Adaptive Hausdorff Estimation of Density Level Sets

Aarti Singh, Clayton Scott and Robert Nowak

Technical Report ECE-07-06
Department of Electrical and Computer Engineering
University of Wisconsin - Madison

Abstract

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 paper 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.

1 Introduction

Level sets provide useful summaries of a function for many applications including clustering [1, 2], anomaly detection [3, 4, 5], functional neuroimaging [6, 7], bioinformatics [8], digital elevation mapping [9, 10], and environmental monitoring [11]. 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 [3, 4, 12, 9, 13, 14, 15, 16] considers error measures related to the symmetric set difference,

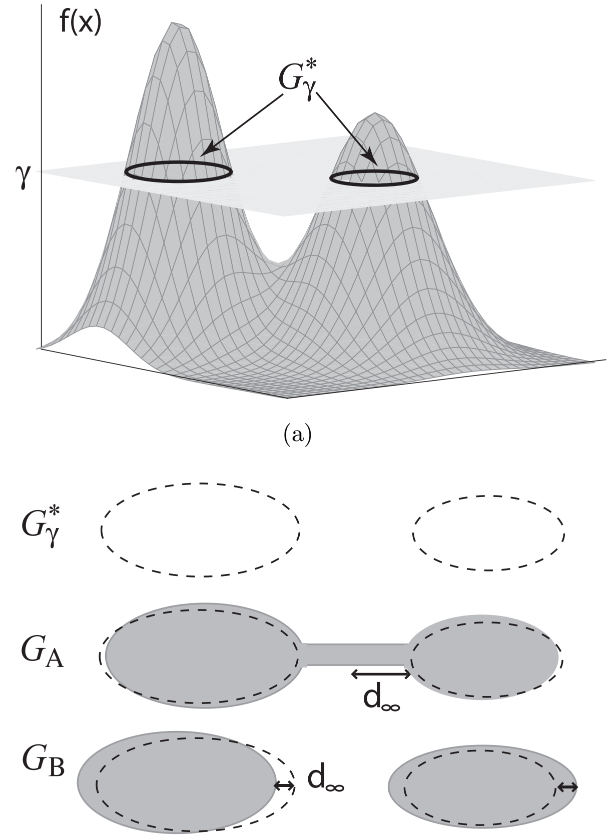
$$G_1 \Delta G_2 = (G_1 \setminus G_2) \cup (G_2 \setminus G_1). \tag{1}$$

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 [3, 13, 14, 16] a probability measure of the symmetric set difference is considered, and in [12, 9, 16] 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, 2, 17], or ensure robustness to outliers in level set based anomaly detection [3, 4, 5] and data ranking [18]. 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 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:

$$d_\infty(G_1, G_2) = \max\{\sup_{x \in G_2} \rho(x, G_1), \sup_{x \in G_1} \rho(x, G_2)\} \tag{2}$$

where

$$\rho(x, G) = \inf_{y \in G} ||x - y||, \tag{3}$$



(a)

(b)

Figure 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 $d_\infty(G_B, G_\gamma^*)$.

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 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 [13, 14, 19] 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 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 [13, 14, 19] 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 [20] 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 paper, 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 [21], 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 [22, 23, 24, 25] 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 [5], 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, 2, 17], 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 [18]. 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 [3, 4, 5]. 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, 27]. As discussed in [28], 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 paper is organized as follows. Section 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 3 discusses the issue with direct Hausdorff estimation and provides motivation for an alternate error measure. In Section 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 4.1 extends the proposed estimator to achieve adaptivity to unknown density regularity. Subsections 4.2-4.4 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 5 and Section 6 contains the proofs.

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 (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 [13, 14, 19, 20]. 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 4.3 and the case $\alpha = 0$ (which corresponds to a jump in the density at level γ) in Subsection 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 [14, 19], except that the upper bound assumption on the density deviation in [14, 19] 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 [14, 19] 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 [13, 14, 19] 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 [14]. Then G_γ^* satisfies the level set regularity assumption [**B**].*

The proof is given in Section 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 [14] 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 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.*

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 paper. 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:

$$\begin{aligned} \mathcal{E}(G) &= \max\{ \sup_{x \in G_\gamma^* \setminus G} (f(x) - \gamma), \sup_{x \in G \setminus G_\gamma^*} (\gamma - f(x))\} && (4) \\ &= \sup_{x \in \mathcal{X}} (\gamma - f(x))[\mathbf{1}_{x \in G} - \mathbf{1}_{x \notin G}] && (5) \end{aligned}$$

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.

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*

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

The proof is given in Section 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 [20] 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* [29], 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

$$\mathcal{E}_\gamma(G) = \sup_{A \in \Pi(G)} (\gamma - \bar{f}(A))[\mathbf{1}_{A \subseteq G} - \mathbf{1}_{A \not\subseteq G}].$$

[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.

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 (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 $\left(\gamma - \bar{f}(A)\right)[\mathbf{1}_{A\subseteq G} - \mathbf{1}_{A\nsubseteq G}] > 0$ 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.

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

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:

$$\widehat{G}_{\Pi_o} = \arg\min_{G:\Pi(G)=\Pi_o} \widehat{\mathcal{E}}_\gamma(G) \tag{6}$$

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.

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

$$\widehat{G}_j = \bigcup_{A \in \mathcal{A}_j : \widehat{f}(A) \geq \gamma} A \tag{7}$$

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 [14, 19]), 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 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 $[\mathbf{A}]$ 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.

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 [21], 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

$$\widehat{G} = \widehat{G}_{\widehat{j}} \tag{8}$$

where

$$\widehat{j} = \arg\min_{0 \le j \le J} \left\{ \widehat{\mathcal{V}}_{\gamma,j} + \Psi_{j'} \right\} \tag{9}$$

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 (9), and $\widehat{G}$ denote the final estimate of (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 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.

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 $[\mathbf{A}]$ 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 (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, $\gamma_1 \le \gamma_2 \le \ldots \le \gamma_K$, then the true level sets will be nested $G^*_{\gamma_1} \supseteq G^*_{\gamma_2} \supseteq \ldots \supseteq G^*_{\gamma_K}$. 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

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

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).

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 [30, 14, 31] 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 6.6 and requires a construction similar to the minimax lower bound derived in [14] for level set estimation.

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

$$\widehat{G}_{0,j} = \bigcup_{A \in \mathcal{A}_j : \widehat{f}(A) > 0} A \tag{10}$$

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 [32] rather than the relative VC inequalities used in the proofs of Theorems 1, 2 for level set estimation. The proof is sketched in Section 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.

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 (7) achieves the minimax Hausdorff performance. To see this, the theoretical analysis needs to be modified a bit and is discussed in Section 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 6.8.

5 Concluding Remarks

In this paper, 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, [21]) 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 [33] or curvelets [34]. The earlier work on Hausdorff accurate level set estimation [13, 14, 19] 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) [22, 23, 24, 25] may provide significant improvements over non-adaptive partitions like histograms for many set learning problems involving a weighted symmetric difference error measure, including classification [25], minimum volume set estimation [4] and level set estimation [9]. 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 paper 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.

6 Proofs

6.1 Proof of Lemma 1

We proceed by recalling the definition of $\mathcal{F}_{SL}(\alpha)$ as defined in [14]. 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 $\boldsymbol{m} = (m_1, m_2, \ldots, m_{d-1}) \in \{1, \ldots, M\}^{d-1}$

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

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 $O(\epsilon)$.

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

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:

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

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

$$\begin{aligned}
\rho(y, \mathcal{I}_\epsilon(G^*_\gamma)) &\leq \sqrt{(g(\boldsymbol{\phi}) - g(\boldsymbol{\theta}))^2 + 4g(\boldsymbol{\phi})g(\boldsymbol{\theta}) \sum_{i=1}^{d-1} \sin^2 \frac{\phi_i - \theta_i}{2}} \\
&\leq |g(\boldsymbol{\phi}) - g(\boldsymbol{\theta})| + 2\sqrt{g(\boldsymbol{\phi})g(\boldsymbol{\theta})} \sum_{i=1}^{d-1} \left| \sin \frac{\phi_i - \theta_i}{2} \right|
\end{aligned}$$

$$
\begin{aligned}
&\leq L||\boldsymbol{\phi}-\boldsymbol{\theta}||_1 + \sum_{i=1}^{d-1}\frac{|\phi_i-\theta_i|}{2} \\
&= (L+1/2)\sum_{i=1}^{d-1}|\phi_i-\theta_i| \\
&\leq (L+1/2)d\frac{\pi}{M} \\
&\leq \frac{9d(L+1/2)}{h}\epsilon := C_3\epsilon
\end{aligned}
$$

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 $f \in \mathcal{F}_{SL}(\alpha)$.

■

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 [14]. Lemma 1 establishes that all densities in $\mathcal{F}_{SL}(\alpha)$ satisfy assumption **[B]**. Further, since the discrete set of densities $\mathcal{F}^D_{SL}(\alpha) \subset \mathcal{F}_{SL}(\alpha)$ used to derive the lower bound using Fano's lemma in [14], satisfy the local density regularity as stated in assumption **[A]** [2], we have

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

for n large enough. The last step follows from proof of Theorem 4 in [14].

■

[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]**.

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), $|f(x) - \gamma| \leq |f(x) - \widehat{f}(x)|$ 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:

$$\sup_{x \in \widehat{G}\Delta G^*_\gamma} \rho(x, \partial G^*_\gamma) \leq \sup_{x \in \widehat{G}\Delta G^*_\gamma} \left(\frac{|f(x) - \gamma|}{C_1}\right)^{1/\alpha} \leq \left(\frac{\mathcal{E}(\widehat{G})}{C_1}\right)^{1/\alpha} =: \epsilon. \tag{11}$$

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,

$$\sup_{x \in \mathcal{I}_{2\epsilon}} \rho(x, \widehat{G} \cap G^*_\gamma) \leq \epsilon. \tag{12}$$

Using Eqs. (11) and (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 (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

$$\sup_{x \in \widehat{G}} \rho(x, G^*_\gamma) \leq \left(\frac{\mathcal{E}(\widehat{G})}{C_1} \right)^{1/\alpha}. \tag{13}$$

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 (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 (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 (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

$$\sup_{x \in G^*_\gamma} \rho(x, \widehat{G}) \leq (2C_3 + 3)\epsilon = (2C_3 + 3)\left(\frac{\mathcal{E}(\widehat{G})}{C_1}\right)^{1/\alpha}. \tag{14}$$

Putting together the bounds from Eqs. (13), (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.

■

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 [35] 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

$$\widehat{P}(A) \leq P(A) + \epsilon\sqrt{\widehat{P}(A)} \Longrightarrow \widehat{P}(A) \leq 2\max(P(A), 2\epsilon^2) \tag{15}$$

and

$$P(A) \leq \widehat{P}(A) + \epsilon\sqrt{P(A)} \Longrightarrow P(A) \leq 2\max(\widehat{P}(A), 2\epsilon^2). \tag{16}$$

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 (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:

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

$\square$

We now analyze the performance of the plug-in histogram-based level set estimator proposed in (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 sidelength 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

$$d_\infty(\widehat{G}_j, G^*_\gamma) = \max\{\sup_{x \in G^*_\gamma} \rho(x, \widehat{G}_j), \sup_{x \in \widehat{G}_j} \rho(x, G^*_\gamma)\}, \tag{17}$$

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

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

where the last step follows from Proposition 4. Thus,

$$\sup_{x\in\widehat{G}_j}\rho(x,G^*_\gamma)\leq\epsilon_j. \tag{18}$$

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 (19), we get:

$$\sup_{x \in G^*_\gamma} \rho(x, \widehat{G}_j) \leq (2C_3+3)\epsilon_j. \tag{19}$$

From Eqs. (18) and (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 $> 1 - 3/n$,

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

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 $> 1 - 3/n$,

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

□

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.)

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

$$
\begin{aligned}
&\le \quad C \max\left\{ s_n^{-d/2\alpha} \left(\frac{n}{\log n}\right)^{-\frac{1}{d+2\alpha}}, s_n \left(\frac{n}{\log n}\right)^{-\frac{1}{d+2\alpha}}, \frac{1}{n} \right\} \\
&\le \quad C s_n \left(\frac{n}{\log n}\right)^{-\frac{1}{d+2\alpha}}.
\end{aligned}
$$

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 $P(\Omega) \le 1$ and since the domain $\mathcal{X} = [0,1]^d$, $\mathbb{E}[d_\infty(\widehat{G}_j, G^*_\gamma)|\bar{\Omega}] \le \sqrt{d}$. The third step follows from Lemma 4 and the fourth one using Lemma 3. The fifth step follows since the chosen resolution $2^{-j} \asymp s_n(n/\log n)^{-\frac{1}{(d+2\alpha)}}$.

■

6.5 Proof of Theorem 2

To analyze the resolution chosen by the complexity penalized procedure of (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 $j \ge 0$:*

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

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.)

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

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,

$$\widehat{\mathcal{V}}_{\gamma,j} - \mathcal{V}_{\gamma,j} = |\gamma - \widehat{f}(A'_{11})| - |\gamma - \bar{f}(A'_{00})|$$

$$
\begin{aligned}
&\leq \ |\gamma - \widehat{f}(A'_{01})| - |\gamma - \bar{f}(A'_{00})| \\
&\leq \ |\bar{f}(A'_{00}) - \widehat{f}(A'_{01})|
\end{aligned}
$$

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,

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

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 (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 (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.

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

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 $1-2/n$,

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

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 $1 - 2/n$,

$$c_1 s_n^{\frac{d}{d+2\alpha}} \left(\frac{n}{\log n}\right)^{-\frac{1}{d+2\alpha}} \leq 2^{-\widehat{j}} \leq c_2 s_n s_n^{\frac{d}{d+2\alpha}} \left(\frac{n}{\log n}\right)^{-\frac{1}{d+2\alpha}}, \tag{20}$$

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 (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.)

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

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 (20).

■

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 [14], 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 [36].

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 $\operatorname{supp}(K) = (-1/2, 1/2)^{d-1}$. 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

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

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

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.

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:

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

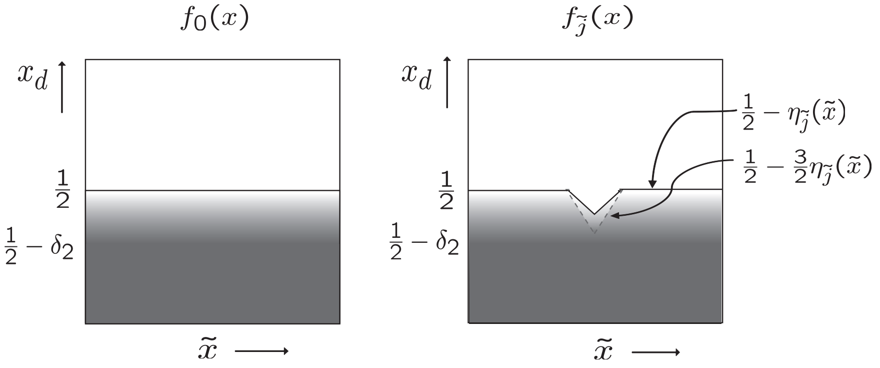


Figure 2: Densities used in the lower bound construction for Hausdorff accurate support set estimation.

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 $\tilde{j} \neq \tilde{k}$,

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

   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.

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

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.

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

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

Finally, we get

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

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

■

6.7 Proof sketch of Theorem 3

We derive an upper bound on the Hausdorff error of the estimator proposed in (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 [32].

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 [32]:

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 $2^j = O((n/\log n)^{1/d})$,

$$c_5 \frac{2^{jd} \log n}{n} \leq \Psi_j^0 \leq c_6 \frac{2^{jd} \log n}{n}. \tag{21}$$

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 (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 $s_n(n/\log n)^{-1/(d+\alpha)}$.

■

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 (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

$$d_\infty(\widehat{G}_j, G_\gamma^*) \leq \max(2C_3 + 3, 8\sqrt{d}\epsilon_o^{-1}) \left[2\sqrt{d}2^{-j}\right]. \qquad (22)$$

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 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:

$$\mathcal{V}_{\gamma,\widehat{j}} \leq \widehat{\mathcal{V}}_{\gamma,\widehat{j}} + \Psi_{\widehat{j}'} \quad \leq \quad C2^{-j^{*'}/2}2^{-j^*\alpha} \leq Cs_n^{-1/2}s_n^{\frac{d(\alpha+1/2)}{d+2\alpha}}\left(\frac{n}{\log n}\right)^{-\frac{(\alpha+1/2)}{d+2\alpha}}.$$

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 $s_n > 8\max(3\epsilon_o^{-1}, 28, 12C_3)\sqrt{d}$,

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

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:

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

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 (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.

■

References

[1] J. A. Hartigan, *Clustering Algorithms.* NY: Wiley, 1975.

[2] 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.

[3] I. Steinwart, D. Hush, and C. Scovel, "A classification framework for anomaly detection," *Journal of Machine Learning Research*, vol. 6, pp. 211–232, 2005.

[4] C. Scott and R. Nowak, "Learning minimum volume sets," *Journal of Machine Learning Research*, vol. 7, pp. 665–704, 2006.

[5] 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.

[6] 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.

[7] R. Willett and R. Nowak, "Level set estimation in medical imaging," in *IEEE Workshop on Statistical Signal Processing (SSP)*, 2005.

[8] 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.

[9] R. Willett and R. Nowak, "Minimax optimal level set estimation," *IEEE Transactions on Image Processing*, vol. 16, no. 12, pp. 2965–2979, 2007. [Online]. Available: http://www.ee.duke.edu/~willett/papers/WillettLevelSetsSubmitted.pdf

[10] 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.

[11] 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.

[12] 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.

[13] A. P. Korostelev and A. B. Tsybakov, *Minimax Theory of Image Reconstruction.* NY: Springer, 1993.

[14] A. B. Tsybakov, "On nonparametric estimation of density level sets," *Annals of Statistics*, vol. 25, pp. 948–969, 1997.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

[19] L. Cavalier, "Nonparametric estimation of regression level sets," *Statistics*, vol. 29, pp. 131–160, 1997.

[20] 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.

[21] 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.

[22] L. Breiman, J. Friedman, R. Olshen, and C. J. Stone, *Classification and Regression Trees*. Belmont, CA: Wadsworth, 1983.

[23] D. L. Donoho, "CART and best-ortho-basis: A connection," *Annals of Statistics*, vol. 25, pp. 1870–1911, 1997.

[24] E. Kolaczyk and R. Nowak, "Multiscale likelihood analysis and complexity penalized estimation," *Annals of Statistics*, vol. 32, no. 2, pp. 500–527, 2004.

[25] 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.

[26] M. Seeger, "Learning with labeled and unlabeled data," Institute for ANC, Edinburgh, UK. URL http://citeseer.ist.psu.edu/seeger01learning.html, Tech. Rep., 2000.

[27] P. Rigollet, "Generalization error bounds in semi-supervised classification under the cluster assumption," *Journal of Machine Learning Research*, vol. 8, pp. 1369–1392, 2007.

[28] A. Singh, R. D. Nowak, and X. Zhu, "Finite sample analysis of semi-supervised learning," University of Wisconsin - Madison, ECE Department. URL http://www.ece.wisc.edu/∼nowak/SSL_TR.pdf, Tech. Rep., 2008.

[29] V. Vapnik, *The Nature of Statistical Learning Theory.* NY: Springer, 1995.

[30] 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.

[31] 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.

[32] C. C. Craig, "On the tchebychef inequality of bernstein," *Annals of Statistics*, vol. 4, no. 2, pp. 94–102, 1933.

[33] D. L. Donoho, "Wedgelets: Nearly-minimax estimation of edges," *Annals of Statistics*, vol. 27, pp. 859–897, 1999.

[34] 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.

[35] L. Devroye and G. Lugosi, *Combinatorial Methods in Density Estimation.* New York: Springer-Verlag, 2001.

[36] A. B. Tsybakov, *Introduction a l'estimation non-parametrique.* Berlin Heidelberg: Springer, 2004.