpha}}\right).
$$

And using Corollary 5,

$$
\mathbb{E}[(f^*(X) - \widehat{f}_{m,n}(X))^2] = O\left(\left(\frac{n}{\log n}\right)^{-\frac{2\alpha}{d+2\alpha}} + \frac{1}{m} + n\left(\frac{m}{(\log m)^2}\right)^{-1/d}\right).
$$

If $m \gg n^{2d}$, then $1/m + n(m/(\log m)^2)^{-1/d} = O((n/\log n)^{-1})$ and we get an upper bound of $O\left((n/\log n)^{-\frac{2\alpha}{d+2\alpha}}\right)$ on the performance of the semi-supervised learner.

If $|\xi| < C_o(m/(\log m)^2)^{-1/d}$, the decision regions are not discernable using unlabeled data and the target regression function is piecewise Hölder-α smooth on each p-connected region. As shown in [5], for piecewise Hölder-α functions, the proposed estimator achieves an error bound of $\max(n^{-2\alpha/(2\alpha+d)}, n^{-1/d})$. Also, notice that the number of resulting p-connected regions cannot be more than $|\mathcal{D}|$ since the procedure can miss detecting where the marginal density jumps, however with high probability it will not declare two points to be p-connected when the marginal density does not jump between them. Thus, the number of p-connected regions are finite, and an overall error bound of $\max(n^{-2\alpha/(2\alpha+d)}, n^{-1/d})$ holds.

■

3.7.4 Supervised Learning Lower Bound

Consider the single cluster class $\mathcal{P}'_{XY}$ with $\text{supp}(p_X) = [0,1]^d$. For this class, it is known [31] that there exists a constant $c > 0$ such that

$$\inf_{f_n} \sup_{\mathcal{P}'_{XY}} \mathbb{E}[(f^*(X) - f_n(X))^2] \geq cn^{-2\alpha/(d+2\alpha)}.$$

Notice that $\mathcal{P}'_{XY} \subset \mathcal{P}_{XY}(\xi)$ for all ξ. Therefore, we get:

$$\inf_{f_n} \sup_{\mathcal{P}_{XY}(\xi)} \mathbb{E}[(f^*(X) - f_n(X))^2] \geq cn^{-2\alpha/(d+2\alpha)}.$$

If $\xi < c_o n^{-1/d}$, where $c_o > 0$ is a constant, we derive a tighter lower bound of $cn^{-1/d}$. Thus, we will have

$$\inf_{f_n} \sup_{\mathcal{P}_{XY}(\xi)} \mathbb{E}[(f^*(X) - f_n(X))^2] \geq cn^{-1/d}.$$

To establish the tighter lower bound of $cn^{-1/d}$, we use the following theorem based on Assouad's lemma (adapted from Theorem 2.10 (iii) in [57]).

Theorem 6. *Let $\Omega = \{0,1\}^q$, the collection of binary vectors of length q. Let $\mathcal{P}_\Omega = \{P^\omega, \omega \in \Omega\}$ be the corresponding collection of 2^q probability measures associated with each vector. Also let $H(\cdot,\cdot)$ denote the Hellinger distance between two distributions, and $\rho(\cdot,\cdot)$ denotes*

the Hamming distance between two binary vectors. If $H^2(P^{\omega'}, P^{\omega}) \leq \kappa < 2$, $\forall \omega, \omega' \in \Omega$: $\rho(\omega, \omega') = 1$, *then*

$$\inf_{\hat{\omega}} \max_{\omega \in \Omega} \mathbb{E}_{\omega}[\rho(\hat{\omega}, \omega)] \geq \frac{q}{2}(1 - \sqrt{\kappa(1 - \kappa/4)})$$

We will construct such a collection of joint probability distributions $\mathcal{P}_{\Omega} \subseteq \mathcal{P}_{XY}(\xi)$ satisfying Theorem 6 with $q = \ell^{d-1}$, where $\ell = \lceil c_6 n^{1/d} \rceil$, $c_6 > 0$ is a constant. Notice that $\mathbb{E}[(f^*(X) - f_n(X))^2] = \mathbb{E}[R(f^*, f_n)]$, where $R(f^*, f_n)$ denotes the mean square error

$$R(f^*, f_n) = \int (f^*(x) - f_n(x))^2 p(x) dx.$$

Since the mean square error is not symmetric, we will first relate it to a semi-distance $d(\cdot, \cdot)$ defined as follows:

$$d^2(f, f_n) = \int (f^*(x) - f_n(x))^2 dx.$$

For $f^* \equiv f^{\omega}$ and $f_n \equiv f^{\hat{\omega}}$, we will show that the mean square error and semi-distance are related as follows:

$$R(f^{\omega}, f^{\hat{\omega}}) \geq b\ [d^2(f^{\omega}, f^{\hat{\omega}}) - 4M^2\xi]. \tag{3.2}$$

We will then show the following lower bound on the semi-distance in terms of the Hamming distance:

$$d^2(f^{\omega}, f^{\hat{\omega}}) \geq c_7 \ell^{-d} \rho(\hat{\omega}, \omega) \tag{3.3}$$

where $c_7 > 0$ is a constant. Thus, we will have

$$\begin{aligned}
\inf_{f_n} \sup_{\mathcal{P}_{XY}(\xi)} \mathbb{E}[(f^*(X) - f_n(X))^2] &= \inf_{f_n} \sup_{\mathcal{P}_{XY}(\xi)} \mathbb{E}[R(f^*, f_n)] \\
&\geq \inf_{f^{\hat{\omega}}} \sup_{\mathcal{P}_{\Omega}} \mathbb{E}[R(f^{\omega}, f^{\hat{\omega}})] = \inf_{\hat{\omega}} \sup_{\omega \in \Omega} \mathbb{E}_{\omega}[R(f^{\omega}, f^{\hat{\omega}})] \\
&\geq b \left(\inf_{\hat{\omega}} \sup_{\omega \in \Omega} \mathbb{E}_{\omega}[d^2(f^{\omega}, f^{\hat{\omega}})] - 4M^2\xi \right) \\
&\geq b \left(c_7 \ell^{-d} \inf_{\hat{\omega}} \sup_{\omega \in \Omega} \mathbb{E}_{\omega}[\rho(\omega, \hat{\omega})] - 4M^2\xi \right)
\end{aligned}$$

$$\begin{aligned} &\geq b\left(c_7\ell^{-d}\frac{q}{2}(1-\sqrt{\kappa(1-\kappa/4)})-4M^2\xi\right) \\ &\geq b\left(\frac{c_7}{2c_6}(1-\sqrt{\kappa(1-\kappa/4)})-4M^2c_o\right)n^{-1/d} \end{aligned}$$

where the last step follows since $q=\ell^{d-1}$, $\ell=\lceil c_6 n^{1/d}\rceil$ and $\xi<c_o n^{-1/d}$. Thus, there exists $c_o\equiv c_o(c_6,c_7,M,\kappa)$, for which we obtain the desired lower bound of $cn^{-1/d}$, where $c>0$ is a constant.

We now construct $\mathcal{P}_\Omega\subseteq\mathcal{P}_{XY}(\xi)$ along the lines of standard minimax construction that satisfies Theorem 6 with $q=\ell^{d-1}$, $\ell=\lceil c_6 n^{1/d}\rceil$, and Equations. (3.2) and (3.3). We construct the elements (p^ω,f^ω) of our collection as follows. Let $x=(\tilde{x},x_d)\in[0,1]^d$, where $\tilde{x}\in[0,1]^{d-1}$ and $x_d\in[0,1]$. Define

$$\tilde{x}_{\tilde{j}}=\frac{\tilde{j}-1/2}{\ell}\quad\text{ and }\quad\eta_{\tilde{j}}(\tilde{x})=\frac{L}{\ell}\zeta(\ell(\tilde{x}-\tilde{x}_{\tilde{j}}))$$

where $\tilde{j}\in\{1,\ldots,\ell\}^{d-1}$ and $\zeta>0$ is a Lipschitz function with Lipschitz constant 1, and $\text{supp}(\zeta)=(-1/2,1/2)^{d-1}$. Now define

$$g_\omega(\tilde{x})=\sum_{\tilde{j}\in\{1,\ldots,\ell\}^{d-1}}\omega_{\tilde{j}}\eta_{\tilde{j}}(\tilde{x})$$

Then $g_\omega(\cdot)$ is a Lipschitz function with Lipschitz constant L. Now define for $\omega\in\Omega$

$$p^\omega(x) \;=\; ap_1^\omega(x)+(1-a)p_2^\omega(x),$$

where $a\leq 1/2$, $p_1^\omega(x)$ is uniform and supported over $C_1^\omega=\left\{x\in[0,1]^d : x_d\geq\frac{1}{2}+\frac{\xi}{2}+g_\omega(\tilde{x})\right\}$ and $p_2^\omega(x)$ is uniform and supported over $C_2^\omega=\left\{x\in[0,1]^d : x_d\leq\frac{1}{2}-\frac{\xi}{2}+g_\omega(\tilde{x})\right\}$. Therefore, the margin is equal to ξ. And

$$f^\omega(x) \;=\; \frac{ap_1^\omega(x)m_1(x)+(1-a)p_2^\omega(x)m_2(x)}{p^\omega(x)}\mathbf{1}_{\{p^\omega(x)\neq 0\}}-M\mathbf{1}_{\{p^\omega(x)=0\}},$$

where $m_1(x)=M$ and $m_2(x)=-M$. Let Y be continuous and bounded, and also assume that $p_1^\omega(Y|X=x),p_2^\omega(Y|X=x)\leq W$, where $W>0$ is a constant. This implies that

$$p^\omega(Y|X=x)=\frac{ap_1^\omega(x)p_1^\omega(Y|X=x)+(1-a)p_2^\omega(x)p_2^\omega(Y|X=x)}{p^\omega(x)}\leq\frac{2BW}{ab}=\frac{p_{\max}W}{p_{\min}}.$$

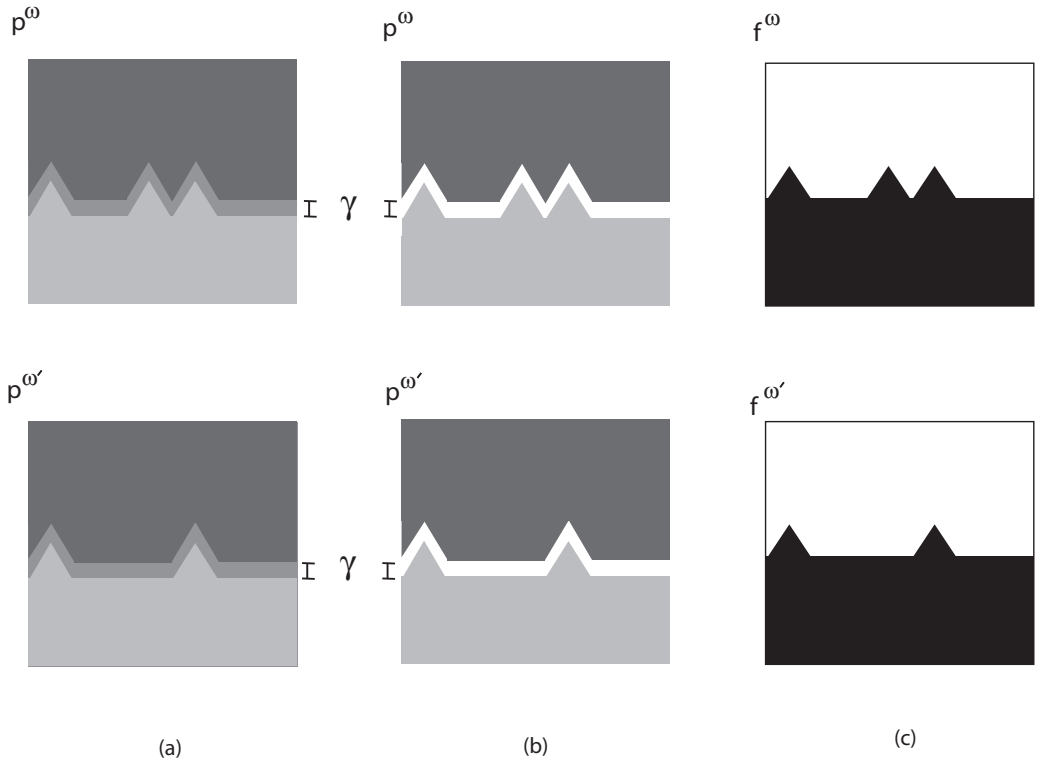


Figure 3.3 Examples of two sets of marginal density functions $p^{\omega}, p^{\omega'}$ for (a) $\xi < 0$, (b) $\xi > 0$ and regression functions $f^{\omega}, f^{\omega'}$ used for minimax construction.

Figure 3.3 shows examples of two marginal density functions $p^{\omega}, p^{\omega'}$ for positive and negative margin, and corresponding regression functions $f^{\omega}, f^{\omega'}$.

Notice that the component densities are supported on compact, connected sets, are Hölder-α smooth for any α, and are bounded from above and below by $b \leq 1$ and $B \geq 4$. To see the latter, notice that

$$p_1^{\omega}(x) = \frac{1}{\mathrm{vol}(C_1^{\omega})} = \frac{1}{\frac{1}{2} - \frac{\xi}{2} - \int g_{\omega}(\tilde{x})d\tilde{x}}, p_2^{\omega}(x) = \frac{1}{\mathrm{vol}(C_2^{\omega})} = \frac{1}{\frac{1}{2} - \frac{\xi}{2} + \int g_{\omega}(\tilde{x})d\tilde{x}}.$$

The lower bound follows since $\mathrm{vol}(C_1^{\omega}), \mathrm{vol}(C_1^{\omega}) \leq 1$, and the upper bound follows since

$$\mathrm{vol}(C_1^{\omega}) \geq \mathrm{vol}(C_1^{\omega}) = \frac{1}{2} - \frac{\xi}{2} - \int g_{\omega}(\tilde{x})d\tilde{x} > \frac{1}{2} - \frac{c_o}{2}n^{-1/d} - \frac{L\|\zeta\|_1}{2c_6}n^{-1/d} \geq 1/4.$$

Here the second last step follows since

$$\int g_{\omega}(\tilde{x})d\tilde{x} = \sum_{\tilde{i} \in \{1,\ldots,\ell\}^{d-1}} \omega_{\tilde{i}}\eta_{\tilde{i}}(\tilde{x}) \leq \ell^{d-1}L\|\zeta\|_1\ell^{-d} = L\|\zeta\|_1\ell^{-1} \leq \frac{L\|\zeta\|_1}{2c_6}n^{-1/d},$$

and the last step holds for $n \equiv n(c_o, c_6, d, L, \|\zeta\|_1)$ large enough. Further, the support sets of the component densities have Lipschitz boundaries with Lipschitz constant L. The component regression functions are uniformly bounded between $-M$ and M, and are Hölder-α smooth for any α. Thus $\mathcal{P}_{\Omega} \subseteq \mathcal{P}_{XY}(\xi)$.

We first establish (3.2).

$$\begin{aligned}
R(f^{\omega}, f^{\hat{\omega}}) &= \int (f^{\omega}(x) - f^{\hat{\omega}}(x))^2 p^{\omega}(x)dx \\
&\geq b\left[\int (f^{\omega}(x) - f^{\hat{\omega}}(x))^2 \mathbf{1}_{\{p^{\omega}(x) \neq 0\}}dx\right] \\
&\geq b\left[\int (f^{\omega}(x) - f^{\hat{\omega}}(x))^2 dx - \int (f^{\omega}(x) - f^{\hat{\omega}}(x))^2 \mathbf{1}_{\{p^{\omega}(x) = 0\}}dx\right] \\
&\geq b\left[d^2(f^{\omega}, f^{\hat{\omega}}) - 4M^2\xi\right]
\end{aligned}$$

Next, we establish (3.3). We will consider two cases:

If $\xi > 0$,

$$\begin{aligned}
d^2(f^{\omega}, f^{\hat{\omega}}) &= \int (f^{\omega}(x) - f^{\hat{\omega}}(x))^2 dx \\
&= 4M^2 \sum_{\tilde{i} \in \{1,\ldots,\ell\}^{d-1}} |\omega_{\tilde{i}} - \hat{\omega}_{\tilde{i}}|^2 \int_{[0,1]^{d-1}} \eta_{\tilde{i}}(\tilde{x})d\tilde{x} = 4M^2L\|\zeta\|_1\ell^{-d}\rho(\hat{\omega}, \omega)
\end{aligned}$$

If $\xi \leq 0$,

$$\begin{aligned} d^2(f^\omega, f^{\hat{\omega}}) &= \int (f^\omega(x) - f^{\hat{\omega}}(x))^2 dx \\ &\geq \min\left(\frac{\frac{2M(1-a)}{\mathrm{vol}(C_2^\omega)}}{\frac{a}{\mathrm{vol}(C_1^\omega)} + \frac{(1-a)}{\mathrm{vol}(C_2^\omega)}}, \frac{\frac{2M(1-a)}{\mathrm{vol}(C_2^{\hat{\omega}})}}{\frac{a}{\mathrm{vol}(C_1^{\hat{\omega}})} + \frac{(1-a)}{\mathrm{vol}(C_2^{\hat{\omega}})}} \right)^2 \cdot \\ &\quad \sum_{\tilde{i} \in \{1,\ldots,\ell\}^{d-1}} |\omega_{\tilde{i}} - \hat{\omega}_{\tilde{i}}|^2 \int_{[0,1]^{d-1}} \eta_{\tilde{i}}(\tilde{x}) d\tilde{x} \\ &\geq 4M^2 a^2 \min\left(\frac{\mathrm{vol}(C_1^\omega)}{\mathrm{vol}(C_2^\omega)}, \frac{\mathrm{vol}(C_1^{\hat{\omega}})}{\mathrm{vol}(C_2^{\hat{\omega}})} \right)^2 L \|\zeta\|_1 \ell^{-d} \rho(\omega, \hat{\omega}) \\ &\geq 4M^2 \frac{a^2}{16} L \|\zeta\|_1 \ell^{-d} \rho(\omega, \hat{\omega}) \end{aligned}$$

The second step follows from the definition of f^ω and p^ω, and the third step follows since $a \leq 1/2 \Rightarrow 1-a \geq a$ and since $\mathrm{vol}(C_1^\omega) \leq \mathrm{vol}(C_2^\omega)$ for all $\omega \in \Omega$. To see the last step, observe that for all $\omega \in \Omega$,

$$\frac{\mathrm{vol}(C_1^\omega)}{\mathrm{vol}(C_2^\omega)} = \frac{\frac{1}{2} - \frac{\xi}{2} - \int g_\omega(\tilde{x}) d\tilde{x}}{\frac{1}{2} - \frac{\xi}{2} + \int g_\omega(\tilde{x}) d\tilde{x}} \geq \frac{\frac{1}{2} - \int g_\omega(\tilde{x}) d\tilde{x}}{\frac{1}{2} - \frac{\xi}{2} + \int g_\omega(\tilde{x}) d\tilde{x}} \geq \frac{\frac{1}{2} - L\|\zeta\|_1 \ell^{-1}}{1 + L\|\zeta\|_1 \ell^{-1}} \geq \frac{1}{4}.$$

Here the third step follows since $\xi \geq -1$ and $\int g_\omega(\tilde{x}) d\tilde{x} \leq L\|\zeta\|_1 \ell^{-1}$, and the last step follows for $n \equiv n(c_6, d, L, \|\zeta\|_1)$ large enough since $\ell = \lceil c_6 n^{1/d} \rceil$. Therefore, for all ξ, we have

$$d^2(f^\omega, f^{\hat{\omega}}) \geq 4M^2 \frac{a^2}{16} L \|\zeta\|_1 \ell^{-d} \rho(\omega, \hat{\omega}) =: c_7 \ell^{-d} \rho(\omega, \hat{\omega}).$$

Thus, (3.3) is satisfied.

Now we only need to show that the condition of Theorem 6 is met, that is, $H^2(P^{\omega'}, P^\omega) \leq \kappa < 2$, $\forall \omega, \omega' \in \Omega : \rho(\omega, \omega') = 1$. Observe that

$$\begin{aligned} H^2(P^{\omega'}, P^\omega) &= H^2(P^{\omega'}(\{X_1, Y_1\}_{i=1}^n), P^\omega(\{X_1, Y_1\}_{i=1}^n)) \\ &= 2\left(1 - \prod_{i=1}^n \left(1 - \frac{H^2(P^{\omega'}(X_i, Y_i), P^\omega(X_i, Y_i))}{2}\right)\right) \end{aligned}$$

We now evaluate

$$\begin{aligned} H^2(P^{\omega'}(X_i, Y_i), P^\omega(X_i, Y_i)) &= \int (\sqrt{p^{\omega'}(X_i, Y_i)} - \sqrt{p^\omega(X_i, Y_i)})^2 \\ &= \int (\sqrt{p_X^{\omega'}(X_i) p_{Y|X}^{\omega'}(Y_i|X_i)} - \sqrt{p_X^\omega(X_i) p_{Y|X}^\omega(Y_i|X_i)})^2 \end{aligned}$$

Recall that $p^{\omega}_{Y|X}(Y_i|X_i) \leq p_{\max}W/p_{\min}$. Since $\rho(\omega,\omega') = 1$, let $\tilde{j}$ denote the index for which $\omega_{\tilde{j}} \neq \omega'_{\tilde{j}}$ and without loss of generality, assume that $\omega_{\tilde{j}} = 1$ and $\omega'_{\tilde{j}} = 0$. Also let $B_{\tilde{j}} = \{x : \tilde{x} \in (\tilde{x}_{\tilde{j}} - \frac{1}{2\ell}, \tilde{x}_{\tilde{j}} + \frac{1}{2\ell})\}$. We will evaluate the Hellinger integral over 4 different regions: (Here we use $\pm$ or $\mp$ to denote that the top sign is for the case $\xi > 0$ and bottom sign is for the case $\xi \leq 0$)

First consider

$$A_1 := \{x : \tilde{x} \in B_{\tilde{j}}, \quad 1/2 \pm \xi/2 \leq x_d < 1/2 \pm \xi/2 + g_\omega(\tilde{x}),$$
$$1/2 \mp \xi/2 < x_d \leq \min(1/2 \mp \xi/2 + g_\omega(\tilde{x}), 1/2 \pm \xi/2)\}$$

Since $p^{\omega'}_X(X_i), p^{\omega}_X(X_i) \in [b, B]$ and $p^{\omega'}_{Y|X}(Y_i|X_i), p^{\omega}_{Y|X}(Y_i|X_i) \in [0, p_{\max}W/p_{\min}]$, for this region, we bound the argument of the integral by $Bp_{\max}W/p_{\min}$.

$$\begin{aligned}
\int_{A_1} (\sqrt{p^{\omega'}_X(X_i)p^{\omega'}_{Y|X}(Y_i|X_i)} - \sqrt{p^{\omega}_X(X_i)p^{\omega}_{Y|X}(Y_i|X_i)})^2 &\leq \frac{Bp_{\max}W}{p_{\min}} \int_{A_1} dx \\
&\leq \frac{Bp_{\max}W}{p_{\min}} 2 \int \eta_{\tilde{j}} d\tilde{x} \\
&= \frac{2Bp_{\max}W}{p_{\min}} L\|\zeta\|_1 \ell^{-d} \\
&\leq \frac{2Bp_{\max}WL\|\zeta\|_1}{p_{\min}(2c_6)^d} n^{-1}
\end{aligned}$$

For $x \notin A_1$, notice that $p^{\omega'}_{Y|X}(Y_i|X_i) = p^{\omega}_{Y|X}(Y_i|X_i) \leq p_{\max}W/p_{\min}$, therefore we have:

$$\int_{x \notin A_1} (\sqrt{p^{\omega'}_X(X_i)p^{\omega'}_{Y|X}(Y_i|X_i)} - \sqrt{p^{\omega}_X(X_i)p^{\omega}_{Y|X}(Y_i|X_i)})^2$$
$$\leq \frac{p_{\max}W}{p_{\min}} \int_{x \notin A_1} \left(\sqrt{p^{\omega'}_X(X_i)} - \sqrt{p^{\omega}_X(X_i)}\right)^2$$

We now evaluate the latter integral over three regions: Before that, we set up some results that will be used in all these cases.

$$|\mathrm{vol}(C^{\omega}_1) - \mathrm{vol}(C^{\omega'}_1)|, |\mathrm{vol}(C^{\omega}_2) - \mathrm{vol}(C^{\omega'}_2)| \leq \int \eta_{\tilde{j}} d\tilde{x} = L\|\zeta\|_1 \ell^{-d} \leq \frac{L\|\zeta\|_1}{(2c_6)^d} n^{-1},$$

Also, we establish that

$$\mathrm{vol}(C^{\omega}_1), \mathrm{vol}(C^{\omega'}_1), \mathrm{vol}(C^{\omega}_2), \mathrm{vol}(C^{\omega'}_2) \geq 1/4.$$

For this, observe that for $n \equiv n(c_o, c_6, d, L, \|\zeta\|_1)$ large enough

$$\mathrm{vol}(C_1^{\omega'}) \geq \mathrm{vol}(C_1^{\omega}) = 1/2 - \xi/2 - \int g_\omega(\tilde{x}) > 1/2 - \frac{c_o}{2} n^{-1/d} - \frac{L\|\zeta\|_1}{2c_6} n^{-1/d} \geq 1/4$$
$$\mathrm{vol}(C_2^{\omega}) \geq \mathrm{vol}(C_2^{\omega'}) = 1/2 - \xi/2 + \int g_{\omega'}(\tilde{x}) \geq 1/2 - \frac{c_o}{2} n^{-1/d} \geq 1/4$$

We are now ready to consider the three regions:

$$A_2 := \{x : x_d \geq 1/2 \pm \xi/2 + g_\omega(\tilde{x})\}$$

Notice that

$$\begin{aligned}
\int_{A_2} \left(\sqrt{p_X^{\omega'}(X_i)} - \sqrt{p_X^{\omega}(X_i)} \right)^2 &= \int_{A_2} \left(\sqrt{\frac{a}{\mathrm{vol}(C_1^{\omega'})}} - \sqrt{\frac{a}{\mathrm{vol}(C_1^{\omega})}} \right)^2 \\
&\leq \frac{a}{4} \int_{A_2} \left(\frac{1}{\mathrm{vol}(C_1^{\omega'})} - \frac{1}{\mathrm{vol}(C_1^{\omega})} \right)^2 \\
&\leq \frac{1}{4} \int_{A_2} \left(\frac{|\mathrm{vol}(C_1^{\omega}) - \mathrm{vol}(C_1^{\omega'})|}{\mathrm{vol}(C_1^{\omega'})\mathrm{vol}(C_1^{\omega})} \right)^2 \leq 4 \frac{L^2 \|\zeta\|_1^2}{(2c_6)^{2d}} n^{-2}
\end{aligned}$$

The second step follows since $2 \leq \sqrt{1/\mathrm{vol}(C_1^{\omega'})} + \sqrt{1/\mathrm{vol}(C_1^{\omega})}$.

$$A_3 := \{x : x_d \leq 1/2 \mp \xi/2 + g_{\omega'}(\tilde{x})\}$$

Notice that

$$\begin{aligned}
\int_{A_3} \left(\sqrt{p_X^{\omega'}(X_i)} - \sqrt{p_X^{\omega}(X_i)} \right)^2 &= \int_{A_3} \left(\sqrt{\frac{1-a}{\mathrm{vol}(C_2^{\omega'})}} - \sqrt{\frac{1-a}{\mathrm{vol}(C_2^{\omega})}} \right)^2 \\
&\leq \frac{1-a}{4} \int_{A_3} \left(\frac{1}{\mathrm{vol}(C_2^{\omega'})} - \frac{1}{\mathrm{vol}(C_2^{\omega})} \right)^2 \\
&\leq \frac{1}{4} \int_{A_3} \left(\frac{|\mathrm{vol}(C_2^{\omega}) - \mathrm{vol}(C_2^{\omega'})|}{\mathrm{vol}(C_2^{\omega'})\mathrm{vol}(C_2^{\omega})} \right)^2 \leq 4 \frac{L^2 \|\zeta\|_1^2}{(2c_6)^{2d}} n^{-2}
\end{aligned}$$

The second step follows since $2 \leq \sqrt{1/\mathrm{vol}(C_2^{\omega'})} + \sqrt{1/\mathrm{vol}(C_2^{\omega})}$.

$$\begin{aligned}
A_4 = \{x : \tilde{x} \notin B_{\tilde{j}}, \quad & 1/2 \mp \xi/2 + g_{\omega'}(\tilde{x}) < x_d < 1/2 \pm \xi/2 + g_{\omega'}(\tilde{x}), \\
\tilde{x} \in B_{\tilde{j}} \quad & \min(1/2 \mp \xi/2 + g_\omega(\tilde{x}), 1/2 \pm \xi/2) < x_d < 1/2 \pm \xi/2\}
\end{aligned}$$

If $\xi > 0$, Notice that

$$\int_{A_4} \left(\sqrt{p_X^{\omega'}(X_i)} - \sqrt{p_X^{\omega}(X_i)} \right)^2 = 0$$

If $\xi \leq 0$, then

$$\begin{aligned}
\int_{A_4} \left(\sqrt{p_X^{\omega'}(X_i)} - \sqrt{p_X^{\omega}(X_i)} \right)^2 = & \int_{A_4} \left(\sqrt{\frac{a}{\mathrm{vol}(C_1^{\omega'})} + \frac{1-a}{\mathrm{vol}(C_2^{\omega'})}} - \sqrt{\frac{a}{\mathrm{vol}(C_1^{\omega})} + \frac{1-a}{\mathrm{vol}(C_2^{\omega})}} \right)^2 \\
\leq & \frac{1}{4} \int_{A_4} \left(\frac{a}{\mathrm{vol}(C_1^{\omega'})} + \frac{1-a}{\mathrm{vol}(C_2^{\omega'})} - \frac{a}{\mathrm{vol}(C_1^{\omega})} - \frac{1-a}{\mathrm{vol}(C_2^{\omega})} \right)^2 \\
\leq & \frac{1}{4} \int_{A_4} \left(\left| \frac{a}{\mathrm{vol}(C_1^{\omega'})} - \frac{a}{\mathrm{vol}(C_1^{\omega})} \right| + \left| \frac{1-a}{\mathrm{vol}(C_2^{\omega'})} - \frac{1-a}{\mathrm{vol}(C_2^{\omega})} \right| \right)^2 \\
\leq & \frac{1}{4} \int_{A_4} \left(\frac{|\mathrm{vol}(C_1^{\omega}) - \mathrm{vol}(C_1^{\omega'})|}{\mathrm{vol}(C_1^{\omega'})\mathrm{vol}(C_1^{\omega})} + \frac{|\mathrm{vol}(C_2^{\omega}) - \mathrm{vol}(C_2^{\omega'})|}{\mathrm{vol}(C_2^{\omega'})\mathrm{vol}(C_2^{\omega})} \right)^2 \\
\leq & 16 \frac{L^2 \|\zeta\|_1^2}{(2c_6)^{2d}} n^{-2}
\end{aligned}$$

The second step follows since $2 \leq \sqrt{\frac{a}{\mathrm{vol}(C_1^{\omega'})} + \frac{1-a}{\mathrm{vol}(C_2^{\omega'})}} + \sqrt{\frac{a}{\mathrm{vol}(C_1^{\omega})} + \frac{1-a}{\mathrm{vol}(C_2^{\omega})}}$.

Therefore, we get that

$$\begin{aligned}
H^2(P^{\omega'}(X_i, Y_i), P^{\omega}(X_i, Y_i)) \leq & \frac{2B p_{\max} W L \|\zeta\|_1}{p_{\min}(2c_6)^d} n^{-1} + 24 \frac{p_{\max} W L^2 \|\zeta\|_1^2}{p_{\min}(2c_6)^{2d}} n^{-2} \\
\leq & \frac{4B p_{\max} W L \|\zeta\|_1}{p_{\min}(2c_6)^d} n^{-1} =: c_8 n^{-1}
\end{aligned}$$

where the second step holds for $n \equiv n(c_6, d, L, \|\zeta\|_1)$ large enough. And $c_8 > 0$ is a constant.

$$H^2(P^{\omega'}, P^{\omega}) \leq 2 \left(1 - \left(1 - \frac{c_8}{2} n^{-1} \right)^n \right) \leq 2(1 - e^{-c_8/2}) =: \kappa$$

where the second step holds for $n \equiv n(c_8)$ large enough. Thus, the conditions of Theorem 6 are met and we have established the desired lower bounds for supervised learning.

■

Chapter 4

Level Set based Approach to fMRI Data Analysis

This chapter describes a new methodology and associated theoretical analysis for rapid and accurate extraction of activation regions from functional MRI data. Most fMRI data analysis methods in use today adopt a hypothesis testing approach, in which the BOLD (Blood Oxygen Level Dependent) signals in individual voxels (volumetric element of brain activation map) or clusters of voxels are compared to a threshold. In order to obtain statistically meaningful results, the testing must be limited to very small numbers of voxels/clusters or the threshold must be set extremely high. Furthermore, voxelization introduces partial volume effects (PVE), which present a persistent error in the localization of activity that no testing procedure can overcome. We abandon the multiple hypothesis testing approach, and instead advocate an approach based on set estimation that aims to control the localization error. To do this, we view the activation regions as level sets of the statistical parametric map (SPM) under consideration. The estimation of the level sets, in the presence of noise, is then treated as a statistical inference problem. We describe a level set estimator and show that the expected volume of the error is proportional to the sidelength of a voxel. Since PVEs are unavoidable and produce errors of the same order, this is the smallest error volume achievable. Experiments demonstrate the advantages of this new theory and methodology, and the statistical reasonability of controlling the localization error rather than the probability of error for multiple hypothesis testing.

4.1 fMRI Analysis

Statistical analysis of fMRI (functional Magnetic Resonance Imaging) data involves translating a time series of MRI brain images, collected while the subject is performing a designated task, to an activation map that identifies regions of the brain that have high activity associated with the task. The time series data is used to generate a statistical parameter map (SPM) in the form of t-statistics, cross-correlation coefficients, z-statistics, or one of a variety of other measures, as described in [65]. First we review the standard hypothesis testing approach used in most studies today, and we point out the limitations and flaws of such methods. Then we formulate the fMRI analysis problem as a level set estimation task, and discuss the virtues of this perspective. Before moving on, we establish basic notation that will be used throughout the chapter. Let y_i, $i = 1, \ldots, n$, denote the elements of the SPM under consideration, n being the number of voxels. In this chapter, for simplicity of presentation, we will assume $y_i \in [-1, 1]$. In general, each statistic is modeled as the sum of a deterministic and stochastic component:

$$y_i = \bar{f}_i + \epsilon_i \ .$$

The deterministic component $\bar{f}_i$ is the mean value of y_i, and is viewed as the average of an underlying continuous activation function over the i-th voxel. The stochastic component ϵ_i is assumed to have a mean value of zero. Assume that the volume of the brain is embedded into the unit cube $[0, 1]^3$, and that each voxel corresponds to $1/n$ of this volume. Let f denote the continuous activation function defined on $[0, 1]^3$. Then $\bar{f}_i = n \int_{V_i} f(x)\, dx$, where V_i is the subcube that is the i-th voxel (the factor of n accounts for normalization by the voxel volume).

4.1.1 Hypothesis Testing for fMRI

Broadly speaking, hypothesis testing procedures are based on thresholding the SPM, or functions of it, at a certain level. Points exceeding the threshold are declared as active. In voxel-wise testing, it is usually assumed that the ϵ_i are zero-mean Gaussian errors. Inactive

voxels are assumed to have $\bar{f}_i = 0$, and the Gaussian distribution of ϵ_i then provides a principled means for selecting an appropriate threshold. Since this involves simultaneous testing of multiple hypotheses (equal to the number of voxels, n), the threshold needs to be adjusted to account for this and provide a family-wise error control, i.e. instead of a 5% false alarm control per voxel, we want that the test should report false activation anywhere only 5% of the times. To accomplish this, either the threshold is increased using Bonferroni correction (BC) (union bounding) by a factor of n, or a sequential p-value[1] method such as FDR (False Discovery Rate) which is defined as the expected proportion of false discoveries (falsely detected activations) [42]. FDR is controlled by a data-dependent threshold based on the observed p-value distribution, and thus is adaptive to the sparsity of the signal [43]. This notion is useful since in multiple hypotheses testing problems, such as fMRI, most of the tests are expected to follow the null hypothesis as there is activation only in very small regions of the brain. However, the activations in the brain are not independent from voxel to voxel, and hence these techniques result in overly conservative thresholds with very weak detection power; typically very few voxels exceed these stringent, albeit statistically sound, thresholds. One way to circumvent this problem is to perform a much smaller number of tests, say on a subset of voxels or larger clusters/groups of voxels [66]. However, a heuristic clustering procedure is proposed in [66] and it is noted that good clustering is essential for the success of the approach as aggregation in homogeneous regions leads to lower variance and hence improved SNR, but aggregation of heterogenous regions leads to higher bias and weakening of the signal. Thus restricting the testing to a subset of voxels or clusters may significantly sacrifice the resolution or regional specificity of fMRI analysis. An excellent discussion of this tradeoff may be found in [67], which also describes a variety of testing approaches based on the theory of Gaussian fields. Also, we mention that other alternatives to voxel-based testing have been proposed. For example, [68] uses hypothesis testing based on the wavelet transform of the SPM to detect activity. Thresholding the wavelet coefficients is helpful in

[1]In statistical hypothesis testing, the p-value is the probability of obtaining a value of the test statistic under the null hypothesis that is at least as extreme as the one that was actually observed.

detecting edges or discontinuities in the SPM. However, brain activation is typically spatially diffused and cannot be accurately characterized as a sharp transition. Thus, wavelet based approaches often yield conservative estimates of brain activity.

Beyond these difficulties associated with hypothesis testing approaches to fMRI, there is also the error due to partial volume effects (PVE). PVEs are always present in fMRI and place a limit on the regional specificity (i.e., accuracy with which activation can be localized). The localization accuracy is proportional to the sidelength of a voxel. For example, suppose that a certain voxel contains a volume of the brain that is activated in one half of the voxel, but not the other. It is quite possible that the SPM value for this voxel could exceed a detection threshold, but subsequently inferring that this implies this entire voxel volume is active is obviously incorrect. One can only safely say that some part of the voxel volume is active. Similar problems arise in cluster- and wavelet-based testing.

4.1.2 Level Set Estimation for fMRI

In light of the challenges associated with controlling the probability of error in fMRI, and the fact that even under such control, errors in regional specificity persist at the voxel level due to PVE, we advocate an alternative to hypothesis testing. The best we can hope for is a localization error whose volume is proportional to voxel side-length, so we aim to localize activation to within this accuracy.

The ideal goal of fMRI is to determine the subset of $[0,1]^3$ where the activation function f exceeds a certain positive level, indicating regions where the BOLD response is especially strong.

Aim 1. *(Ideal): For a given level $\gamma > 0$, determine the level set*

$$G^*_\gamma \equiv \{x \in [0,1]^d : f(x) \geq \gamma\} \ \subset \ [0,1]^3.$$

This is similar in spirit to testing approaches based on the theory of Gaussian random fields [67, 69]. In those approaches, the SPM is viewed as a voxelated representation of an underlying continuous Gaussian field. The function f in our set-up would be the mean of

such a field. Based on the Gaussian field assumption, one can probabilistically characterize the *excursion set*, which is closely related to our notion of a level set. The excursion set is a level set of the SPM, whereas G^*_γ is the gamma-level set of only the *deterministic* component of the SPM.

The ideal aim is not achievable for two reasons. First, the PVE artifact limits the accuracy of any approach to this problem to $n^{-1/3}$, the side-length of a voxel. Second, the stochastic component of the SPM introduces another source of error. Remarkably, a careful statistical analysis shows that the effect of the stochastic error can be controlled to order $n^{-1/(3+\eta)}$. The extra price, as reflected by the η term in the exponent is related to the smoothness of the activation function around the level of interest. If the activation function changes sharply, localization is easier and η is small. On the other hand if the activation is diffused, it is harder to identify the level set and hence η is large. In anticipation of this, we pose a second, more useful, aim:

Aim 2. *(Practical): For a given level $\gamma > 0$, construct an estimator of the γ-level set, denoted $\widehat{G}$, that satisfies the following error bound:*

$$\mathbb{E}\left[\text{vol}\left(\widehat{G}\Delta G^*_\gamma\right)\right] \propto n^{-1/(3+\eta)}, \tag{4.1}$$

where $\eta \geq 0$ is related to the smoothness of f around the level γ, $\mathbb{E}$ denotes the expectation over the random stochastic errors, vol *stands for volume, and $\widehat{G}\Delta G^*_\gamma$ denotes the symmetric difference between the sets G^*_γ and $\widehat{G}$, and is defined in (2.1), Chapter 2.*

Similar quantifications of the size of the error can be made in terms of probability, rather than expectation, but the expectation bound most clearly illustrates the performance.

Remark: As discussed in Chapter 2, controlling the volume of the symmetric set difference may result in set estimates that deviate significantly from the true set. This is undesirable as the reconstructed brain activation patterns are used for further inference regarding the functioning of the brain. Thus it is of interest to control the Hausdorff or worst-case error as defined in (2.2) that guarantees a spatially uniform confidence interval, and can preserve the shape and connectivity of the activation regions. The results of Chapter 2 indicate that

a Hausdorff control of $n^{-1/(3+\eta)}$ can be provided on the maximal distance deviation between the true and estimated activation regions. However, since the resolution of fMRI images is limited to 64×64 pixels/slice, the current histogram based technique described in Chapter 2 may not yield practically impressive performance as it requires sufficiently high number of samples to yield useful results. The symmetric set difference based estimator that we describe in this chapter uses spatially adaptive partitions that yield practically effective results. Thus, we will focus on average error control as offered by the volume of the symmetric set difference, though it will be desirable to guarantee worst-case control, once Hausdorff accurate spatially adapted estimators are developed.

4.1.3 The Level Set Approach vs. Gaussian Field Approaches

As pointed out above, there is a close connection in the formulation of level set estimation and testing based on Gaussian random field models. Our main criticism of the latter is that they are based on a strong assumption about the spatial dependencies in the BOLD signal, namely that the SPM is a realization of a smooth Gaussian process. While there is little doubt that dependencies exist, it is very unclear that a Gaussian process is a reasonable model for them. The assumed smoothness of the field, W in [67], strongly influences the shapes, sizes and boundary smoothness of the excursion sets. Even the best choice of W may not yield excursion sets that are good models for real activation patterns. In contrast, the level set approach assumes only that the boundary of G_γ^* is lower dimensional (e.g., a two dimensional surface when studying a three dimensional volume). No further assumptions are placed on the shape or smoothness of the level set.

4.2 Level Set Activation Detection in fMRI

4.2.1 Error metrics

The symmetric set difference measure does not have a natural empirical counterpart, and hence careful selection of an error metric is the first step in designing an effective level set estimator. In particular, we use the method presented in [22] that is designed to minimize

the *weighted* symmetric difference [2]

$$\mathcal{E}(G, G^*_\gamma) \equiv \int_{G \Delta G^*_\gamma} |\gamma - f(x)| dx. \tag{4.2}$$

We show that, under mild assumptions on the local regularity of the activation function around the level of interest, this metric can provide control over the volume of the symmetric set difference. First, consider activation functions with $f(x) < \gamma_{\min}$ in inactive regions and $f(x) > \gamma_{\max}$ in active regions. Note that if $\gamma_{\min} < \gamma < \gamma_{\max}$, then there exist constants $c, C > 0$ such that $c \leq |\gamma - f(x)| \leq C$. It follows that

$$c \,\mathrm{vol}(G \Delta G^*_\gamma) \leq \int_{G \Delta G^*_\gamma} |\gamma - f(x)| dx \leq C \,\mathrm{vol}(G \Delta G^*_\gamma) \ ,$$

and therefore minimizing $\mathcal{E}(G, G^*_\gamma)$ minimizes the volume of the erroneous region, as desired. Now consider the case when the activation function does not exhibit a sharp transition but is smooth as characterized by the following assumption[3] as introduced in [37].

Definition 3. *For any $\gamma > 0, \kappa \geq 1$, a function f is said to have κ-exponent at level γ with respect to a probability measure P if there exist constants $c_0 > 0$ and $\epsilon_0 > 0$ such that, for all $0 < \epsilon \leq \epsilon_0$,*

$$P(\{x : |f(x) - \gamma| \leq \epsilon\}) \leq c_0 \epsilon^{1/(\kappa - 1)}.$$

Observe that $\kappa = 1$ corresponds to a jump in the activation function at the level γ and increasing κ implies increasing smoothness of f around level γ. Under this local smoothness assumption, Proposition 2.1 from [37] implies that there exists $C > 0$ such that

$$\mathrm{vol}(G \Delta G^*_\gamma) \leq C (\mathcal{E}(G, G^*_\gamma))^{1/\kappa}. \tag{4.3}$$

Thus, control over $\mathcal{E}(G, G^*_\gamma)$ implies control over the volume of the erroneous region.

The quantity $\mathcal{E}(G, G^*_\gamma)$ cannot be directly evaluated without knowledge of G^*_γ (the level set we wish to estimate.) However, as shown in [22], the weighted symmetric difference can be decomposed as $\mathcal{E}(G, G^*_\gamma) = \mathcal{R}(G) - \mathcal{R}(G^*_\gamma)$, where

$$\mathcal{R}(G) \equiv \int \frac{\gamma - f(x)}{2} \left[\mathbf{1}_{x \in G} - \mathbf{1}_{x \in G^c} \right] dx \tag{4.4}$$

[2] This metric is also known as the excess mass deficit, see [37].

[3] This is analogous to Tsybakov's *noise margin assumption* [70, 71] for the classification setting.

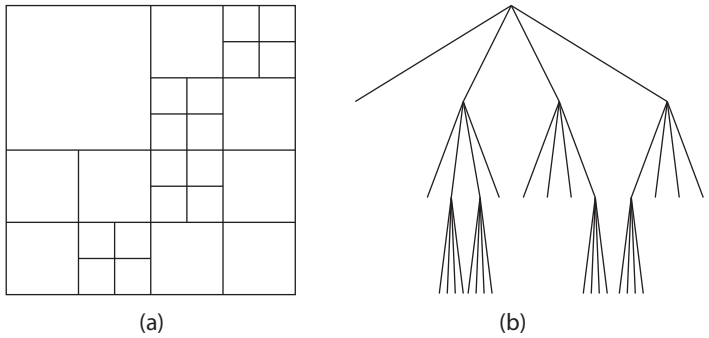


Figure 4.1 (a) An example Recursive Dyadic Partition (RDP) of a 2-d function domain and (b) associated tree structure.

and $\mathbf{1}_A = 1$ if event A is true and 0 otherwise. Thus minimizing $\mathcal{E}(G, G^*_\gamma)$ is equivalent to minimizing $\mathcal{R}(G)$, since $\mathcal{R}(G^*_\gamma)$ is a constant, and the value of $\mathcal{R}(G)$ is independent of G^*_γ. Furthermore, $\mathcal{R}(G)$ has an empirical counterpart that can be computed from the SPM:

$$\widehat{\mathcal{R}}_n(G) = \frac{1}{n}\sum_{i=1}^{n} \frac{\gamma - y_i}{2} \left[\mathbf{1}_{x_i \in G} - \mathbf{1}_{x_i \in G^c}\right]. \tag{4.5}$$

Note that $\mathcal{R}(G) = \mathbb{E}\left[\widehat{\mathcal{R}}_n(G)\right]$.

4.2.2 Estimation via Trees

Following [22], we estimate the level set of the activation function f based on the SPM by using a tree-pruning method akin to CART [48] or dyadic decision trees [51]. The estimator is built over a recursive dyadic partition (RDP) of the domain. An RDP can be identified with a tree structure, where each leaf of the tree corresponds to a cell of the dyadic partition (see Fig. 4.1). Trees are utilized for a couple of reasons. First, they provide a simple means of generating a spatially adaptive partition, yielding an automatic data aggregation in regions estimated to be strictly above or below the γ level. This adaptive and automatic aggregation effectively boosts the signal-to-noise ratio. Second, the optimal partition can be computed very rapidly using a simple bottom-up pruning scheme. Here we summarize the tree-based estimator proposed in [22].

Let $\mathcal{T}_n$ denote the collection of all 8-ary trees in three dimensions (8-ary trees are based on recursively partitioning cubic volumes into 8 sub-cubes). For example, consider a $64 \times 64 \times 64$

voxel volume. The voxels represent the limit of the 8-ary partition process, generated by a 6 level 8-ary tree ($2^6 = 64$). In this example, $n = 64^3$ and the side-length of each voxel is $n^{-1/3} = 1/64$ (assuming the brain volume is normalized to be the unit cube). Note, however, that is not necessary to complete the partitioning process to the voxel level; using less than 6 levels will result in partition cells composed of groups of voxels. Moreover, the level of the tree can vary spatially, yielding smaller cells in certain areas and larger cells in others. A spatially adapted partition is crucial in level set estimation, since ideally we wish to aggregate the SPM wherever $f(x)$ strictly exceeds or falls below the target level of γ to boost the signal-to-noise ratio.

Let $\pi(T)$ denote the partition induced on $[0,1]^3$ by the 8-ary tree T. Each leaf node of the tree corresponds to a cell of the partition. A zero or one is assigned to each leaf node of T (equivalently, to each cell $L \in \pi(T)$), and the union of leafs with label one form a set denoted G_T. Let $|L|$ denote the volume of the leaf L. Based the theory of tree-based level set estimators developed in [22], we have the following result.

Theorem 7 (Lemma 2, Willett et al. [22])**.** *With probability at least* $1 - 1/n$*,*

$$\mathcal{R}(G_T) \leq \widehat{\mathcal{R}}_n(G_T) + \Phi_n(T) \quad \forall\, T \in \mathcal{T}_n,$$

where

$$\Phi_n(T) \equiv \sum_{L \in \pi(T)} \sqrt{\frac{2|L|}{n} \log\left(\frac{4n}{|L|^{4/3}}\right)}\,. \tag{4.6}$$

The proof of this theorem utilizes a union bound over the leaves coupled with Hoeffding's inequality (a concentration of measure inequality for bounded random variables), which bounds the contribution of the risk from each individual leaf. Theorem 7 shows that the ideal quantity, $\mathcal{R}(G_T)$, that we wish to minimize is upper bounded by the sum of empirical version $\widehat{\mathcal{R}}_n(G_T)$ and $\Phi_n(T)$, both of which are easily computed; i.e., although $\mathcal{R}(G_T)$ requires exact knowledge of $f(x)$, the upper bound does not. Thus, the set that minimizes $\widehat{\mathcal{R}}_n(G_T) + \Phi_n(T)$ also makes $\mathcal{R}(G_T)$ (and hence the volume error) small with very high probability. Intuitively, we can interpret $\Phi_n(T)$ as a regularization term which discourages excessively complex partitions.

Theorem 7 leads directly to the following level set estimator:

$$\widehat{G}_T = \arg \min_{G_T : T \in \mathcal{T}_n} \left\{ \widehat{\mathcal{R}}_n(G_T) + \Phi_n(T) \right\}. \tag{4.7}$$

In addition, the estimator defined by (4.7) and (4.6) is rapidly computable. In fact, a translation-invariant estimator can be computed in $O(n \log n)$ operations [22]. Furthermore, as shown in the following section, the estimator is nearly optimal in terms of the expected volume of the erroneous region.

4.2.3 Performance Analysis

In Theorem 6 of [22] an upper bound on $\mathcal{E}(\widehat{G}_T, G^*_\gamma)$, the expected weighted symmetric difference error of the estimator proposed in (4.7), is derived. Coupled with (4.3), we can translate the performance bound on the weighted symmetric difference error to an upper bound on the expected volume of the erroneous region (total volume of missed activation and falsely detected activation). In particular, the following theorem shows that for a broad class of functions f, the proposed method is nearly optimal in terms of the expected volume of the erroneous region. Here it is assumed that the boundaries of G^*_γ are two-dimensional surfaces (i.e., the boundary of a volume is a surface); please refer to [22] for details.

Theorem 8.

$$\mathbb{E}\left[\mathrm{vol}\left(\widehat{G}_T \Delta G^*_\gamma\right)\right] \propto \mathbb{E}\left[(\mathcal{E}(\widehat{G}_T, G^*_\gamma))^{1/\kappa}\right] \propto \left(\frac{\log n}{n}\right)^{1/(3+2(\kappa-1))}.$$

This shows that the expected volume of the error is nearly equal to the minimum dictated by partial volume effects, $n^{-1/3}$, except for a term in the exponent that depends on the smoothness of the activation map in the vicinity of the desired level γ. For $\kappa = 1$, the expected localization error is exactly $n^{-1/3}$ except for a logarithmic factor, which is relatively negligible. Furthermore, the estimator automatically adapts to the smoothness of the underlying activation map, as well as form and extent of the level set G^*_γ, without prior constraints on shape or size of the activation regions. The level set estimator is also computationally efficient.

4.3 Improved Estimation via Cross-Validation

The above theoretical analysis demonstrates that the estimator in (4.7) controls the expected volume of the error. However, in practice, because the regularization term $\Phi_n(T)$ is based on worst-case analysis techniques, the resulting level set estimate can be over-regularized (i.e. over-smoothed). Attenuating the regularization term can result in improved performance. It can be shown that choosing this attenuation factor using a cross-validation procedure is both empirically effective and retains the optimality of the original estimator.

In particular, we assume that we have access to two data sets or runs collected under the same conditions, and we denote the two resulting SPMs as $\{y_i^T\}_{i=1}^n$ for training data and $\{y_i^V\}_{i=1}^n$ for validation data. Such data is often available in fMRI studies [66]. For an attenuation factor $\lambda \in [0,1]$, we have the estimator

$$\widehat{G}_\lambda = \arg \min_{G_T : T \in \mathcal{T}_n} \left\{ \widehat{\mathcal{R}}_n^T(G_T) + \lambda \Phi_n(T) \right\}, \tag{4.8}$$

where $\widehat{\mathcal{R}}_n^T$ (4.5) is evaluated using the training data. We then choose the optimal attenuation λ as

$$\widehat{\lambda} = \arg \min_{\lambda \in \Lambda} \widehat{\mathcal{R}}_n^V(\widehat{G}_\lambda),$$

where $\widehat{\mathcal{R}}_n^V$ is (4.5) evaluated using the validation data, and $\Lambda \subset [0,1]$ such that the size of the regularization parameter search space $|\Lambda|$ is polynomial in n and $1 \in \Lambda$. The latter condition ensures that the theoretically optimal estimator without attenuation ($\lambda = 1$) of the regularization term is contained within the search space. Set the final estimate to be

$$\widehat{G} = \widehat{G}_{\widehat{\lambda}}.$$

From here it is possible to derive the following theorem.

Theorem 9.

$$\begin{aligned} \mathbb{E}\left[\mathcal{E}(\widehat{G}, G_\gamma^*)\right] &\leq \min_{\lambda \in \Lambda} \mathbb{E}\left[\mathcal{E}(\widehat{G}_\lambda, G_\gamma^*)\right] + \sqrt{\frac{2\log(|\Lambda| n)}{n}} + \frac{1}{n} \\ &\propto \left(\frac{\log n}{n}\right)^{\kappa/(3+2(\kappa-1))}. \end{aligned}$$

Proof. The proof of the theorem follows through a straightforward application of the results in [72] for cross-validation using a generic risk function. We provide a brief sketch here. Applying Hoeffding's concentration inequality and union bound, we have

$$P\left(\max_{\lambda\in\Lambda}\mathcal{R}(\widehat{G}_\lambda)-\widehat{\mathcal{R}}_n^V(\widehat{G}_\lambda)>\epsilon\right)\leq|\Lambda|e^{-n\epsilon^2/2}.$$

Using this we get, with probability $>1-1/n$,

$$\mathcal{R}(\widehat{G})\ \leq\ \widehat{\mathcal{R}}_n^V(\widehat{G})+\sqrt{2\frac{\log(|\Lambda|n)}{n}}\ \leq\ \widehat{\mathcal{R}}_n^V(\widehat{G}_\lambda)+\sqrt{2\frac{\log(|\Lambda|n)}{n}}.$$

The last step holds for every $\{\widehat{G}_\lambda\}_{\lambda\in\Lambda}$. Taking expectation with respect to the validation and training data sets,

$$\mathbb{E}[\mathcal{R}(\widehat{G})]\leq\mathbb{E}[\mathcal{R}(\widehat{G}_\lambda)]+\sqrt{2\frac{\log(|\Lambda|n)}{n}}+\frac{1}{n}.$$

Since this is true for every $\{\widehat{G}_\lambda\}_{\lambda\in\Lambda}$, the first result follows by subtracting $\mathcal{R}(G_\gamma^*)$ from both sides. The error bound in terms of n follows since the unscaled estimator corresponding to $\lambda=1$ is contained in the search space, and using Theorem 8 that provides an upper bound on the performance of the unscaled estimator. □

Since the bounds hold with high probability and not just in expectation, it can be shown that $\mathbb{E}\left[\mathrm{vol}\left(\widehat{G}\Delta G_\gamma^*\right)\right]\propto(\log n/n)^{1/(3+2(\kappa-1))}$. Thus, Theorem 9 implies that the error volume of the cross-validated estimator is also near-optimal.

4.4 Experimental Results

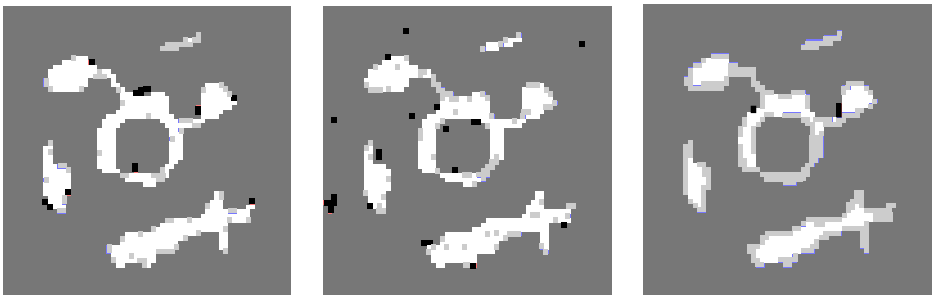
4.4.1 Simulated Data

We first use simulated data to perform our holdout method for parameter tuning. The simulated activation pattern consists of regions of different activation level, representing levels of neurological activity. Additive unit-variance zero-mean Gaussian noise is used to corrupt this underlying activity to form our simulated data. The SPM is then computed using this known $\mathcal{N}(0,1)$ noise distribution. In Fig. 4.2, we compare the results obtained by

the proposed level set method to those obtained by traditional pre-smoothing and thresholding (without using Bonferroni correction) at level $\gamma = 0.7$. Figure 4.2(a) shows the level set estimate obtained by the hold-out procedure described in the last section. In Fig. 4.2(b), we use a Gaussian smoothing kernel, where the smoothing bandwidth of 1.2 voxels FWHM (Full-Width at Half-Maximum) was chosen *clairvoyantly* via directly minimizing $\mathrm{vol}(\widehat{G}\Delta G^*_\gamma)$ on the holdout data set, and then threshold at level $\gamma = 0.7$. Since this clairvoyant estimator is based on knowledge of the true simulated activation pattern, the performance of this conventional approach will exceed what is possible in practice when the true activation pattern G^*_γ is unknown. In Fig. 4.2(c), we also show a more typical example where a bandwidth of 3 voxels FWHM is chosen, demonstrating a severe degradation in performance due to oversmoothing of the SPM. Note that in comparison to simple smoothing and thresholding, the errors of the level set method are well-localized along the boundary of the level set, as the theory predicts. Averaging over 100 realizations of the SPM, we have an average symmetric difference volume of 0.024 using level sets, 0.030 using thresholding with clairvoyant pre-smoothing, and 0.077 using thresholding with a fixed typical amount of pre-smoothing. This shows that the level set estimation procedure outperforms pre-smoothing and thresholding estimate *even with the best clairvoyant choice of the smoothing bandwidth.*

4.4.2 fMRI Data

We now consider real fMRI data. This data set consists of 64×64 axial slices with 122 time samples taken during a finger-tapping experiment. By splitting the data into two temporal halves, we generate two independent sets of data for the cross-validation method: the first 61 time samples for training, and the remaining 61 for validation. Shown in Fig. 4.3 are the results of the holdout based level set procedure and thresholding with pre-smoothing, using two different levels of γ. For pre-smoothing, we used a Gaussian smoothing kernel with a 1.2 voxel FWHM, corresponding to the clairvoyant choice of bandwidth in the simulated experiment of the previous section. By picking γ, we can examine different levels of activation in fMRI studies. Figure 4.3 (a) and (c) show the results of the proposed level



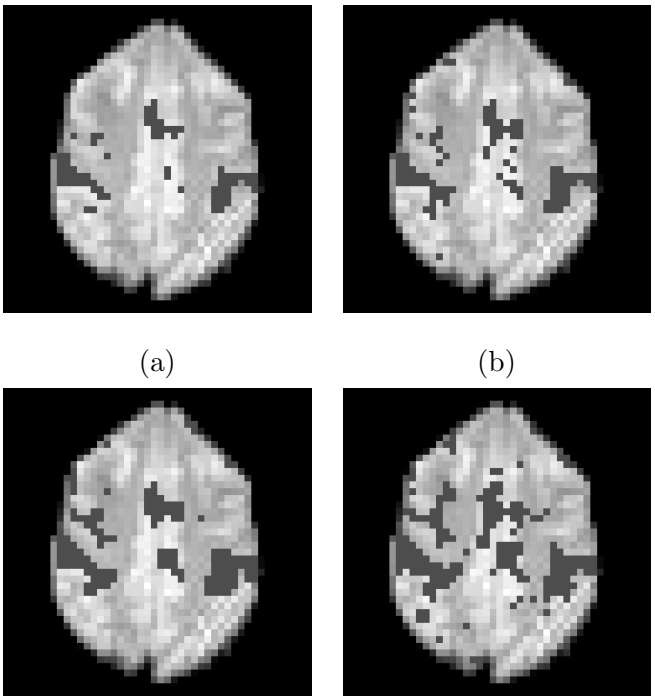
(a) (b) (c)

Figure 4.2 Comparison of $\gamma = 0.7$ level set estimation, thresholding with clairvoyant pre-smoothing, and thresholding with typical pre-smoothing for a simulated activation pattern. Black regions are false positives, light gray regions indicate false negatives, their union comprises the erroneous region $\widehat{G}\Delta G^*_\gamma$. (a) Level set estimate; $\mathrm{vol}(\widehat{G}\Delta G^*_\gamma) = 0.024$. (b) Voxel-wise threshold estimate with clairvoyant pre-smoothing; $\mathrm{vol}(\widehat{G}\Delta G^*_\gamma) = 0.030$. (c) Voxel-wise threshold estimate with typical pre-smoothing; $\mathrm{vol}(\widehat{G}\Delta G^*_\gamma) = 0.075$.

set based method for $\gamma = 0.95$ and $\gamma = 0.9$, respectively. Comparing these results to voxel-wise thresholding with pre-smoothing (without Bonferroni correction) in Fig. 4.3 (b) and (d), we can see that the detected regions using the proposed level set estimation method show significantly fewer spurious detected areas compared to conventional methods, yielding better localization of neural activity. For statistically sound thresholding procedures, typically Bonferroni correction or FDR is used to address the multiplicity of the hypothesis testing problem, however applying these corrections to the fMRI data set resulted in overly conservative thresholds and no significant activation was detected.

4.5 Concluding Remarks

In this chapter, we abandoned standard hypothesis testing fMRI analysis in favor of a level set approach that aims to control the volume of the erroneous region. We argued that controlling the error volume is more natural in light of partial volume effects, which result in an unavoidable error in any event. The proposed level set estimator produces an estimate of active regions that is guaranteed to have an error that is commensurate with partial volume effects. In other words, the error volume, and hence the error in regional specificity, is



(a) (b)

(c) (d)

Figure 4.3 Estimates of neural activation regions. The dark gray regions overlaid onto the anatomic brain image represent the declared activity. With $\gamma = 0.95$: (a) level set estimate, (b) voxel-wise threshold estimate with pre-smoothing. With $\gamma = 0.90$: (c) level set estimate, (d) voxel-wise threshold estimate with pre-smoothing.

controlled to nearly the absolute minimum. One matter that we have not investigated here is the choice of level, γ. The higher the level γ, the smaller the level set G^*_γ. In effect, γ gauges the strength of the activation in the set G^*_γ. The selection of γ could be based on a secondary criterion, such as the false discovery rate. Alternatively, it may be informative to examine several different level sets. These sets will be nested, with lower level sets containing higher level sets, providing a more descriptive view of the activation patterns.

Chapter 5

Adaptive Mobile Sensing for Environmental Monitoring

This chapter investigates data-adaptive path planning schemes for wireless networks of mobile sensor platforms. We focus on applications of environmental monitoring, in which the goal is to reconstruct a spatial map of environmental factors of interest. Traditional sampling deals with data collection processes that are completely independent of the target map to be estimated, aside from possible a priori specifications reflective of assumed properties of the target. We refer to such processes as passive learning methods. Alternatively, one can envision sequential, adaptive data collection procedures that use information gleaned from previous observations to guide the process. We refer to such feedback-driven processes as active learning methods. Active learning is naturally suited to mobile path planning, in which previous samples are used to guide the motion of the mobiles for further sampling. We investigate capabilities of active learning methods for mobile sensing and based on recent advances in assessing active learning performance, we characterize the optimal tradeoffs between latency, path lengths, and accuracy. Adaptive path planning methods are developed to guide mobiles in order to focus attention in interesting regions of the sensing domain, thus conducting spatial surveys much more rapidly while maintaining the accuracy of the estimated map. The theory and methods are illustrated in the application of water current mapping in a freshwater lake.

5.1 Introduction

Many environmental monitoring applications require fine resolution and high fidelity mapping of a spatial phenomenon distributed over a vast physical extent, e.g. aquatic and terrestrial ecosystem studies, detection of toxic biological and chemical spreads, oil spills and weather patterns. This requires an impractically large number of sensing elements to be distributed according to a pre-computed sampling strategy over the given area. Mobile sensor networks offer an alternative by trading off the cost of large number of static sensors in exchange for increased latency, and provide flexible sampling opportunities. Efficient path planning is necessary to harness these benefits of mobile sensing systems and obtain the best possible resolution in minimum time.

Traditional data sampling methods use a pre-computed, fixed strategy that involves uniform sampling of the field at regular intervals, and the sampling pattern is not updated as measurements are collected. We refer to such strategies as *passive learning* methods. When the spatial phenomenon of interest is smoothly varying, passive methods are known to perform well. However, many environmental variables exhibit highly localized features in the spatial map like changepoints or edges that need to be accurately captured and tracked. For such cases, one can envision sequential, adaptive path planning where information gleaned from previous samples is used to focus mobile paths towards regions of sharp changes in the environmental field where more intensive sampling is required. Such feedback-driven approaches where future sampling locations are determined by past sampling locations and observations are referred to as *active learning* methods. Active learning has been successfully applied to standard problems in statistical inference and machine learning, however there has been very little work aimed at harnessing the power of active learning for designing efficient sampling paths for mobile sensing networks [30, 73]. We use active learning methods to design adaptive paths for mobile sensors that achieve minimax efficiency of field reconstruction and latency for environmental monitoring type applications. When the environmental phenomenon exhibits localized features, for example a sharp level transition, the proposed

adaptive path algorithms significantly outperform traditional non-adaptive sampling path strategies. It is shown that for reconstruction of a $d \geq 2$ dimensional piecewise constant field to a desired mean square error (MSE) of ϵ, active path designs require the mobile sensor to cover a pathlength $\propto \epsilon^{-\frac{d-1}{d}}$ and result in a latency that scales as $\epsilon^{-(d-1)}$. Notice that both the pathlength and latency increase as the desired error level is decreased, as one would expect. On the other hand, a non-adaptive uniform sampling path scheme would require the sensor to move over a pathlength $\propto \epsilon^{-1}$ and result in latency that scales as ϵ^{-d}. In the non-adaptive case, the pathlength and latency both increase much more rapidly as the error level is decreased. In fact, adaptive path planning provides a significant reduction in pathlength and latency (by factors of $\epsilon^{\frac{1}{d}}$ and ϵ, respectively), that translates directly to network resource savings or better capabilities for tracking the environmental phenomena of interest. Also notice that the rate exponent gain is equivalent to a dimensionality reduction of the estimation task. Since the features of interest are located in a $d-1$ dimensional subset of the original domain, one can see that active learning results in paths that adapt to the effective dimension of the data. Even more dramatic improvements are possible in one-dimensional scenarios (e.g., vertical measurements through a forest canopy or marine environment).

The path planning strategies developed here can be used for applications like the NIMS (Networked Infomechanical Systems) architecture [20,30]. The NIMS system uses controlled mobility of sensors for spatio-temporal sampling of various environmental variables like solar radiation intensity, atmospheric water vapor, temperature, and chemical composition over the vast expanse of a forest. The sampling paths of these sensors need to be designed so that the scale of motion matches the environmental monitoring needs in terms of accuracy, fidelity and tracking capabilities. Many of the phenomena of interest exhibit changepoints, for example, the solar radiation intensity map exhibits edges due to shadows cast by the vegetation. Further, measurements of different environmental variables require diverse sensors with different sampling duration and velocity requirements. For example, the sampling duration required to achieve a certain SNR (Signal-to-Noise Ratio) may vary from a few

seconds for solar intensity measurements to a few minutes for CO_2 sampling. We present path planning algorithms that accommodate this wide range of sensing capabilities.

Another potential application arises in the North Temperate Lakes Long Term Ecological Research program in Wisconsin [74]. This project involves a sensor network system equipped on boats to take measurements of temperature, water currents, dissolved gases, algae and other biological species concentrations across the lake and at various depths. Continuous monitoring of these phenomena and change detection requires the boats to sample the 3D transect in the most efficient manner. The results presented in this chapter provide guidelines for planning the path, velocity and number of boats required to achieve a certain desired accuracy of measurement in the least time. To verify the effectiveness of our methods and theoretical results, we present an illustrative example that uses adaptive and passive sensor mobility to estimate the water current velocity map for a freshwater lake in Wisconsin.

Some other applications that can benefit from fast spatial surveying adapted to the regions of interest include landscape scanning using sensor equipped aerial vehicles, estimating boundary of an oil spill using aerial or buoy sensors, prediction of weather patterns by monitoring atmospheric pressure, and geological fault-line detection [75].

While several research papers have explored the use of mobile sensors and proposed a variety of path planning algorithms (e.g., game-theoretic approaches, pursuit evasion, sensor exposure etc. [76–78]), the research has mainly being focused on target detection and tracking. We develop path planning algorithms for mobile sensors used in environmental monitoring type applications that involve mapping of a spatially distributed field. We leverage the benefits of active learning methods to design feedback-driven adaptive sensing paths that achieve minimax optimal efficiency in terms of mean square error distortion and latency under varying sensor capabilities (velocity and sampling frequencies). A theoretical framework is provided for the analysis of tradeoffs in latency, accuracy and path lengths. We also investigate cooperative strategies for networked systems of multiple mobile sensors or combination of sensors with and without mobility. It is shown that the advantages of task

distribution under cooperative strategies can also be attained while requiring minimal coordination between sensors. Such a scheme is robust to sensor failures and degrades gracefully as one or more sensors turn faulty.

The chapter is organized as follows. Section 5.2 reviews the active learning methods developed for function estimation and associated theoretical results. Section 5.3 presents the application of active learning methods to adaptive path planning for a mobile sensing network and discusses the tradeoffs in accuracy, latency and path lengths. In Section 5.4 we present the illustrative simulation study of water current velocity in Lake Wingra that is located in Madison, Wisconsin, using both passive and adaptive sensing paths. We conclude in Section 5.5.

5.2 Active Learning for Spatial Mapping

Consider the task of spatial mapping of an environmental variable by mobile sensors that sample the given transect at n locations $\{X_i\}_{i=1}^n$. Denote the observed (scalar) sample values by $\{Y_i\}_{i=1}^n$, that are assumed to obey the measurement model:

$$Y_i = f^*(X_i) + W_i \qquad i \in \{1, \ldots, n\},$$

where the function f^* is the field of interest and the sensor measurement noise W_i is characterized as iid random variables that are independent of the sample locations $\{X_i\}_{i=1}^n$. The task is to reconstruct a map of the spatial field f^* from sample locations and noisy observations $\{X_i, Y_i\}_{i=1}^n$. Classical (passive) sampling techniques consider uniform deterministic or random sampling locations that are specified prior to gathering the observations. Active learning methods, on the other hand, select the future sampling locations online based on past locations and observations, i.e., X_i depends deterministically or randomly on the past sample locations and observations $\{X_j, Y_j\}_{j=1}^{i-1}$ [4, 5, 75, 79–81].

While the concept of active learning is not new, there is very little theoretical evidence to support the effectiveness of active learning, and existing theories often hinge on restrictive assumptions that are questionable in practice. However, there are a handful of key results

that can be leveraged to aid in the design of mobile wireless sensing systems. Burnashev and Zigangirov [79] investigated the problem of changepoint detection in a 1-d function using sequential sampling. They show that the error in the location of the changepoint decays exponentially in n, the number of samples, as opposed to n^{-1} for non-adaptive sampling. More recently, Castro et al. [4, 5] investigate the fundamental limits of active learning for various nonparametric function classes including spatially homogeneous Hölder smooth functions, and piecewise constant functions that are constant except on a $d-1$ dimensional boundary set or discontinuity embedded in the d-dimensional function domain. They reveal that significantly faster rates of convergence, in the minimax sense, are achievable using active learning in cases involving a function whose complexity is highly concentrated in small regions of space. In this section we review the active learning methods presented in [79] and [4, 5].

In [79], Burnashev and Zigangirov address the problem of estimating the confidence interval of an unknown parameter from n controlled observations. This problem is equivalent to estimating the changepoint in a 1-d function by adaptive sampling. The problem can be stated formally as follows. Consider the class of functions

$$\mathcal{F}^* = \{f^* : [0,1] \to \mathbb{R} | f^*(x) = 1_{[0,\theta)}(x)\}$$

where $\theta \in [0,1]$. The goal is to estimate the changepoint θ from observations $\{Y_i\}_{i=1}^n$, where

$$Y_i = \begin{cases} f^*(X_i) & \text{with probability } 1-p \\ 1-f^*(X_i) & \text{with probability } p \end{cases} = f^*(X_i) \oplus W_i.$$

Here $\oplus$ indicates a sum *modulo* 2 and W_i represents Bernoulli noise. Clearly, if there is no noise ($p = 0$) it is easy to design an estimator $\hat{\theta}_n$ using binary bisection that attains exponential error probability i.e. $|\theta - \hat{\theta}_n| = O(2^{-n})$. Burnashev and Zigangirov show that even when $p > 0$, a similar probabilistic bisection approach can be used to get $\mathbb{E}[|\hat{\theta}_n - \theta|] = O(2^{-n})$. Notice that this adaptive sampling rate is much faster than the passive rate of $O(n^{-1})$ (under both the noiseless and noisy scenarios). The probabilistic bisection method is based on a bayesian posterior update. A Bayesian posterior update method that does not restrict the noise W_i to be Bernoulli was given by [3] and is presented in Figure 5.1. In

---

PROBABILISTIC BISECTION SAMPLING

1. *Initialization*: Assume uniform prior for θ.

$$p_\theta^0(x) = \text{uniform}([0,1])$$

2. *Repeat* i = 1,. . . ,n

   Sample selection: The sample location x_i is selected to be the median of the distribution of θ.

$$x_i : \int_0^{x_i} p_\theta^{i-1}(x)dx = 1/2$$

   Record noisy observation $Y_i = f^*(x_i) \oplus W_i$.

   Posterior update: Update the distribution of θ based on the observed sample Y_i at location X_i according to Bayes rule.

$$p_\theta^i(x) = \frac{p_{Y_i}(y|\theta = x)p_\theta^{i-1}(x)}{p_{Y_i}(y)},$$

   where $y \in \{0,1\}$.

3. *Estimator*:

$$\hat{\theta}_n = \arg\max_{x\in[0,1]} p_\theta^n(x).$$

Figure 5.1 Probabilistic bisection sampling strategy (Horstein [3])

practice, a discretized distribution for θ (e.g. a histogram form) is considered, as described in [79]. The next sampling location is randomly chosen to be either end point of the interval in which the median lies by flipping a coin with head/tail probability that ensures the average value is the actual median. A corresponding modification is required in the distribution update, refer to [79] for details. In the next section, we present an illustrative example using this method to design a mobile sensor's path assuming sensor readings contaminated with Gaussian noise (measurement error) and Bernoulli noise (sensor fault).

For multidimensional settings, Castro et al. [4, 5] developed an active learning method based on dyadic partitioning that provides near-minimax optimal error convergence for certain classes of nonparametric multivariate functions, and assuming the noise satisfies certain moment conditions. One of the main results of that work shows that any piecewise constant d-dimensional function separated by a d-1 dimensional boundary can be accurately estimated through adaptive sampling methods using far fewer samples than required by conventional non-adaptive sampling. For $d = 1$ they obtain the same exponential rates as shown by Burnashev. Their main results (for $d \geq 2$) can be summarized in the following three theorems:

Set up: An estimation strategy consists of a pair $(\hat{f}_n, S_n)$, where $\hat{f}_n$ denotes an estimator of the d-dimensional function f^* using n noisy observations $\{Y_i\}_{i=1}^n$, and S_n denotes a sampling strategy for choosing the sampling locations $\{X_i\}_{i=1}^n$. Let Θ_{active} denote the set of all active estimation strategies.

Theorem 10 (Theorems 3 and 8, Castro et al. [5])**.** *Let $H(\alpha)$ denote the class of spatially homogenous Hölder-α smooth functions*[1]*, then*

$$\inf_{(\hat{f}_n, S_n) \in \Theta_{active}} \sup_{f^* \in H(\alpha)} \mathbb{E}_{f^*, S_n}[||\hat{f}_n - f^*||^2] \asymp n^{-\frac{2\alpha}{2\alpha+d}}.$$

The notation $a_n \asymp b_n$ means that $a_n = O(b_n)$ and $b_n = O(a_n)$, ignoring logarithmic factors. This rate is the same as the minimax learning rate with passive methods, hence for

[1]The function has $\lfloor \alpha \rfloor$ continuous derivatives, where $\lfloor \alpha \rfloor$ is the maximal integer $< \alpha$, and the function can be well approximated by degree-$\lfloor \alpha \rfloor$ Taylor polynomial approximation. A formal definition is given in Chapter 3.

this class of smooth functions that do not contain highly localized features, both passive and active methods perform equally well in terms of rate of error convergence.

Theorem 11 (Theorems 5 and 12, Castro et al. [5]). *Let PC denote the class of piecewise constant functions with d-1 dimensional boundaries, then*

$$\inf_{(\hat{f}_n, S_n) \in \Theta_{active}} \sup_{f^* \in PC} \mathbb{E}_{f^*, S_n}[||\hat{f}_n - f^*||^2] \asymp n^{-\frac{1}{d-1}}.$$

Thus the error rate is determined by the effective dimension of the function. On the other hand, in this situation the non-adaptive sampling error decays like $n^{-1/d}$, and thus we see that active methods can lead to significant improvements when the target function contains localized features.

Theorem 12 (Theorem 6, Castro et al. [5]). *Let $PS(\alpha)$ denote the class of more general piecewise Hölder-α smooth functions with d-1 dimensional boundaries, then*

$$\inf_{(\hat{f}_n, S_n) \in \Theta_{active}} \sup_{f^* \in PS(\alpha)} \mathbb{E}_{f^*, S_n}[||\hat{f}_n - f^*||^2] \geq c \max\{n^{-\frac{2\alpha}{2\alpha+d}}, n^{-\frac{1}{d-1}}\}.$$

where $c > 0$ is a constant.

It is conjectured in [5] that the above result should characterize the optimal rate, that is, an upper bound matching the minimax lower bound should exist, though this is not formally proved. This theorem follows directly from the previous two and essentially states that the rate of convergence is determined by the dominating cause of function complexity - dimensionality of the edge or complexity of function derivatives away from the boundary.

Castro et al. also present an active learning algorithm that nearly achieves these minimax rates; the exact performance bound on the MSE of the proposed algorithm is discussed later. In Figure 5.2, we first review the strategy proposed in [4, 5] for piecewise constant functions. This procedure produces a non-uniform sampling pattern in which more samples are concentrated near boundaries in the field; half of the samples are focused in a small region about the boundary. Since accurately estimating f^* near the boundary set is key to

MULTISCALE ADAPTIVE SAMPLING

1. *Preview Step*: $n/2$ samples are collected at uniformly spaced points and a coarse estimate of the field f^* is generated from these data using a complexity-regularized tree pruning procedure. This provides a rough indication of where boundaries may exist.
2. *Refinement Step*: $n/2$ additional samples are collected at points in regions near the rough location of boundaries detected in the Preview Step. A similar tree pruning procedure is used in these regions to obtained refined estimates in the vicinity of boundaries.
3. *Fusion Step*: The estimates from Preview and Refinement Steps are combined to produce an overall estimate.

Figure 5.2 Multiscale adaptive sampling strategy, Castro et al. [4, 5]

obtaining faster rates, we expect such a strategy to outperform a strategy based on passive sampling.

The estimators used in the first two steps are built over recursive dyadic partitions (RDPs) of the function domain. An RDP can be identified with a tree structure, where each leaf of the tree corresponds to a cell of the dyadic partition (see Figure 4.1 in Chapter 4). An estimate is generated on each leaf of the RDP by averaging the observations collected in each leaf. If the field is Hölder-α smooth on either side of the boundary, the estimator consists of a least square degree-$\lfloor\alpha\rfloor$ polynomial fit to the observations [5] rather than simple averaging, as for piecewise constant field. If we consider Π to be the set of all RDPs that can be generated on the input domain, the best RDP is chosen according to the following complexity-regularized estimation rule:

$$\hat{\pi} = \arg\min_{\pi\in\Pi} \left\{ \sum_{i=1}^{n} (Y_i - \hat{f}_n^{(\pi)}(X_i))^2 + \lambda(|\pi|) \right\}, \tag{5.1}$$

where $\lambda(\cdot)$ is a penalty term that depends on $|\pi|$, the number of leaves in RDP π and $\hat{f}_n^{(\pi)}$ is the estimator containing average of all observations in each leaf of the RDP. The final

estimate for each step, $\hat{f}_n$ is given by

$$\hat{f}_n = \hat{f}_n^{(\hat{\pi})}. \tag{5.2}$$

The computation of $\hat{f}_n$ can be done efficiently using bottom-up tree pruning algorithms [48]. In the Preview Step, leaves that are not pruned back indicate regions of the field where further averaging (through pruning and aggregation) would have led to large data fitting errors. Thus, these leaves indicate regions that probably contain boundaries or other sharply varying characteristics of the field, and these are the focus regions for sampling in the Refinement Step. There are a few additional subtleties involved in the procedure [5], but this gives one the main idea of the method.

The MSE of $\hat{f}_n$ can be shown to decay like $n^{-1/(d-1+1/d)}$ for piecewise constant functions, much faster than the best passive rate of $n^{-1/d}$. Furthermore, performing repetitive refinement steps leads to an improved rate of $n^{-1/(d-1+\delta)}$, where $0 < \delta \leq 1/d$ decreases with the number of refinements. Thus, this method can be arbitrarily close to the minimax rate of Theorem 11.

We use this algorithm to plan the path of a current velocity sensor equipped boat and generate a spatial map of the water current velocity profile for a freshwater lake in Wisconsin. Simulation results reveal that huge savings in time can be achieved for reconstructing the spatial map to same accuracy as a passive path. In the next section we show how the active learning methods discussed here can be used to design fast adaptive spatial survey paths for mobile sensing networks and explore the tradeoffs in path length, accuracy and latency under varying sensor capabilities.

5.3 Adaptive Sampling Paths for Mobile Sensing

We now apply the active learning methods discussed in the last section to design efficient adaptive paths for a mobile sensing network that is required to generate a high fidelity, fine resolution estimate of a spatially varying environmental phenomenon.

5.3.1 Single mobile sensor

We start with the simplest case of a single mobile sensor that needs to efficiently sample a 1-d field containing a change point. This situation is well-suited, for example, to a NIMS [20] like sensor suspended on cableways that measures solar radiation intensity within the forest transect with the shadow cast by an object constituting a changepoint. We model a toy example (Figure 5.3) for estimating this 1-d function under two scenarios:

I. **Measurement noise** - Typically the sensor measurements are noisy due to environmental fluctuations and slight uncertainties in sensor readings. Such noise can be modeled as a Gaussian random variable. Figure 5.3 shows the sampling locations and the corresponding noisy observations made by the mobile sensor with (a) adaptive sampling using probabilistic bisection and (b) passive sampling. In the adaptive case, though the sensor takes longer excursions for the initial samples, it quickly "homes-in" on the more interesting feature of the field concentrating most of its measurements around the changepoint with only a few samples where the field is constant. For a 100 m transect containing a changepoint at 70 m that needs to be mapped to a resolution of 0.1 m and assuming noise variance of 0.1, the number of samples required in the adaptive scheme are exponentially less - only about 10 as opposed to 700 samples for passive case. A NIMS sensor typically moves at 1 m/s and takes 1 sec to record a sample of the solar intensity, which implies that an adaptive mobile sensor would take 2 mins instead of 13 mins to accomplish the task.

II. **Sensor fault** - We model this case by assuming the noise W_i is a Bernoulli(p) random variable with p denoting the probability that either the sensor fails to record a reading or the reading is corrupted and has to be discarded. We simulate the same example as before for Bernoulli noise with $p = 0.2$. It was observed that 20 samples were required with the adaptive path to achieve the desired resolution of 0.1 m, as opposed to nearly 700 for non-adaptive path.

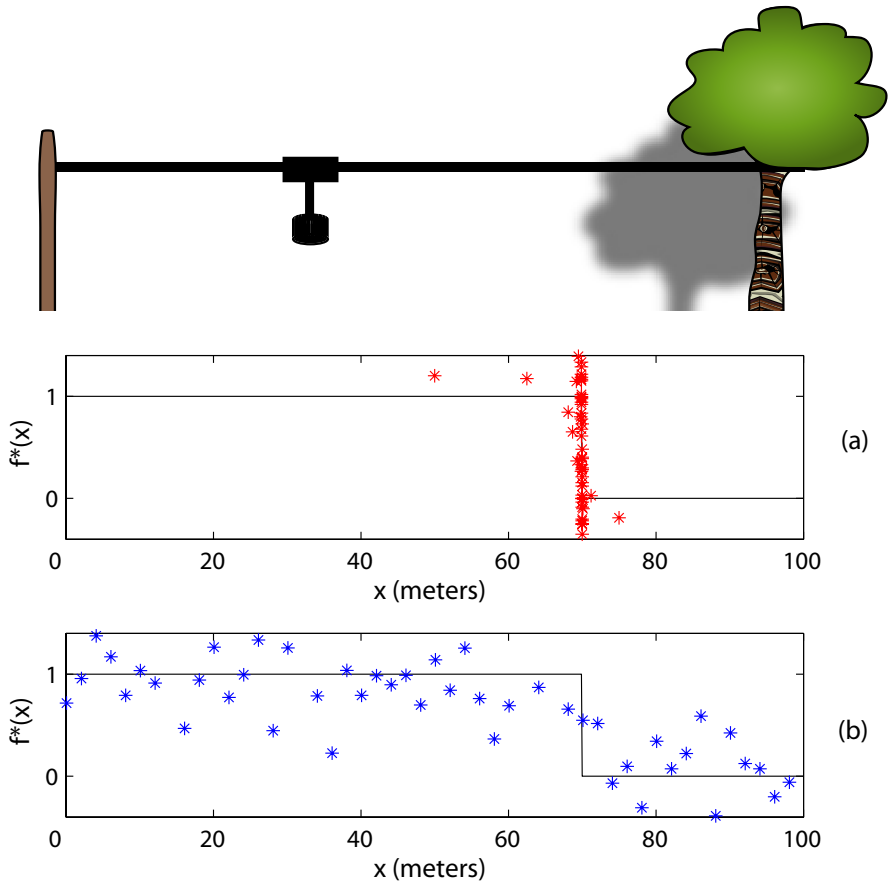


Figure 5.3 A toy example of estimating a 1-d solar intensity map containing a single changepoint at 70 m due to a shadow, using a NIMS type mobile sensor. The sample locations and observations are shown in (a) for adaptive survey and (b) for passive survey.

For a 2-d transect, the multiscale adaptive sampling technique developed by Castro et al. [4, 5] can be used for designing adaptive paths for a mobile sensor. In the following, we first translate their algorithm and analysis to the mobile sensing setting. This will provide a characterization of how the path length and latency incurred by a mobile sensor scale with a desired fidelity (MSE). Then, we discuss the implications of these results under varying sensor capabilities. We start with the case of a single mobile sensor and later extend the result to multiple mobile sensing network and a network system comprised of heterogeneous sensors with and without mobility.

The design of the sensing path for a single mobile sensor can be described as follows. For simplicity, we assume that the coordinates are scaled so that the area to be monitored can be described as a unit square or hypercube $[0, 1]^d$. Also, we only describe the case where the field is piecewise constant, i.e. the variable of interest exhibits a level transition taking on constant values on either side of the boundary. If the field is not constant, but say Hölder-α smooth on either side of the boundary, the path planning is same as described for the piecewise constant case, except that the estimator involves a least square degree-$\lfloor\alpha\rfloor$ polynomial fit to the observations rather than simple averaging.

1. *Coarse survey* - With no prior knowledge of the field, the mobile sensor starts by doing a coarse passive survey of the field in a raster scan fashion. The sensor collects $n/2$ samples at regular intervals along the coarse survey path of length ℓ. A complexity penalized estimate $\hat{f}^c$ is constructed using the $n/2$ samples over recursive dyadic partitions of the field according to equations (5.1) and (5.2), with penalty $\lambda(|\pi|) = C\sigma^2\log(n/2)|\pi|$, that is proportional to the number of leaves $|\pi|$ and sensor measurement noise σ^2. $C > 0$ is a constant that depends on the dimension d and smoothness of the field [5]. Notice that this penalizes RDPs with fine partitions and leaves at maximum depth are retained only if pruning by averaging the observations would lead to large data fitting errors. The estimator averages out the noise where the field is smooth (reduces the variance where bias is low), while performing no averaging

on the samples around the boundary (where bias is high). This yields an RDP estimate of the field with leaves at the greatest depth J providing a rough location of the boundary:

$$\mathbb{B} \quad = \quad \{\text{Regions where the coarse estimate contains leaves at maximum depth}\}$$

Since at each depth in the tree, the sidelength of cells is halved, at maximum depth J the leaves have sidelength 2^{-J}, which is equal to the finest resolution of $1/\ell$ provided by a pathlength of ℓ. The volume at this maximum depth is ℓ^{-d}, which implies that the maximum possible number of leaves at this depth is ℓ^d. Of these, only $O(\ell^{d-1})$ intersect the boundary since the boundary occupies a d-1 dimensional subspace in the monitored region. Thus, $|\mathbb{B}| = O(\ell^{d-1})$.

2. *Refinement pass* - The mobile sensor is now guided along the regions identified as containing the boundary in the coarse survey (set $\mathbb{B}$) to collect an additional $n/2$ samples in these regions, again traversing a pathlength ℓ. A complexity penalized estimator $\hat{f}^r$ is generated on each region in the set $\mathbb{B}$, similar to the estimator built on the entire area in the coarse survey. This confines the boundary location (and hence bias) to even smaller regions, while averaging out the noise and lowering the variance in the smooth regions of set $\mathbb{B}$.

3. *Field reconstruction* - A final estimate of the field is now generated by fusing the estimates obtained in the coarse survey and the refinement pass as follows:

$$\hat{f}_{active}(x) = \begin{cases} \hat{f}^r(x) & \text{if } x \in \mathbb{B} \\ \hat{f}^c(x) & \text{otherwise} \end{cases}$$

Figure 5.4 shows the two stages of the adaptive mobile sensing path. The advantage of a two-step method is that the computation involved can be done either on the mobile platform or if enough processing power is not available on board the mobile, the data can be dumped to a fusion center after each pass where processing and design of future paths may be carried out.

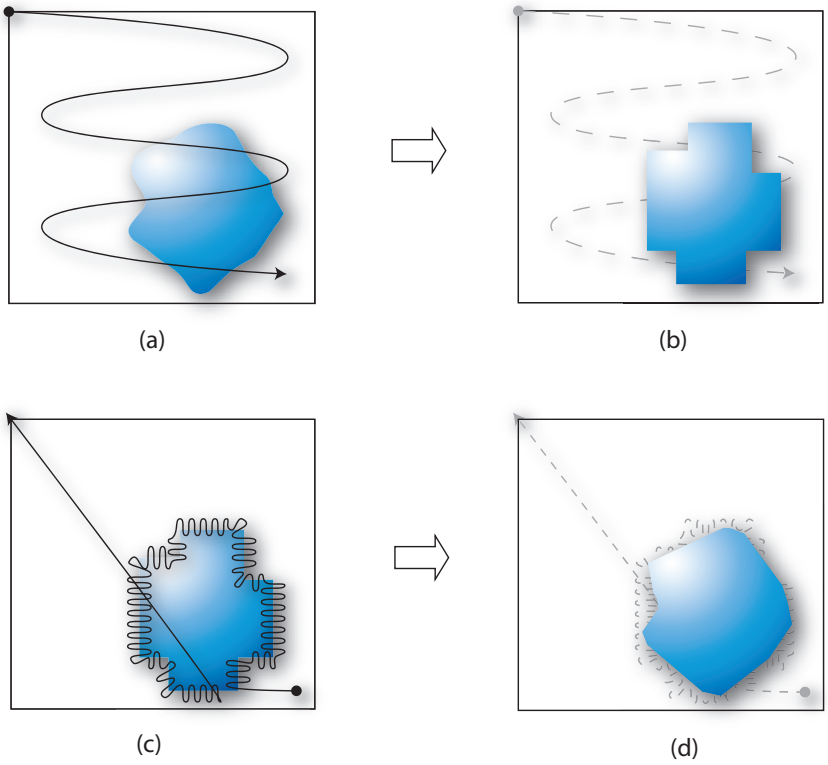


Figure 5.4 Multiscale adaptive path of a mobile sensor for mapping a field containing a boundary. In the first step, the mobile sensor follows a coarse survey path (a) and produces a rough estimate of the field (b). In the refinement pass (c), the mobile follows a path adapted to the interesting regions of the field and produces a fine resolution estimate (d).

The mean square error (MSE) of the estimator can be upper bounded using the Craig-Bernstein inequality as shown in [4, 5], and for piecewise constant fields this yields:

$$\mathbb{E}[||\hat{f}_{active} - f^*||^2] \leq C_1 \left(\frac{\log n'}{n'}\right)^{1/d} 2^{-J} + C_2 2^{J(d-1)} \frac{\log n}{n}$$

where J denotes the maximum depth of the RDP tree estimator in the coarse step and $n' = n/(2|\mathbb{B}|)$ denotes the number of samples in each preview leaf collected in the refinement pass and $C_1, C_2 > 0$ are constants. Since $1/\ell = 2^{-J}$ as discussed before, in terms of pathlength the above bound can be expressed as:

$$MSE \leq C_1 \left(\frac{\log n'}{n'}\right)^{1/d} \ell^{-1} + C_2 \ell^{d-1} \frac{\log n}{n} \tag{5.3}$$

where $n' = n/(2\ell^{(d-1)})$, $n/2$ samples distributed over $|\mathbb{B}| = O(\ell^{d-1})$ leaves that contain rough location of the boundary. The first term denotes the error incurred in the refinement pass in regions close to the boundary and the second term denotes the error incurred in the coarse survey in regions away from the boundary. The two terms can be balanced by appropriately choosing the number of samples and path length:

$$\ell \;\sim\; O\left(n^{\frac{d-1}{(d-1)^2+d}}\right), \tag{5.4}$$

where $\sim$ denotes polynomial order dependence, dropping any log terms. This represents the classical bias-variance tradeoff since the length ℓ primarily determines the bias in estimating the boundary location, while the number of samples n determines the variance. If a spatial map with a MSE accuracy ϵ is desired, then by setting the MSE bound in (5.3) to ϵ and combining with (5.4), we find that the minimum number of samples required and the optimal sensing path length would scale with the desired accuracy as:

$$n_{opt} \sim O(\epsilon^{-(d-1+\frac{1}{d})})$$

and,

$$\ell_{opt} \sim O(\epsilon^{-\frac{d-1}{d}}),$$

where the latter expression is obtained by substituting n_{opt} into (5.4). This implies that for a 2-d field, as in the water current mapping in a lake, increasing the accuracy by a factor

of 10 would require the mobile sensor to traverse roughly 3 times longer pathlength and collect about 30 times more samples. This is a significant improvement over the passive approach, where the mobile would have to cover 10 times longer path and collect 100 times more samples.

If the mobile sensor moves at a velocity v and requires time T to record one sample of the environmental variable to a desired Signal to Noise Ratio (SNR), the total time required to generate the estimate is given by

$$t = nT + \frac{\ell}{v}$$

Thus for a fixed sampling time T and fixed velocity v,

$$\begin{aligned} t &\sim O(\max\{n_{opt}, l_{opt}\}) \\ &\sim O(\epsilon^{-(d-1+\frac{1}{d})}) \end{aligned}$$

for all $0 \leq \epsilon < 1$.

If the mobile sensor was to move along a passive path, the time required to achieve the same accuracy would be $O(\epsilon^{-d})$, and the mobile would need to traverse a length of $O(\epsilon^{-1})$. Thus, the improvement provided by adaptive path planning is essentially equivalent to a dimensionality reduction of the field estimation problem. This shows that active learning adapts to the complexity of the problem, since the effective dimension of the field in the piecewise constant case is the dimension of the boundary (d-1). This represents a huge reduction in latency and pathlengths, that directly translates to network resource savings as well as allows for more accurate tracking of the temporal changes in the spatial map of the environmental variable.

In practice, sensors for measuring different environmental variables may have sampling time anywhere from a few seconds for measuring water current velocity to few minutes for measuring CO_2 level in a forest canopy. Also sensors might be restricted in mobility to a certain maximum velocity. For example, NIMS type sensors that are suspended on cableways are limited to a maximum velocity of about 1 m/s, on the other hand sensors mounted on

aerial vehicles can move at hundreds of m/s. Thus for a desired small accuracy of ϵ, we identify two distinct regimes of operation based on sensor capabilities.

I. *Sampling time constrained* - If the sensor mobility is constrained by the sampling time, the latency incurred is primarily due to the time spent in recording samples. This represents the **variance limited** regime since the noise in sensor measurements limits achievable accuracy, while the sensor can move at high enough velocity to cover the path length required to reduce bias. In this situation the latency scales according to

$$t \sim O(n_{opt}) \sim O(\epsilon^{-(d-1+\frac{1}{d})})$$

Balancing the two terms contributing to latency ($nT = \ell/v$), we find that the minimum velocity required for a sampling time constrained sensor scales like

$$v_{opt} \sim O(\epsilon^{d-2+\frac{2}{d}}).$$

The sensor can move at any velocity greater than or equal to the optimal velocity. However if the sensor's velocity is limited to less than v_{opt}, we are in the *velocity constrained* regime.

II. *Velocity constrained* - If the sensor mobility is velocity limited (i.e., if the platform is not capable of moving at the minimum required velocity above), then the time to move from one sampling location to the other is the primary cause of latency. This corresponds to the **bias limited** regime since enough samples can be collected to reduce the variance, while the path length that the sensor can cover and hence reduction in bias, is limited. In this regime the latency scales like

$$t \sim O(\ell_{opt}) \sim O(\epsilon^{-(\frac{d-1}{d})})$$

Again this rate is faster than that for passive paths $O(\epsilon^{-1})$, leading to time savings of the same order as in the variance limited regime where the sensor is sampling time constrained.

Notice that even for the **noise-free case** (i.e. when the sensor reading has high enough SNR), latency scales with pathlength since the variance is negligible (ideally zero) and the error is completely due to bias in locating the boundary. In this case, no error is incurred in cells away from the boundary, and the pathlength determines the resolution or bias. Thus the adaptive rates for bias-limited regime also characterize the noise-free case and one sees that multiscale adaptive scheme performs better than passive whether noise is present or not.

It should be noted that if the environmental variable does not exhibit much variability and the spatial map is smooth, the path planning strategies described above will yield performance similar to a uniform passive scan of the field. Thus active learning paths are data adaptive, resulting in shorter localized paths if the interesting phenomena are confined to certain regions of space and longer uniform paths if the phenomena are spatially well distributed.

5.3.2 Mobile Sensing Network

So far we have developed strategies to design adaptive paths for a single mobile sensor. Now we consider a network of mobile sensors. Intuitively, it is clear that if there are k mobile sensors, the field estimation task can be distributed amongst them leading to reduced latency (by a factor of k). However, the way the task distribution is done can have significant impact on the robustness of the system. For example, an obvious approach is to divide the field into k equal regions and require each mobile to perform field estimation on one region. Assuming, for example, that the mobile sensors have a fixed sampling time and are moving at the optimal velocity, the MSE distortion is reduced by a factor of $k^{-1/(d-1+\frac{1}{d})}$ for a desired latency, or equivalently, for a given distortion it yields a factor k improvement in latency.

However, this approach is clearly not robust to sensor failure. We wish to devise a path planning strategy that would degrade gracefully with node failures. Certainly, as long as there is at least one working sensor, the error should be no larger than the error in the single mobile sensor case. A cooperative strategy needs to be followed where sensors coordinate

their movements so that even with sensor failures there is adequate sampling of the entire field, though at a reduced resolution. We outline such a collaborative strategy in Figure 5.5. If the failures are random, with p the probability of a sensor failure, it can be shown that under this cooperative strategy the distortion will be reduced by a factor of $(pk)^{-1/(d-1+\frac{1}{d})}$, with pk reflecting the effective number of active sensors. Thus this cooperative strategy is robust and degrades gracefully with sensor failures. The coordination requirements of this scheme during the data collection process can actually be reduced. It can be shown that if the mobile sensors start at uniformly random locations and collect uniformly random samples independent of other sensors in each pass, we can still guarantee distortion improvement by the same factor, provided the estimates in each step are formed using collective measurements and the fusion center transmits the rough location of the boundary to all sensors after the first step.

It is envisioned that in many cases mobile sensors will be used to complement a static network of sensors. In this case, the static sensor network can be deployed uniformly across the region of interest to form a low resolution estimate of the spatial map. This would provide continuous monitoring of the environmental variable, and if an event of interest occurs the mobile sensor(s) can be dispatched to collect guided measurements and form a refined estimate of the changepoints.

5.4 Case Study: Lake Mapping

In this section, we use the path planning strategies presented above for a real-life problem of water current mapping in a freshwater lake. The Lake Ecology research group in Wisconsin [74] is interested in studying hydrodynamics in lakes and how the biophysical setting of the lake influences these dynamics. One of the primary requirements for such studies is the spatial mapping of the water current velocity in the lake. Currently, such measurements are taken by a sensor called Acoustic Doppler Current Profiler mounted on a boat that moves around the lake taking samples at regular intervals. However, the time requirements for producing a spatial map of the entire lake have restricted the researchers to base their inference

---

COOPERATIVE STRATEGY FOR k MOBILE SENSORS

1. *Initialization* - The mobile sensors start at uniformly spaced locations in the field. This provides robustness if sensors are likely to fail with progression in time.

2. *Coarse survey* - The mobile sensors survey the whole field in a raster scan fashion moving along paths that are interleaved such that spacing between adjacent mobiles' paths is $1/\ell$, where ℓ is the pathlength a sensor would traverse if acting alone (see Fig. 5.4). Notice that rows of each raster scan are now k/ℓ apart since there are k sensors. Thus, each mobile covers a k times shorter (and hence less dense) path and collects n/k measurements. The measurements are transmitted to a fusion center where a coarse RDP estimate of the field is constructed using measurements from all k mobiles. The rough location of the boundary is then conveyed back to all the mobiles.

3. *Refinement pass* - The mobile sensors move along coordinated paths, similar to the coarse survey (each path is again widely spaced by a factor of k), along the regions identified as possibly containing the boundary to collect additional n/k measurements each. The final refined estimate is then generated using the collective measurements.

Figure 5.5 Cooperative strategy for k mobile sensors

Figure 5.6 Simulated low resolution water current velocity profile in Lake Wingra. Notice there are two distinct regions characterized by low or high velocity, with significant gradient between them.

on field measurements collected from a small part of the lake, and at coarse resolutions. The techniques presented in this chapter for designing adaptive paths that yield low latency, high resolution spatial map can greatly benefit such research efforts.

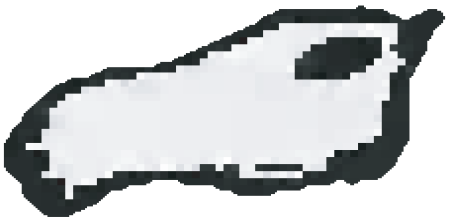
The water current velocity in a lake is influenced by many factors including the lake bathymetry (depth profile), littoral-zone vegetation, stratification induced by temperature and the wind profile generated by the lake surroundings. The circulation velocity map is shown in Figure 5.6 for a freshwater lake. The map exhibits two distinct regions with significant velocity gradient between the two. Hence, the field can be characterized as a level transition with an irregular boundary. This situation is ideally suited for the multiscale adaptive sampling based path planning approach developed in the last section. We investigate the performance of the multiscale adaptive approach against a passive approach by designing the path for the sensor-carrying boat using both strategies and comparing the accuracies obtained for the reconstructed maps.

To apply our algorithms to the lake model, we consider a square region around the lake with the values at the edges of the lake interpolated to fill the square area. This is necessary to prevent the algorithm from focusing samples along the boundary of the lake itself, rather than the boundary of the velocity gradient that we really seek. We use the data from a three-dimensional non-hydrostatic and stratified flow model (3DNHYS) [82] for our experiments. For simplicity we model the lake with a piecewise constant function. In general, platelet or polynomial fits can be used to approximate the velocity on either side of the boundary. The

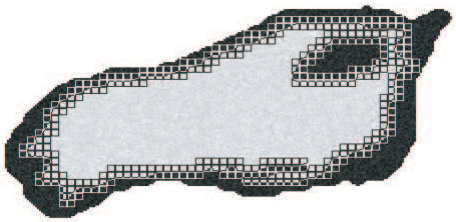
platelet fit has been used in [83] for a network of static sensors and is shown to provide good approximation to a smooth field using the active learning method of [4].

To study the circulation pattern, spatial map of the water current velocity at a resolution of $\sim 10\,\mathrm{m}$ is required. The original discrete model data provided by the limnology group is for a $2\,\mathrm{km} \times 1\,\mathrm{km}$ lake with a resolution of $25\,\mathrm{m}$ between samples. We interpolate this data using nearest neighbor interpolation to get a sampling resolution of $25/3 = 8.3\,\mathrm{m}$. These samples can be thought of as the data collected using a sensor mounted on a boat that moves along a dense uniform passive path around the entire lake. Of the 256×256 samples in our square region, 37323 samples actually lie in the lake region and are meaningful to evaluate the accuracy, latency and pathlength of a boat recording these measurements. In the field experiments done by the lake research group, the sampling rate of data collection using the acoustic doppler current profile sensor is $2\,\mathrm{Hz}$ (i.e. $0.5\,\mathrm{secs/sample}$), and the boat speed is about $2\,\mathrm{m/s}$. This implies that the boat needs to cover an extensive pathlength of $310\,\mathrm{km}$ and takes nearly $48\,\mathrm{hrs}$ to collect all the samples. Analysis of the latency shows that only $5\,\mathrm{hrs}$ are spent in the measurement process, while most of the time is consumed in moving around the lake. Thus for this problem, we are in the "velocity-constrained" regime. The reconstructed estimate for the passive case is shown in Figure 5.7(d) with MSE $= 3 \times 10^{-4}$.

In the two-step adaptive approach, the boat makes an initial coarse survey collecting samples at $33.2\,\mathrm{m}$ resolution. Figure 5.7(a) shows the estimate obtained after the coarse survey and 5.7(b) shows the cells corresponding to the rough boundary locations (as indicated by the coarse estimator), superimposed on the noisy field. In the refinement pass, the boat is guided along the rough boundary location identified by the coarse survey to collect additional samples at finer resolution of $8.3\,\mathrm{m}$. The final reconstructed map is shown in 5.7(c). The adaptive approach achieves nearly the same accuracy as the passive with MSE $= 4 \times 10^{-4}$ and requires only 8077 samples in the lake. The pathlength is thus reduced by a factor of 3 to $92\,\mathrm{km}$ and the time reduced to nearly $14\,\mathrm{hrs}$. This represents huge savings in time, and makes the desired task much more feasible. Multiple sensors, if available, can be used to reduce the latency further, e.g. 5 such sensors would reduce the time to less than $3\,\mathrm{hrs}$.



(a) Estimate generated after the coarse survey in a 2-step adaptive path approach.



(b) Cells denoting rough location of the boundary as indicated by the coarse estimate, superimposed on the noisy field.

(c) Final estimate generated after the refinement pass over the cells in 5.7(b). MSE $= 4 \times 10^{-4}$, latency = 14 hrs.

(d) Estimate generated by a boat moving along a uniform passive path. MSE $= 3 \times 10^{-4}$, latency = 48 hrs.

Figure 5.7 Comparison of passive and adaptive path planning approaches for water current velocity mapping in a freshwater lake. The adaptive strategy requires only 14 hrs for mapping the nearly 2 km$\times$ 1 km lake to a resolution of < 10 m, as opposed to 48 hrs using the passive method.

5.5 Concluding Remarks

In this chapter, we proposed adaptive path planning approaches to design paths of mobile sensors used for mapping a spatially distributed environmental phenomenon. Based on recent advances in the active learning literature, fast spatial surveying methods are developed to guide the mobile sensors along interesting regions of the sensing domain. A theoretical framework is developed for evaluating the accuracy, pathlength and latency tradeoffs under varying sensor capabilities. It is shown that the proposed data-adaptive path planning procedures achieve significant improvements in latency and pathlength, for a given desired accuracy, over a non-adaptive raster-scan approach if the environmental phenomenon exhibits a sharp transition along a lower dimensional boundary. Application of the developed methods to water current mapping in a lake results in latency improvement from 48 hrs to 14 hrs.

The two-step path-planning scheme proposed in this chapter has provable optimal guarantees in terms of how the path length and latency scale with desired fidelity, and an advantage of a two-step method is that the computation involved can be done either on the mobile platform or if enough processing power is not available on board the mobile, the data can be dumped to a fusion center after each pass where processing and design of future paths may be carried out. However, in scenarios where there is enough on-board processing power on the mobile sensor, or the span of the field to be monitored is small so that the sensor can easily contact the base station at frequent intervals, path planning strategies that can modify the path in an online fashion as new samples are collected are more desirable. While some solutions to online path design are available in the robotics literature, these are either heuristic or place strong a priori assumptions on the field to be monitored. It is of interest to develop online path planning schemes under a nonparametric framework that can exploit optimal tradeoffs between pathlength, latency and fidelity.

Chapter 6

Summary and Future Directions

This thesis addresses some open theoretical questions and critical applications that involve statistical learning and inference using sets in the nonparametric framework. Specifically, it (a) extends the applicability of Hausdorff accurate set reconstruction that ensures a spatially uniform confidence interval around a region of interest, which is particularly relevant to problems that require robustness of the set estimate or statistical guarantees based on connectivity of sets. A data-adaptive method is proposed that can provide minimax optimal Hausdorff guarantees for density level set estimation over a more general class than considered in previous literature, without assuming a priori knowledge of any parameters and requiring only local regularity of the density in the vicinity of the desired level. This thesis also (b) provides a solid theoretical foundation for analyzing the semi-supervised learning setting where unlabeled data are used to learn the decision sets, and characterizes the relative value of labeled and unlabeled data. This analysis suggests that conventional evaluation tools such as minimax rates of error convergence or asymptotic analysis may be inadequate for characterizing certain situations, and reinforces the value of finite sample error bounds to quantify the performance of a learning algorithm. Finally, this thesis also contributes to two applications in neuroimaging and sensor networks. It (c) proposes a novel level set based approach to fMRI data analysis that can exploit structural information by adaptively aggregating neighboring voxels to boost detection of weak brain activity patterns and (d) suggests feedback-drive adaptive mobile sensing paths for environmental monitoring that focus on regions where the environmental phenomenon exhibits a transition, and characterizes

the optimal tradeoffs between path length, latency and fidelity under various mobile sensing capabilities.

The concluding remarks at the end of each chapter suggest some fine-grained extensions and open problems related to the work presented. I now elaborate on some of these and also provide some broader directions for future work.

Statistical Guarantees for Clustering : One of the main motivating applications for studying the Hausdorff metric was clustering. Since clustering is typically used as a first step in exploratory data analysis, it is desirable to develop a nonparametric distribution-free framework for clustering that places very mild assumptions on the data distribution. Identification of arbitrary shaped clusters relies on accurate recovery of the connectivity of constituent components, and hence Hausdorff accurate set reconstruction is particularly relevant. Observe that a Hausdorff control of ϵ between two sets ensures that the boundary of each set is contained within an ϵ-tube around the boundary of the other set. This follows from the following argument:

$$d_\infty(G_1, G_2) \leq \epsilon \quad \Longrightarrow \quad G_1 \subseteq G_2^\epsilon, \;\; G_2 \subseteq G_1^\epsilon,$$

where G^ϵ denotes the set obtained by dilation of set G by an ϵ-ball. This implies that Hausdorff control can preserve the shape and connectivity of connected components of a set up to erosion or dilation by ϵ. Thus, it offers the potential to provide statistical guarantees for clustering. Recently, the importance of distance between cluster boundaries to guarantee clustering stability was realized in [84]. Here we sketch an alternative approach to clustering and the associated statistical guarantees. We define a *cluster* as a connected component of the density at some level. Since it is not possible to detect arbitrarily small clusters based on a finite sample from the distribution, we pose the following problem: Detect all clusters containing at least $\nu > 0$ mass. In order to detect every cluster of mass $\geq \nu$, it suffices to select density levels $\Gamma = \{\gamma_k\}_{k=1}^K$ that are close enough so that the difference in mass between consecutive density level sets is less than ν, for example we may require that $|P(G^*_{\gamma_{k+1}}) - P(G^*_{\gamma_k})| \leq \nu/2$ for all k. This ensures that no ν-mass cluster will be missed and in fact, we will detect at least $m - \nu/2$ of the mass of any cluster with mass $m \geq \nu$.

Appropriate levels can be empirically selected based on control over discrepancy of true and empirical mass of sets, and the fact that Hausdorff control between two sets also guarantees control over the deviation of mass of the sets. As an example, we propose a bisection based density level search. The highest density level of interest $n\nu$ corresponds to ν mass present in the smallest resolvable volume of $1/n$, given n samples.

BISECTION LEVEL SEARCH

1. *Initialization:* $\Gamma = \{0, n\nu\}$
2. *Iterate:* Generate Hausdorff accurate level set estimates $\{\widehat{G}_\gamma\}_{\gamma\in\Gamma(i)}$.

   For $i = 1 : end - 1$

   If $\widehat{P}(\widehat{G}_{\Gamma(i)}) - \widehat{P}(\widehat{G}_{\Gamma(i+1)}) > \nu/4$, add $(\Gamma(i) + \Gamma(i+1))/2$ to Γ.

   If Γ does not change

   Stop.

   Else

   Sort Γ in ascending order.

   Repeat.

Theorem 13. *Consider the class of densities $\mathcal{F}_2^*(\alpha)$ as defined in Chapter 2. Let $\{\widehat{G}_{\gamma_k}\}_{k=1}^K$ be the level set estimates generated by the bisection search algorithm, then there exists $C > 0$ such that for all n with probability at least $1 - 3/n$*

$$|P(G^*_{\gamma_{k+1}}) - P(G^*_{\gamma_k})| \leq \nu/4 + \epsilon$$

where

$$\epsilon \leq C s_n^2 \left(\frac{n}{\log n}\right)^{-\min\left\{\frac{1}{d+2\alpha_2}, \frac{\alpha_1}{d+2\alpha_1}\right\}}$$

Here $\alpha_1 = \min_k \alpha_k, \alpha_2 = \max_k \alpha_k, C \equiv C(C_1, C_2, C_3, C_4, \epsilon_o, f_{\max}, \delta_0, d, \{\alpha_k\}_{k=1}^K)$.

Proof. See Appendix. □

So for sufficiently large n, the procedure selects density levels such that difference in mass between consecutive level sets is less than ν, and hence no ν-mass cluster is missed. The approach suggested here corresponds to attempting to recover the cluster-tree as introduced in [1], minus the branches that lead to clusters with mass less than ν.

The main statistical guarantees for clustering can now be stated as follows. Assume that the regularity of any density level can be characterized by $\alpha_k \in [\alpha_1, \alpha_2]$. Then there exist

$$\epsilon_1 \leq C s_n^2 \left(\frac{n}{\log n}\right)^{-\min\left\{\frac{1}{d+2\alpha_2}, \frac{\alpha_1}{d+2\alpha_1}\right\}} \quad \text{and} \quad \epsilon_2 \leq C s_n^2 \left(\frac{n}{\log n}\right)^{-\frac{1}{d+2\alpha_2}}$$

such that with high probability (at least $1 - 3/n$)

> *Detection* - The algorithm detects at least $m - \nu/4 - \epsilon_1$ of the mass of all clusters with mass $m \geq \nu$.
>
> *Estimation* - The estimated clusters preserve the shape and connectivity of corresponding true clusters up to erosion or dilation by an ϵ_2-ball.

As ϵ_1, ϵ_2 depends on α, we see that the regularity parameter also characterizes the *"discernability"* of clusters and hence the complexity of the clustering problem.

While this can provide a nice theoretical characterization of clustering that places very mild assumptions on the cluster shapes, devising a practically useful algorithm requires an efficient Hausdorff accurate set estimation method. As we discussed at the end of Chapter 2, spatially adapted partitions may result in better practical estimates, as these can adapt to the local density smoothness around each connected component at a level. We believe that it is possible to devise an approach based on spatially adapted partitions using the vernier or a modified Lepski method.

SSL Analysis under the Manifold Assumption : In Chapter 3 we suggested that the analysis techniques and arguments presented for the cluster assumption should also be applicable to analyze semi-supervised learning in the manifold setting. Under the manifold assumption, the target function lies on a low-dimensional manifold and is smooth with respect to the geodesic distance along the manifold. If the manifold can be learnt, that is,

geodesic distances can be recovered using unlabeled data, then the supervised learning task reduces to taking a majority vote on the manifold in classification, or computing the average (or polynomial fit) within geodesic balls along the manifold in regression. A result that can be leveraged in this regard is presented in [85] that states that geodesic distances on manifolds can be well-approximated using distances on graphs constructed using unlabeled data, where the quality of the approximation depends on certain geometric properties of the manifold, and is summarized below:

Theorem 14 (Main Theorem A, Bernstein et al. [85])**.** *Let M denote a geodesically convex manifold with minimum radius of curvature r_0 and minimum branch separation s_0, where the latter is defined to be the largest positive number for which $\|x-y\| < s_0$ implies $d_M(x,y) \leq \pi r_0$, $\|\cdot\|$ denotes Euclidean distance and d_M denotes geodesic distance. Assume that the sampling density of the unlabeled data points is high enough so that $\forall x \in M$, there exists an unlabeled data point X_i such that $d_M(x, X_i) \leq \delta$, where $\delta > 0$.*[1] *Let $d_G(\cdot,\cdot)$ denote the shortest distance between two points on the graph constructed using ϵ-rule on the unlabeled data, that is, two unlabeled data points have an edge if the Euclidean distance between them is no more than ϵ.*

Then given positive real numbers $\lambda_1, \lambda_2 < 1$, if $\epsilon < \min(s_0, (2/\pi)r_0\sqrt{24\lambda_1})$ and $\delta \leq \lambda_2\epsilon/4$, then for all unlabeled data pairs X_1, X_2,

$$(1-\lambda_1)d_M(X_1,X_2) \leq d_G(X_1,X_2) \leq (1+\lambda_2)d_M(X_1,X_2).$$

Thus, in the manifold case, the minimum radius of curvature r_0 and minimum branch separation s_0 play the role of the margin, and it can be argued that if these properties are resolvable using unlabeled data but not using labeled data, then semi-supervised learning can yield improved error bounds. For example, assuming a uniform marginal density and that the boundary of the manifold is regular, the sampling condition is satisfied with $\delta \sim m^{-1/d}$

[1] This condition is satisfied with high probability if the marginal density is bounded from below and the boundary of the manifold satisfies the following condition: $\forall x \in M$, $\mathrm{vol}_M(B_x(r)) \geq cr^d$ for all $r \leq r_1$, where vol_M denotes the d-dimensional volume measured along the manifold with intrinsic dimension d, $B_x(r)$ denotes a geodesic ball of radius r around the point x and $c, r_1 > 0$ are constants.

with probability at least $1 - 1/m$. This implies that the manifold structure is resolvable (the geodesic distances can be learnt using graph distance) provided s_0, r_0 are large compared to $m^{-1/d}$. Derivation of bounds for regression or classification in the manifold setting require slightly careful analysis since the geodesic distances are only learnt approximately, and hence the problem is tantamount to solving an error-in-variables problem [86].

Other Directions for Semi-Supervised learning : The semi-supervised learning literature contains many computationally efficient algorithms for using unlabeled data [58], however they provide no or very weak guarantees. The theory developed in this thesis helps characterize and understand the improvements possible using unlabeled data in favorable situations, and the next step would be to identify good practical algorithms based on these insights that are both statistically and computationally efficient. Moreover, while we focus on the cluster and manifold assumptions, it is also useful to identify other favorable situations where semi-supervised learning is helpful, for example, based on identifiability of mixtures as pioneered by the work of Castelli and Cover [59, 60]. There are some recent results on identifying component distributions in a multivariate mixture [87] that might be leveraged in this direction. Lastly, since semi-supervised learning is expected to yield performance gains only in favorable situations where there exists a link between the conditional and marginal densities and in practice it is not known a priori that such a link exists, it is important to safeguard against such situations and develop procedures that are *agnostic*. One possible approach to an agnostic solution is to develop a minimax optimal supervised learner $\widehat{f}_n$ along with the semi-supervised learner $\widehat{f}_{m,n}$ and choose the one that has smaller error on a hold-out data set. In the case where the size of the candidate classes used for both supervised and semi-supervised learning is finite, we can specify the final learner $\widehat{f}$ as

$$\widehat{f} = \begin{cases} \widehat{f}_{m,n} & \text{if } \widehat{R}(\widehat{f}_{m,n}) \leq \widehat{R}(\widehat{f}_n) + \text{pen}_{SL} + \text{pen}_{SSL} \\ \widehat{f}_n & \text{otherwise} \end{cases}.$$

Here $\widehat{R}$ denotes empirical risk on a hold-out data set, pen_{SL} is a penalty term that depends on the size of the candidate class used for supervised learning $\mathcal{F}_{SL}$ and bounds the difference $|R(f) - \widehat{R}(f)|$ for all $f \in \mathcal{F}_{SL}$ with high probability (at least $1 - n^{-c}$, for any $c > 0$)

and pen_{SSL} is a penalty term that depends on the size of the candidate class $\mathcal{F}_{SSL}$ used for semi-supervised learning and bounds the difference $|R(f) - \widehat{R}(f)|$ for all $f \in \mathcal{F}_{SSL}$ with high probability (at least $1 - n^{-c}$, for any $c > 0$). The rationale is that in favorable situations, $R(\widehat{f}_{m,n}) \leq R(\widehat{f}_n)$ which implies that with probability at least $1-2n^{-c}$, $\widehat{R}(\widehat{f}_{m,n}) \leq \widehat{R}(\widehat{f}_n) + \text{pen}_{SL} + \text{pen}_{SSL}$. And if the situation is not favorable, then the semi-supervised learner is replaced by the supervised learner. This implies that the overall error incurred is never much worse than the best supervised learner, and is much smaller in favorable situations.

Exploiting structure in multiple hypothesis testing problems : The fMRI data analysis problem considered in this thesis is one example of multiple hypothesis testing problems where it is useful to exploit structural information to enhance detection capabilities. Similar problems arise in testing for significance of gene expression in microarray analysis [9] and detection of weak patterns of spatially distributed activity in a network [88]. While the activation at a single node or gene may not be statistically significant, aggregation of activations based on structural information can boost detection power. Notice that the method proposed in this thesis may not apply here, since the data are unordered and lie on a graph. Thus, an important open problem is to exploit structure for improved detection in a fashion that is adaptive to the unknown size and shape of the structural pattern. We would like to point out that these multiple hypothesis testing problems can also be viewed as detection of sparse activations, where sparsity can be defined in an appropriate structural basis. While there is a flurry of recent research activity on addressing sparsity, identifying the appropriate sparsifying basis adaptively remains an open problem. Recent attempts at developing efficient representation of functions defined on graphs such as diffusion wavelets [89] seem particularly promising in this direction.

APPENDIX
Proof Sketch of Theorem 13

The proposed algorithm for clustering yields $|\widehat{P}(\widehat{G}_{\gamma_{k+1}}) - \widehat{P}(\widehat{G}_{\gamma_k})| \leq \nu/4$. If we can show that $|P(G^*_{\gamma_k}) - \widehat{P}(\widehat{G}_{\gamma_k})| \leq \epsilon/2$ for all k, then it follows that $|P(G^*_{\gamma_{k+1}}) - P(G^*_{\gamma_k})| \leq \nu/4 + \epsilon$. First observe that

$$|P(G^*_{\gamma_k}) - \widehat{P}(\widehat{G}_{\gamma_k})| \leq |P(G^*_{\gamma_k}) - P(\widehat{G}_{\gamma_k})| + |P(\widehat{G}_{\gamma_k}) - \widehat{P}(\widehat{G}_{\gamma_k})|$$

We now establish bounds on the two terms of the right hand side. To bound the first term, we prove that Hausdorff control also provides mass control under regularity assumptions on the density. From Corollary 1, we have $\forall k \in K$

$$d_\infty(G^*_{\gamma_k}, \widehat{G}_{\gamma_k}) \leq C s_n^2 \left(\frac{n}{\log n}\right)^{-\frac{1}{d+2\max_k \alpha_k}} := \epsilon'$$

So $\forall x \in \widehat{G}_{\gamma_k} \setminus G^*_{\gamma_k}$, $\rho(x, \partial G^*_{\gamma_k}) \leq \epsilon'$. If the boundary $\partial G^*_{\gamma_k}$ is locally Lipschitz, it implies that the length of the boundary is a constant, say ℓ. Therefore, we have

$$\mu(\widehat{G}_{\gamma_k} \setminus G^*_{\gamma_k}) \leq \ell\epsilon'$$

And $\forall x \in G^*_{\gamma_k} \setminus \widehat{G}_{\gamma_k}$, $\rho(x, \partial \widehat{G}_{\gamma_k}) \leq \epsilon'$. In fact it is also true that $\rho(x, \partial G^*_{\hat{\gamma}_k}) \leq \epsilon'$ (recall Proposition 4). Since the length of the boundary of the true level set is a constant, we have

$$\mu(G^*_{\gamma_k} \setminus \widehat{G}_{\gamma_k}) \leq \ell\epsilon'$$

Translating this to mass control, (Here C may denote a different constant from line to line):

$$\begin{aligned}
|P(G^*_{\gamma_k}) - P(\widehat{G}_{\gamma_k})| &\leq P(G^*_{\gamma_k} \setminus \widehat{G}_{\gamma_k}) + P(\widehat{G}_{\gamma_k} \setminus G^*_{\gamma_k}) \\
&\leq (\gamma_k + C_2\epsilon'^{\alpha_k})\mu(G^*_{\gamma_k} \setminus \widehat{G}_{\gamma_k}) + \gamma_k \mu(\widehat{G}_{\gamma_k} \setminus G^*_{\gamma_k}) \\
&\leq C\gamma_k \mu(G^*_{\gamma_k} \setminus \widehat{G}_{\gamma_k}) + \gamma_k \mu(\widehat{G}_{\gamma_k} \setminus G^*_{\gamma_k}) \\
&\leq C\epsilon'
\end{aligned}$$

where the second step invokes the assumption on the density regularity that the deviation in density from the level γ_k scales as the α_k power of the distance from $\partial G^*_{\gamma_k}$.

For the second term, recall that for all sets G defined by collections of cells from $\mathcal{A}_{0:J}$, that is let G be a collection of cells with sidelength $2^{-j(G)}$ in a regular partition of $[0,1]^d$,

$$|P(G) - \hat{P}(G)| \leq \sum_{A \in G} |P(A) - \hat{P}(A)| \quad = \quad \sum_{A \in G} |\bar{f}(A) - \widehat{f}(A)| \mu(A) \leq \Psi_{j(G)}$$

where the last step follows from proof of Lemma 2. So we have:

$$\begin{aligned} |P(\widehat{G}_{\gamma_k}) - \widehat{P}(\widehat{G}_{\gamma_k})| \leq \Psi_{\hat{j}} \quad &\leq \quad C \left(\frac{n}{\log n} \right)^{-\frac{\alpha_k}{d+2\alpha_k}} \\ &\leq \quad C \left(\frac{n}{\log n} \right)^{-\frac{\min_k \alpha_k}{d+2\min_k \alpha_k}} := \epsilon'' \end{aligned}$$

The second step follows from Theorem 2 using the bounds established on the chosen sidelength.

The result of Theorem 13 follows by taking $\epsilon = 2(\max\{\epsilon', \epsilon''\})$.

□

REFERENCES

[1] W. Stuetzle, "Estimating the cluster tree of a density by analyzing the minimal spanning tree of a sample," *Journal of Classification*, vol. 20, no. 5, pp. 25–47, 2003.

[2] C. Scott and E. Kolaczyk, "Annotated minimum volume sets for nonparametric anomaly discovery," in *IEEE Workshop on Statistical Signal Processing (SSP)*, 2007.

[3] M. Horstein, "Sequential decoding using noiseless feedback," *IEEE Transactions on Information Theory*, vol. 9, no. 3, pp. 136–143, 1963.

[4] R. Castro, R. Willett, and R. Nowak, "Faster rates in regression via active learning," in *Advances in Neural Information Processing Systems 19 (NIPS)*, 2005.

[5] R. Castro, R. Willett, and R. Nowak, "Faster rates in regression via active learning," Tech. Rep. ECE-05-3, University of Wisconsin - Madison. URL http://homepages.cae.wisc.edu/∼rcastro/publications/ECE-05-3.pdf, 2005.

[6] B. Silverman, *Density Estimation for Statistics and Data Analysis*. London: Chapman & Hall, 1986.

[7] D. Scott, *Multivariate Density Estimation: Theory, Practice, and Visualization*. New York: Wiley, 1992.

[8] L. Devroye and G. Lugosi, *Combinatorial Methods in Density Estimation*. New York: Springer-Verlag, 2001.

[9] J. H. Friedman, R. Tibshirani, and T. Hastie, *The Elements of Statistical Learning*. New York: Springer-Verlag, 2003.

[10] L. Wasserman, *All of Nonparamteric Statistics*. New York: Springer-Verlag, 2005.

[11] W. Hardle, A. Chesher, and M. Jackson, *Applied Nonparamteric Regression*. Cambridge Univeristy Press, 1992.

[12] J. A. Hartigan, *Clustering Algorithms*. NY: Wiley, 1975.

[13] A. Cuevas, M. Febrero, and R. Fraiman, "Cluster analysis: a further approach based on density estimation," *Computational Statistics and Data Analysis*, vol. 36, pp. 441–459, 2001.

[14] M. Ester, H. P. Kriegel, J. Sander, and X. Xu, "A density based algorithm for discovering clusters in large spatial databases with noise," in *Proceedings of 2nd International Conference on Knowledge Discovery and Data Mining (KDD)*, 1996.

[15] R. Y. Liu, J. M. Parelius, and K. Singh, "Multivariate analysis by data depth: Descriptive statistics, graphics and inference," *Annals of Statistics*, vol. 27, no. 3, pp. 783–858, 1999.

[16] I. Steinwart, D. Hush, and C. Scovel, "A classification framework for anomaly detection," *Journal of Machine Learning Research*, vol. 6, pp. 211–232, 2005.

[17] R. Vert and J.-P. Vert, "Consistency and convergence rates of one-class svms and related algorithms," *Journal of Machine Learning Research*, vol. 7, pp. 817–854, 2006.

[18] R. Willett and R. Nowak, "Level set estimation in medical imaging," in *IEEE Workshop on Statistical Signal Processing (SSP)*, 2005.

[19] Z. Harmany, R. Willett, A. Singh, and R. Nowak, "Controlling the error in fMRI: Hypothesis testing or set estimation?," in *Fifth IEEE International Symposium on Biomedical Imaging (ISBI)*, 2008.

[20] "http://www.cens.ucla.edu/portal/nims/."

[21] A. Singh, R. Nowak, and P. Ramanathan, "Active learning for adaptive mobile sensing networks," in *Information Processing in Sensor Networks (IPSN)*, 2006.

[22] R. Willett and R. Nowak, "Minimax optimal level set estimation," *IEEE Transactions on Image Processing*, vol. 16, no. 12, pp. 2965–2979, 2007.

[23] B. Efron and R. Tibshirani, "On testing the significance of sets of genes," *Annals of Applied Statistics*, vol. 1, no. 1, pp. 107–129, 2007.

[24] Y. H. Yang, M. Buckley, S. Dudoit, and T. Speed, "Comparision of methods for image analysis on cdna microarray data," *Journal of Computational and Graphical Statistics*, vol. 11, pp. 108–136, 2002.

[25] I. Steinwart, D. Hush, and C. Scovel, "Density level detection is classification," in *Advances in Neural Information Processing Systems 17 (NIPS)*, pp. 1337–1344, 2005.

[26] P. Rigollet, "Generalization error bounds in semi-supervised classification under the cluster assumption," *Journal of Machine Learning Research*, vol. 8, pp. 1369–1392, 2007.

[27] J. Lafferty and L. Wasserman, "Statistical analysis of semi-supervised regression," in *Advances in Neural Information Processing Systems 20 (NIPS)*, pp. 801–808, 2007.

[28] M. Seeger, "Learning with labeled and unlabeled data," tech. rep., Institute for ANC, Edinburgh, UK. URL http://citeseer.ist.psu.edu/seeger01learning.html, 2000.

[29] R. Szewczyk, E. Osterweil, J. Polastre, M. Hamilton, A. Mainwaring, and D. Estrin, "Habitat monitoring with sensor networks," *Communications of the ACM*, vol. 47, no. 6, pp. 34–40, 2004.

[30] M. Batalin, M. Rahimi, Y. Yu, D. Liu, A. Kansal, G. Sukhatme, W. Kaiser, M. Hansen, G. Pottie, M. Srivastava, and D. Estrin, "Call and response: Experiments in sampling the environment," in *Proceedings of ACM SenSys*, 2004.

[31] C. J. Stone, "Optimal rates of convergence for nonparametric estimators," *Annals of Statistics*, vol. 8(6), pp. 1348–1360, 1980.

[32] C. Scott and R. Nowak, "Learning minimum volume sets," *Journal of Machine Learning Research*, vol. 7, pp. 665–704, 2006.

[33] C. Scott and M. Davenport, "Regression level set estimation via cost-sensitive classification," *IEEE Transactions on Signal Processing*, vol. 55, no. 6, pp. 2752–2757, 2007.

[34] A. P. Korostelev and A. B. Tsybakov, *Minimax Theory of Image Reconstruction*. NY: Springer, 1993.

[35] A. B. Tsybakov, "On nonparametric estimation of density level sets," *Annals of Statistics*, vol. 25, pp. 948–969, 1997.

[36] W. Polonik, "Measuring mass concentrations and estimating density contour cluster-an excess mass approach," *Annals of Statistics*, vol. 23, no. 3, pp. 855–881, 1995.

[37] P. Rigollet and R. Vert, "Fast rates for plug-in estimators of density level sets, url http://www.citebase.org/abstract?id=oai:arxiv.org:math/0611473," 2006.

[38] L. Cavalier, "Nonparametric estimation of regression level sets," *Statistics*, vol. 29, pp. 131–160, 1997.

[39] A. P. Korostelev and A. B. Tsybakov, "Estimation of the density support and its functionals," *Problems of Information Transmission*, vol. 29, no. 1, pp. 1–15, 1993.

[40] S. Ben-David, T. Lu, and D. Pal, "Does unlabeled data provably help? worst-case analysis of the sample complexity of semi-supervised learning," in *21st Annual Conference on Learning Theory (COLT)*, 2008.

[41] P. Niyogi, "Manifold regularization and semi-supervised learning: Some theoretical analyses," Tech. Rep. TR-2008-01, Computer Science Department, University of Chicago. URL http://people.cs.uchicago.edu/∼niyogi/papersps/ssminimax2.pdf, 2008.

[42] Y. Benjamini and Y. Hochberg, "Controlling the false discovery rate: a practical and powerful approach to multiple testing," *Journal of the Royal Statistical Society: Series B*, vol. 57, pp. 289–300, 1995.

[43] F. Abramovich, Y. Benjamini, D. L. Donoho, and I. M. Johnstone, "Adapting to unknown sparsity by controlling the false discovery rate," *Annals of Statistics*, vol. 34, no. 2, pp. 584–653, 2006.

[44] M. Pacifico, C. Genovese, I. Verdinelli, and L. Wasserman, "False discovery control for random fields," *J. Amer. Statist. Assoc.*, vol. 99, no. 468, pp. 1002–1014, 2004.

[45] A. Sole, V. Caselles, G. Sapiro, and F. Arandiga, "Morse description and geometric encoding of digital elevation maps," *IEEE Transactions on Image Processing*, vol. 13, no. 9, pp. 1245–1262, 2004.

[46] A. Cuevas, W. G. Manteiga, and A. R. Casal, "Plug-in estimation of general level sets," *Australian and New Zealand Journal of Statistics*, vol. 48, no. 1, pp. 7–19, 2006.

[47] O. V. Lepski, E. Mammen, and V. G. Spokoiny, "Optimal spatial adaptation to inhomogeneous smoothness: An approach based on kernel estimates with variable bandwidth selectors," *Annals of Statistics*, vol. 25, no. 3, pp. 929–947, 1997.

[48] L. Breiman, J. Friedman, R. Olshen, and C. J. Stone, *Classification and Regression Trees*. Belmont, CA: Wadsworth, 1983.

[49] D. L. Donoho, "CART and best-ortho-basis: A connection," *Annals of Statistics*, vol. 25, pp. 1870–1911, 1997.

[50] E. Kolaczyk and R. Nowak, "Multiscale likelihood analysis and complexity penalized estimation," *Annals of Statistics*, vol. 32, no. 2, pp. 500–527, 2004.

[51] C. Scott and R. Nowak, "Minimax-optimal classification with dyadic decision trees," *IEEE Transactions on Information Theory*, vol. 52, no. 4, pp. 1335–1353, 2006.

[52] V. Vapnik, *The Nature of Statistical Learning Theory*. NY: Springer, 1995.

[53] W. Hårdle, B. U. Park, and A. B. Tsybakov, "Estimation of non-sharp support boundaries," *Journal of Multivariate Analysis*, vol. 5, pp. 205–218, 1995.

[54] C. C. Craig, "On the tchebychef inequality of bernstein," *Annals of Statistics*, vol. 4, no. 2, pp. 94–102, 1933.

[55] D. L. Donoho, "Wedgelets: Nearly-minimax estimation of edges," *Annals of Statistics*, vol. 27, pp. 859–897, 1999.

[56] E. Candés and D. L. Dohono, "Curvelets: A surprisingly effective nonadaptive representation for objects with edges," *Curves and Surfaces, Larry Schumaker et al., Ed. Vanderbilt University Press, Nashville, TN*, 1999.

[57] A. B. Tsybakov, *Introduction a l'estimation non-parametrique*. Berlin Heidelberg: Springer, 2004.

[58] X. Zhu, "Semi-supervised learning literature survey," Tech. Rep. TR 1530, Computer Sciences, University of Wisconsin-Madison. URL http://www.cs.wisc.edu/∼jerryzhu/pub/ssl_survey.pdf, 2005.

[59] V. Castelli and T. M. Cover, "On the exponential value of labeled samples," *Pattern Recognition Letters*, vol. 16, no. 1, pp. 105–111, 1995.

[60] V. Castelli and T. M. Cover, "The relative value of labeled and unlabeled samples in pattern recognition," *IEEE Transactions on Information Theory*, vol. 42, no. 6, pp. 2102–2117, 1996.

[61] P. J. Bickel and B. Li, "Local polynomial regression on unknown manifolds," in *IMS Lecture NotesMonograph Series, Complex Datasets and Inverse Problems: Tomography, Networks and Beyond*, vol. 54, pp. 177–186, 2007.

[62] A. Korostelev and M. Nussbaum, "The asymptotic minimax constant for sup-norm loss in nonparametric density estimation," *Bernoulli*, vol. 5(6), pp. 1099–1118, 1999.

[63] O. Chapelle and A. Zien, "Semi-supervised classification by low density separation," in *10th International Workshop on Artificial Intelligence and Statistics (AISTATS)*, pp. 57–64, 2005.

[64] A. Singh, R. Nowak, and C. Scott, "Adaptive hausdorff estimation of density level sets," in *21st Annual Conference on Learning Theory (COLT)*, 2008.

[65] J. Xiong, J. Gao, J. Lancaster, and P. Fox, "Assessment and optimization of functional MRI analyses," *Human Brain Mapping*, vol. 4, pp. 153–167, 1996.

[66] R. Heller, D. Stanley, D. Yekutieli, N. Rubin, and Y. Benjamini, "Cluster-based analysis of fMRI data," *NeuroImage*, vol. 33, no. 2, pp. 599–608, 2006.

[67] K. J. Friston, A. Holmes, J.-B. Poline, C. J. Price, and C. D. Frith, "Detecting activations in PET and fMRI: Levels of inference and power," *Neuroimage*, vol. 40, pp. 223–235, 1996.

[68] D. V. D. Ville, T. Blu, and M. Unser, "Integrated wavelet processing and spatial statistical testing of fMRI data," *NeuroImage*, vol. 23, no. 4, pp. 1472–1485, 2004.

[69] D. O. Siegmund and K. J. Worsley, "Testing for a signal with unknown location and scale in a stationary gaussian random field," *Annals of Statistics*, vol. 23, pp. 608–639, 1994.

[70] E. Mammen and A. B. Tsybakov, "Smooth discrimination analysis," *Annals of Statistics*, vol. 27, pp. 1808–1829, 1999.

[71] A. B. Tsybakov, "Optimal aggregation of classifiers in statistical learning," *Annals of Statistics*, vol. 32, no. 1, pp. 135–166, 2004.

[72] P. Bartlett, S. Boucheron, and G. Lugosi, "Model selection and error estimation," *Machine Learning*, vol. 48, pp. 85–113, 2002.

[73] M. Rahimi, R. Pon, W. J. Kaiser, G. S. Sukhatme, D. Estrin, and M. Srivastava, "Adaptive sampling for environmental robotics," in *Proceedings of IEEE International Conference on Robotics and Automation (ICRA)*, 2004.

[74] "http://limnosun.limnology.wisc.edu/."

[75] P. Hall and I. Molchanov, "Sequential methods for design-adaptive estimation of discontinuities in regression curves and surfaces," *Annals of Statistics*, vol. 31, no. 3, pp. 921–941, 2003.

[76] T. D. Parsons, "Pursuit-evasion in a graph," in *In Y. Alani and D. R. Lick, editors, Theory and Application of Graphs*, pp. 426–441, Springer-Verlag, 1976.

[77] J. P. Hespanha, H. J. Kim, and S. Sastry, "Multiple-agent probabilistic pursuit-evasion games," in *Proceedings of the Conference on Decision and Control*, December 1999.

[78] G. Kesidis, T. Konstantopoulos, and S. Phoha, "Surveillance coverage of sensor networks under a random mobility strategy," in *Proceedings of IEEE Sensors*, pp. 961–965, 2003.

[79] M. Burnashev and K. S. Zigangirov, "An interval estimation problem for controlled observations," *Problems of Information Transmission*, vol. 10, pp. 223–231, 1974.

[80] A. Korostelev, "On minimax rates of convergence in image models under sequential design," *Statistics and Probability Letters*, vol. 43, pp. 369–375, 1999.

[81] G. Golubev and B. Levit, "Sequential recovery of analytic periodic edges in the binary image models," *Mathematical Methods of Statistics*, vol. 12, pp. 95–115, 2003.

[82] H. Yuan and C. Wu, "An implicit 3d fully non-hydrostatic model for free-surface flows," *International Journal for Numerical Methods in Fluids*, vol. 46, pp. 709–733, 2004.

[83] R. Willett, A. Martin, and R. Nowak, "Backcasting: Adaptive sampling for sensor networks," in *Fifth International Conference on Information Processing in Sensor Networks (IPSN)*, 2004.

[84] S. Ben-David and U. von Luxburg, "Relating clustering stability to properties of cluster boundaries," in *21st Annual Conference on Learning Theory (COLT)*, 2008.

[85] M. Bernstein, V. de Silva, J. Langford, and J. Tenenbaum, "Graph approximations to geodesics on embedded manifolds," tech. rep., Stanford University. URL http://citeseer.ist.psu.edu/bernstein00graph.html, 2000.

[86] S. M. Schennach, Y. Hu, and A. Lewbel, "Nonparametric identification of the classical errors-in-variables model without side information," Tech. Rep. 674, Boston College, Department of Economics. URL http://ideas.repec.org/p/boc/bocoec/674.html, 2007.

[87] P. Hall and X.-H. Zhou, "Nonparametric estimation of component distributions in a multivariate mixture," *Annals of Statistics*, vol. 31, no. 1, pp. 201–224, 2003.

[88] A. Lakhina, M. Crovella, and C. Diot, "Diagnosing network-wide traffic anomalies," *SIGCOMM Computer Communication Review*, vol. 34, no. 4, pp. 219–230, 2004.

[89] R. Coifman and M. Maggioni, "Diffusion wavelets," *Applied and Computational Harmonic Analysis*, vol. 21, pp. 53–94, July 2006.