Minimax Optimal Level Set Estimation

R. M. Willett, *Member, IEEE,* and R. D. Nowak, *Senior Member, IEEE*

Abstract

This paper describes a new methodology and associated theoretical analysis for rapid and accurate extraction of level sets of a multivariate function from noisy data. The identification of the boundaries of such sets is an important theoretical problem with applications for digital elevation maps, medical imaging, and pattern recognition. This problem is significantly different from classical segmentation because level set boundaries may not correspond to singularities or edges in the underlying function; as a result, segmentation methods which rely upon detecting boundaries would be potentially ineffective in this regime. This issue is addressed in this paper through a novel error metric sensitive to both the error in the location of the level set estimate and the deviation of the function from the critical level. Hoeffding's inequality is used to derive a novel regularization term that is distinctly different from regularization methods used in conventional image denoising settings. Building upon this foundation, it is possible to derive error performance bounds for the proposed estimator and demonstrate that it exhibits near minimax optimal error decay rates for large classes of level set problems. The proposed method automatically adapts to the spatially varying regularity of both the boundary of the level set and the underlying function.

I. Level Set Estimation

Level set estimation is the process of using noisy observations of a d-dimensional function f defined on the unit hypercube to estimate the region(s) in $[0,1]^d$ where f exceeds some critical value γ; *i.e.* $S^* \equiv \left\{x \in [0,1]^d : f(x) \geq \gamma\right\}$. Accurate and efficient level set estimation plays a crucial role in a variety of scientific and engineering tasks, including the following examples.

- **Geospatial Data:** Recent developments in the field of digital terrain elevation data (DTED) compression hinges on the knowledge of the precise location of key topographic features such as contour levels [1].
- **Bioinformatics:** Microarrays are typically preprocessed using segmentation schemes to identify sets (referred to as "spots" in the microarray literature) where the intensity of gene expression exceeds some level relative to the background [2].
- **Environmental Studies:** Contours of sunlight, rainfall and other key environmental factors are critical to understanding biosystem ecology [3].

Supported by the National Science Foundation, grants CCR-0310889 and ANI-0099148, and the Office of Naval Research, grant N00014-00-1-0390. R. Willett (willett@duke.edu) is with the Department of Electrical and Computer Engineering at Duke University, and R. Nowak is with the Department of Electrical and Computer Engineering at the University of Wisconsin-Madison.

In these and many other image processing applications, level sets are of principal importance, while the amplitude of the function (i.e. the image) away from the level set boundary is secondary, if not irrelevant. This paper presents a methodology and associated theoretical analysis for level set estimation. As noted above, the problem arises in several practical image processing contexts and many methods have been devised for level set estimation [4], [5], [6], yet there is very little theoretical analysis of the basic problem in the literature. One of the key results of the analysis in this paper is that regularization terms required for minimax optimal level set estimation are distinctly different from regularization terms required for minimax optimal image estimation and denoising.

Because set estimation is intrinsically simpler than function estimation, explicit level set estimation methods can potentially achieve higher accuracy than "plug-in" approaches based on computing an estimate of the entire function and thresholding the estimate to extract a level set. This is because function estimates aim to minimize the total error, integrated or averaged spatially over the entire function. This does little to control the error at specific locations of interest, such as in the vicinity of the level set. In part, plug-in approaches to can perform poorly because they tend to produce overly smooth estimates in the vicinity of the boundary of the level set.

Significant volumes of research have been dedicated to the estimation of functions containing singularities, edges, or more generally, lower-dimensional manifolds embedded in a higher-dimensional observation space; see [7], [8], [9], [10] for a few examples. In the context of level set estimation, however, the lower-dimensional manifold is an artificial feature which may not correspond to any form of singularity in the function. Related results from the classification literature suggest that unless the underlying function f is guaranteed to lie in a restrictive global smoothness class (a highly unrealistic assumption in many typical applications), conventional function estimation methods are neither appropriate nor optimal in this context [11], [12]. This is because the rates at which plug-in estimates converge to the true level set may be slowed down by the complexity of the function away from the boundary.

The above observations indicate that accurate level set estimation necessitates the development of new error metrics, methodologies, and error bounding techniques. In this paper, we develop such methods and theoretically characterize their performance. In particular, the estimator proposed in this paper exhibits several key properties:

- nearly achieves the minimax optimal error decay rate,
- automatically adapts to the regularity of the level set boundary,
- automatically adapts to the regularity of the underlying function f in the vicinity of the level set boundary,
- admits a computationally efficient implementation, and

- possesses enough flexibility to be useful in a variety of applications and contexts.

Each of these properties will be discussed in detail in the following sections.

A. Paper structure

In this paper, we describe a new method designed explicitly for minimax optimal level set estimation. The basic idea is to design an estimator of the form

$$\widehat{S} = \arg\min_{S \in \mathcal{S}} \widehat{\mathcal{R}}_n(S) + \Phi_n(S),$$

where $\mathcal{S}$ is a class of candidate level set estimates, $\widehat{\mathcal{R}}_n$ is an empirical measure of the level set estimation error based on n noisy observations of the function f, and Φ_n is a regularization term which penalizes improbable level sets. We describe choices for $\widehat{\mathcal{R}}_n$, Φ_n, and $\mathcal{S}$ which make $\widehat{S}$ rapidly computable and minimax optimal for a large class of level set problems. In particular, a novel error metric, which is ideally suited to the problem at hand, is proposed in Section II. We examine several of its key properties, and describe how it can be modified slightly to solve the closely related problems of (a) simultaneously extracting multiple level sets, and (b) density (as opposed to regression) level set estimation. In Section III, we introduce the regularization term Φ_n, describe its derivation from fundamental probability concentration inequalities, and develop a dyadic tree-based framework which can be used to minimize the proposed objective function. Trees are utilized for a couple of reasons. First, they both restrict and structure the space of potential estimators in a way that allows the global optimum to be both rapidly computable and very close to the best possible (not necessarily tree-based) estimator. Second, they allow us to introduce a spatial adaptivity to the estimator selection criterion which appears to be critical for the formation of provably optimal estimators. The nature of this spatial adaptivity and its role in achieving minimax optimal estimators is detailed in Section IV. That section contains derivations of error decay rates under different assumptions on the function f. In these first sections, we assume that the locations of the observations are randomly distributed over the domain of f with an unknown distribution. In Section V-A, we show that similar performance characterizations hold when the observations are deterministically and uniformly spaced over the domain of f, which is the case in many applications of interest. Section VI tackles computational issues and describes a method for cycle-spinning (or "voting over shifts") level set estimators for improved practical performance. Finally, Section VII contains several simulation results and Section VIII recapitulates key results and discusses avenues for further exploration.

B. Relationship to existing work

Related work in this field was conducted by Mammen and Tsybakov in [13], but their work focused on using binary observations to estimate a boundary between two constant-valued regions, an edge detection problem which is a special case of the more general level set estimation problem presented in this paper. The advantage of the regression level set estimation method proposed in this paper is that it is capable of utilizing additional information available from non-binary observations and from extracting level sets which do not correspond to edges or "jumps" in the amplitude of the function. Cavalier [14] examined a regression level set estimation problem similar to the one discussed here and based on the work of Tsybakov [15] for density level set estimation using piecewise polynomials. The estimators proposed in these works, however, are not computable and place stronger assumptions on S. Specifically, it must be "star-shaped" about the origin, an unrealistic assumption that allows the problem to be case in a function estimation setting. The work presented in this paper is most closely related to Dyadic Decision Trees, a binary classification method described in [16]. One of the key contributions of this paper is the establishment of a connection between the problems of binary classification and level set estimation. We demonstrate that the bounding techniques first developed in the context of classification are portable to more general settings and useful in a variety of contexts. The relationship between the two problems will be highlighted throughout the course of this paper.

Dyadic partitioning schemes similar in spirit to the one discussed in Section III of this paper have also been used in the context of estimating the support of a uniform density [17]. However, it is important to point out that in the support set estimation problem studied by [17] the boundary of the set corresponds to discontinuity of the density, and therefore more standard complexity-regularization and tree pruning methods commonly employed in regression problems suffice to achieve near minimax rates. In contrast, the approach we propose for level set estimation here is based on dyadic decision trees [16], and near minimax rates are achievable for all level sets whose boundaries belong to certain smoothness classes regardless of whether or not there is a discontinuity at the given level.

It is important to note here that the level set problem addressed in this paper can sometimes be approached using active contour models or snakes [4], [5], [6]. In particular, the proposed approach and active contours have similar high level objectives: active contours minimize an energy functional which measures contour roughness due to bending plus "image forces" such as image intensity and gradient in the vicinity of the contour, while the proposed level set estimation method minimizes an objective function which measures certain smoothness properties of the set boundary plus a risk function which measures the image intensity deviation (from the desired level) in the vicinity of the set boundary. However, the two methods are derived from significantly different perspectives and have different advantages and disadvantages. One of the chief

advantages of the active contour setup is that it is extremely flexible and can be used to solve a large variety of image processing and computer vision problems, including level set estimation, motion tracking, and image segmentation. However, in part due to the flexibility of these models, it is often very difficult to characterize the theoretical performance of the method. Furthermore, due to the complex nature of some active contour objective functions, it can be very difficult, if not impossible, to ensure that the globally optimal solution is obtained or to extend existing methods to high dimensional data. In contrast, the tree based method described in this paper is based on an objective function which has been derived from fundamental statistical principles and admits a thorough theoretical analysis. However, the proposed method is designed for the specific and well-defined problem of level set estimation, and its extension to related problems typically handled by active contour methods may not be trivial.

II. Regression Level Set Estimation

Let X and Y be random variables jointly distributed according to some unknown probability law $\mathbb{P}$. In this paper we consider the problem of estimating the γ-level set of the regression function $f(x) \equiv E[Y|X = x]$. Assume that the range of X is the unit hypercube in d dimensions, denoted $[0,1]^d$, and that the range of Y is the bounded interval $[-A, A]$, $A > 0$. Our goal is to estimate the set

$$S^* \;\equiv\; S(f,\gamma) \;=\; \Big\{x \in [0,1]^d : f(x) > \gamma\Big\} \tag{1}$$

based on n independent and identically distributed (i.i.d.) samples drawn from $\mathbb{P}$. We denote the samples by $\{X_i, Y_i\}_{i=1}^n$. The classical "signal+noise" model $Y = f(X)+W$, where W is a zero-mean noise independent of X, results in a special case of our problem. In the general set-up, the difference or "noise" $Y - f(X)$ may depend on X, but is nonetheless zero-mean. Practically speaking, we can interpret the general problem to be that of finding the level set of a function f based on point observations of f contaminated with independent, but not necessarily identically distributed, zero-mean noise. For example, in the image processing application of finding critical contour lines from a noisy digital terrain elevation data map, we have $d = 2$, the X_i's are the pixel locations, the Y_i's are the noisy observed elevations, and A is the maximum absolute elevation.

For the majority of this paper, we will focus on the case where both the locations $\{X_i\}$ and the observations $\{Y_i\}$ are random. As we will see later in this paper, however, the analysis framework presented here can be used in cases where the $\{X_i\}$ are deterministic (e.g., a uniform sampling grid). For now, let $\mathbb{P}^n$ be the n-fold product measure on $(X_i, Y_i)_{i=1}^n$ induced by $\mathbb{P}$. Let $\mathbb{E}$ and $\mathbb{E}^n$ denote expectation with respect to $\mathbb{P}$ and $\mathbb{P}^n$, respectively. And let $\mathbb{P}_X$ denote the marginal distribution of $\mathbb{P}$ with respect to X.

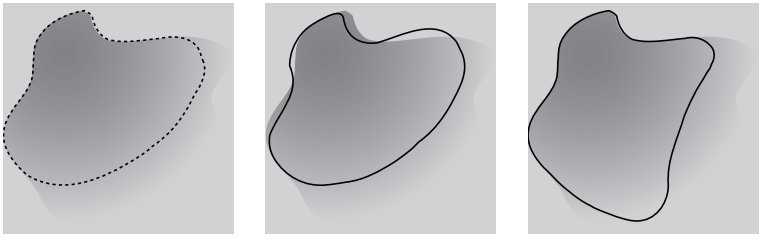
A. Error metric selection

Careful selection of an error metric is the first step in designing an effective level set estimator. While the goal of function estimation is typically to minimize the mean squared error between the true function and the estimate, in level set estimation it is more appropriate to minimize the symmetric difference between the level set of interest, S^*, and a candidate set, S, weighted by severity of the error over the symmetric difference:

$$\frac{1}{A}\int_{\Delta(S^*,S)} |\gamma - f(x)|\, d\mathbb{P}_X \tag{2}$$

where $\Delta(S^*, S) \equiv \{x \in (S^* \cap \bar{S}) \cup (\bar{S^*} \cap S)\}$ denotes the symmetric difference, and $\bar{S}$ is the complement of S. The normalization factor $1/A$ makes the error metric a unit-less quantity.

In effect, this metric penalizes two types of errors: first, the probability that a new observation will be located in a region which has been incorrectly designated by S as inside or outside the level set defined by S^*; second, the distance between the function value and the level of interest at each of these locations. For example, if level set estimation were being used on digital elevation data to assist in the navigation of unmanned helicopters, this metric would weight level set boundary localization errors in the vicinity of a cliff edge more heavily than localization errors in a plain, and thus minimize the potential for serious damage to the helicopter. This is illustrated in Figure 1. On the left is drawn a contour outlining the level set S^*. The center and rightmost figures show the boundary of two different candidate level sets. There is only a small symmetric difference between the set in the center image and the truth, but the distance of the function from the level γ is large in this region. In contrast, there is a large symmetric difference between the set in the rightmost image and the truth, but the distance of the function from the level γ is relatively small in that region.



(a) (b) (c)

Fig. 1. Behavior of level set error metric. (a) Function f and level set S^*. (b) Level set estimate (solid line) with a small symmetric difference but large errors within the symmetric difference region. (c) Second level set estimate (solid line) with same error as estimate in (b); this estimate has a large symmetric difference but small errors within the symmetric difference region. Despite these differences, these two set estimates could have the same weighted symmetric difference risk.

While the expression in (2) is not directly computable (since S is unknown), we will nevertheless be able

to minimize it through the following definitions. Let the *risk* of an candidate set S be defined as:

$$\mathcal{R}(S) \equiv \int \frac{\gamma - f(x)}{2A} \left[\mathbb{I}_{\{x \in S\}} - \mathbb{I}_{\{x \in \bar{S}\}} \right] d\mathbb{P}_X \tag{3}$$

where $\mathbb{I}_{\{E\}} = 1$ if event E is true and 0 otherwise. The loss function measures the distance between the function, f, and the threshold, γ, and weights the distance at location x by plus or minus one according to whether $x \in S$. (The $1/2A$ term is used for normalization. Its role will become more evident in the proceeding analysis sections, but for now it is sufficient to note that it is a constant scaling factor and does not change with S.) Thus regions where $x \in S$ but $f(x) < \gamma$ (that is, $x \in \bar{S}$) will contribute positively to the risk function. From here, we can define the *excess risk* to be

$$\mathcal{R}(S) - \mathcal{R}(S^*) = \frac{1}{A} \int_{\Delta(S^*,S)} |\gamma - f(x)| d\mathbb{P}_X, \tag{4}$$

which yields the desired weighted symmetric difference. Note that $\mathcal{R}(S^*)$ is a constant, so minimizing $\mathcal{R}(S)$ also minimizes the excess risk.

An additional advantage of the proposed metric is that it is simple to define an *empirical loss* metric for a candidate level set S as

$$\widehat{e}_S(X_i, Y_i) = \frac{\gamma - Y_i}{2A} \left[\mathbb{I}_{\{X_i \in S\}} - \mathbb{I}_{\{X_i \in \bar{S}\}} \right] \tag{5}$$

such that $\mathcal{R}(S) = \mathbb{E}\left[\widehat{e}_S\right]$. This results in the empirical risk function

$$\widehat{\mathcal{R}}_n(S) = \frac{1}{n} \sum_{i=1}^{n} \widehat{e}_S(X_i, Y_i). \tag{6}$$

This metric is distinctly different from the global L_p norms typically encountered in function estimation and image denoising.

B. Connection to error metrics in classification

As noted in the introduction, the problems of level set estimation and classification are closely linked, and the connection between the two can be illuminated by comparing and contrasting the proposed level set estimation metric with the generalization error of binary classification. For example, one might approach the level set estimation problem by simply thresholding the noisy observations at γ and then estimating the set where the probability of the thresholded observations being one is greater than 50% using binary classification techniques. More specifically, define $Z \equiv \mathbb{I}_{\{Y \geq \gamma\}}$. One might pose the level set estimation problem as estimating the one-half level set of the function $\eta(x) \equiv \mathbb{P}(Z = 1 | X = x)$. The Bayes classifier

is $S_C^* \equiv \{x \in [0,1]^d : \eta(x) \geq 1/2\}$. In this context, the excess risk is defined as the difference between

$$\mathcal{R}^C(S) \equiv \int \mathbb{I}_{\{Z \neq \mathbb{I}_{\{x \in S_C\}}\}} d\mathbb{P}_X$$

and $\mathcal{R}^C(S_C^*)$. Note that from here, the excess classification risk can be written as

$$\mathcal{R}^C(S) - \mathcal{R}^C(S_C^*) = \int_{\Delta(S_C^*, S_C)} |1/2 - \eta(x)| d\mathbb{P}_X,$$

which is highly analogous to the excess risk defined in (4). In fact, the two are equivalent up to a constant factor when $Y - f(X)$ obeys a uniform distribution over $[-B, B]$ for some $B > 0$. This analogy allows us to explore the difference between a classification-type approach as described in this subsection and the level set estimation method proposed in this paper. Tackling this problem from the classification standpoint and minimizing the excess classification risk would result in an estimate which minimizes the probability of a region in $[0,1]$ being incorrectly classified as in or out of the γ-level set of interest, but it would not weight this by the magnitude of the deviation between the function value and γ. The proposed level set estimation error metric, in contrast, does perform this weighting as described in the beginning of this section.

From another perspective, supposed we wished to design a classifier from training data, but rather than the traditional setup where every observation is labeled zero or one, we instead have observations labeled by our *confidence* that they lie in one of the classes. This setup arises in many practical situations, such as when each observation is the result of an experiment, and the class of the result is ambiguous. The proposed level set estimation method immediately leads to an optimal classifier which uses these confidence levels in a principled manner [18].

C. Alternative error norms

The excess risk proposed in (4) facilitates balancing between errors in level set *localization* and the *magnitude* of the deviation between f and γ in misclassified regions. It may be desirable to shift the fulcrum of this lever in some applications, however. This can be accomplished by setting

$$\begin{aligned}
\widehat{e}_S^q(X_i, Y_i) &\equiv \left(\frac{\gamma - Y_i}{2A}\right)^q \left(\mathbb{I}_{\{X_i \in S\}} - \mathbb{I}_{\{X_i \in \bar{S}\}}\right) \\
\mathcal{R}^q(S) &\equiv \mathbb{E}\left[\widehat{e}_S^q\right] = \mathbb{E}\left[\frac{1}{(2A)^q} \sum_{k=0}^{q} \binom{q}{k} (f(X_i) - Y_i)^k (\gamma - f(X_i))^{q-k} \left(\mathbb{I}_{\{X_i \in S\}} - \mathbb{I}_{\{X_i \in \bar{S}\}}\right)\right] \quad (7) \\
&= \frac{1}{(2A)^q} \sum_{k=0}^{q} \binom{q}{k} \left[\int_S \mu_k(x) (\gamma - f(x))^{q-k} d\mathbb{P}_X - \int_{\bar{S}} \mu_k(x) (\gamma - f(x))^{q-k} d\mathbb{P}_X\right] \quad (8) \\
\widehat{\mathcal{R}}_n^q(S) &\equiv \frac{1}{n} \sum_{i=1}^{n} \widehat{e}_S^q(X_i, Y_i).
\end{aligned}$$

where $q \geq 1$ is an odd integer and $\mu_k(x) \equiv \int (f(x) - y)^k d\mathbb{P}_{Y|X=x}$ can be considered the k^{th} moment of the "noise". Equation (7) follows from the binomial theorem. Note that if the distribution of $Y|X$ is such that $\mu_k(x) = 0$ for all x when k is an odd integer (*i.e.* if the noise distribution is always symmetric about the origin), then we would have an excess risk of

$$\mathcal{R}^q(S) - \mathcal{R}^q(S^*) = 2 \sum_{k=0}^{(q-1)/2} \binom{q}{2k} \int_{\Delta(S^*,S)} \frac{\mu_k(x)}{(2A)^q} \, |\gamma - f(x)|^{q-2k} \, d\mathbb{P}_X.$$

For $q = 1$, this is precisely the excess risk defined in (4). For $q = 0$, this reduces to binary classification as discussed above. By choosing a larger q, we emphasize correct localization more heavily in regions where f varies significantly in the vicinity of the level set. Performance bounds similar to the ones detailed later in this paper can be derived for these alternative error metrics; for details, see our technical report [19].

III. Estimation via Trees

We propose to estimate the level set of a function from noisy observations by using a tree-pruning method akin to CART [20] or dyadic decision trees [16]. Trees are emphasized for a number of reasons. First, they allow us to rapidly compute a globally optimal estimate. In other words, it adds a structure to the space of candidate sets which allows us to organize our search through the solution space and minimize the computational burden. Furthermore, it allows us a convenient mechanism for penalizing different estimates and incorporating prior knowledge (such as the presence of noise and our subsequent desire to avoid improbably complex level set estimates). As we will see later in this paper, an additional key feature of trees is that they allow us to incorporate the idea of spatial adaptivity into our estimation procedure.

In this section, we will define a collection of candidate tree-based sets, $\mathcal{T}_M$, and a "penalty" or regularization term for each tree $T \in \mathcal{T}_M$, $\Phi_n(T)$. We will then define our estimator to be

$$\widehat{T}_n = \arg\min_{T \in \mathcal{T}_M} \widehat{\mathcal{R}}_n(T) + \Phi_n(T). \tag{9}$$

Careful selection of $\mathcal{T}_M$ and $\Phi_n(T)$ are critical to the performance of the proposed method. The collection $\mathcal{T}_M$ must be big enough that some $T \in \mathcal{T}_M$ is a close approximation to S, yet small and structured enough that a computationally tractable optimization algorithm is feasible. The penalty $\Phi_n(T)$ must be designed to favor estimates which correspond to our prior knowledge – that the observations are bounded, and that our objective is to localize the boundary of a set rather than perform signal estimation or regression. The goal of this section is to propose and justify selections of $\mathcal{T}_M$ and $\Phi_n(T)$. Sections IV and VI will then explore the resulting error performance and computational efficiency, respectively, in detail.

Before discussing regularization methods, we first specify the class of tree-based estimators under consideration and define the associated notation. When the observations lay in a relatively low dimensional space

(*i.e.* $d = 2$ or 3), a 2^d-ary tree can be an effective tool. In general, however, 2^d-ary trees are only appropriate in regimes where the number of observations is much larger than the dimensionality of the observations; for large d, growing a 2^d-ary tree with just one level results in 2^d cells, which leads to several computational and statistical problems. To avoid this issue, we will focus on binary trees, where each internal node in the tree corresponds to a dyadic split of a hyperrectangular subset of $[0,1]^d$ in one of the d coordinate directions.

Each node of a tree T corresponds to a dyadic hyperrectangle in $[0,1]^d$, with the root of T corresponding exactly to $[0,1]^d$. If node v of T corresponds to the hyperrectangle $R = \prod_{r=1}^{d}[a_r, b_r]$, and v is split in coordinate direction s to form its children v_{left} and v_{right}, then v_{left} corresponds to the hyperrectangle $R^{s,\text{left}} = \{x \in R : x_s \leq (a_s + b_s)/2\}$, where x_s is the s^{th} coordinate of x, and v_{right} corresponds to the hyperrectangle $R^{s,\text{right}} = R \setminus R^{s,\text{left}}$. Let $\pi(T)$ denote the partition induced on $[0,1]^d$ by the binary tree T. That is, if T has k leaf nodes (denoted $|T| = k$), then $\pi(T) = \{L_1, L_2, \ldots, L_k\}$ so that each $L_i \in \pi(T)$ is a dyadic hyperrectangular cell for some j. The location and size of each L_i is determined by the sequence of coordinate directions assigned to the ancestors of the corresponding leaf node in T. Note that $\bigcup_i L_i = [0,1]^d$, and $\lambda(L_i \cap L_j) = 0$ for $i \neq j$. (Recall λ denotes Lebesgue measure.) A zero or one is assigned to each leaf node of T (equivalently, to each $L \in \pi(T)$), denoted $\ell(L)$, to indicate whether that cell of the partition is estimated to be in $\widehat{S}$. A sample tree and corresponding partition in displayed in Figure 2. We will consider all binary trees with at most $M = 2^J$ leaf nodes, for some nonnegative integer J, and denote this collection $\mathcal{T}_M$. Note that the sidelength of each cell must then necessarily be no less than 2^{-J}.

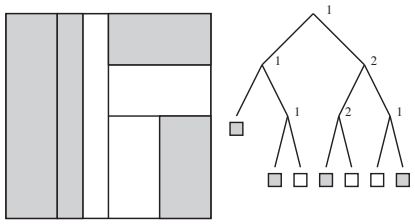


Fig. 2. Sample partition and corresponding tree.

Ideally, the penalty term $\Phi_n(T)$ in (9) will produce an estimate $\widehat{T}_n$ such that the risk $\mathcal{R}(\widehat{T}_n)$ is as small as possible. If we choose $\Phi_n(T)$ such that $\mathcal{R}(T) \leq \widehat{\mathcal{R}}_n(T) + \Phi_n(T)$ for all $T \in \mathcal{T}_M$, then the estimator in (9) will minimize an upper bound on the risk and potentially result in an effective estimator. Towards that end, first note that it is possible to express the difference between the true risk and the empirical risk as

$$\mathcal{R}(T) - \widehat{\mathcal{R}}_n(T) = \sum_{L \in \pi(T)} \mathcal{R}(L) - \widehat{\mathcal{R}}_n(L),$$

where

$$\begin{aligned}\widehat{e}_L(X_i,Y_i) &= \frac{\gamma - Y_i}{2A}\left[\mathbb{I}_{\{\ell(L)=1\}} - \mathbb{I}_{\{\ell(L)=0\}}\right]\mathbb{I}_{\{X_i\in L\}} \\ \mathcal{R}(L) &= \mathbb{E}\left[\widehat{e}_L\right] = \int \frac{\gamma - f(x)}{2A}\left[\mathbb{I}_{\{\ell(L)=1\}} - \mathbb{I}_{\{\ell(L)=0\}}\right]\mathbb{I}_{\{x\in L\}} d\mathbb{P}_X \\ \widehat{\mathcal{R}}_n(L) &= \frac{1}{n}\sum_{i=1}^{n}\widehat{e}_L(X_i,Y_i).\end{aligned}$$

These definitions are simply local counterparts of the risk and loss expressions defined in equations (3), (5), and (6). In addition, define $p_L \equiv \int_L d\mathbb{P}_X$ and $\widehat{p}_L \equiv \frac{1}{n}\sum_{i=1}^{n}\mathbb{I}_{\{X_i\in L\}}$. We can now bound the difference between $\mathcal{R}(L)$ and $\widehat{\mathcal{R}}_n(L)$ to achieve a bound leading to near-optimal error decay rates. In particular, we can use the *relative* form of Hoeffding's inequality [21] to show that for a single tree leaf L, with high probability the risk $\mathcal{R}(L)$ is bounded by $\widehat{\mathcal{R}}_n(L)$ plus a small quantity that depends on δ_L:

Lemma 1 *Let $\delta_L \in [0,1]$. With probability at least $1-\delta_L$ for any $T \in \mathcal{T}_M$ and any $L \in \pi(T)$,*

$$\mathcal{R}(L) - \widehat{\mathcal{R}}_n(L) < \sqrt{\frac{8\log(1/\delta_L)p_L}{n}} - (p_L - \widehat{p}_L)\,.$$

This is proved in Section X. This lemma shows that for a given tree leaf L, the empirical risk $\widehat{\mathcal{R}}_n(L)$ can be used to estimate the true risk $\mathcal{R}(L)$ with a certain degree of accuracy and confidence. Now let $\mathcal{L}_M$ be the collection of all L such that $L \in \pi(T)$ for some $T \in \mathcal{T}_M$. We wish to show that a similar bound holds for all $L \in \mathcal{L}_M$ simultaneously. This is necessary because we ultimately are going to perform a search over all $T \in \mathcal{T}_M$ and hence consider all $L \in \mathcal{L}_M$; conducting this search is a valid approach only if the bound holds for each leaf considered during the search procedure. To do so, set $\delta_L = \delta 2^{-(\llbracket L\rrbracket+1)}$, where $\llbracket L\rrbracket$ denotes the number of bits required to encode the position of L. Specifically, consider the prefix code proposed in [16] for $L \in \pi(T)$. If L is at level j in the binary tree T, then $j+1$ bits must be used to describe the depth of L, j bits must be used to describe whether each branch is a left or right branch, and $j\log_2 d$ bits must be used to describe the coordinate direction of each of the j branches. This results in a total of $j(\log_2 d + 2) + 1$ bits, and this expression is denoted as $\llbracket L\rrbracket$. From here we can derive the following lemma, which proposes a candidate regularization term $\Phi'_n(T)$ and states that $\mathcal{R}(T) \le \widehat{\mathcal{R}}_n(T) + \Phi'_n(T)$ with high probability:

Lemma 2 *Let*

$$\Phi'_n(T) \equiv \sum_{L\in\pi(T)} \sqrt{\frac{8\left(\log(2/\delta) + \llbracket L\rrbracket\log 2\right)p_L}{n}}. \tag{10}$$

With probability at least $1-\delta$, $\mathcal{R}(T) \le \widehat{\mathcal{R}}_n(T) + \Phi'_n(T) \quad \forall\, T \in \mathcal{T}_M$.

This is proved in Section X. In effect, this expression penalizes a tree-based estimate by independently

penalizing each leaf in the tree, where each leaf's penalty is proportional to the square root of the probability measure of the leaf (i.e. the likelihood of an observation being collected in the function domain associated with the leaf) and proportional to the square root of the depth of the leaf in the tree. Intuitively, this penalty is designed to favor unbalanced trees which focus on the location of the manifold defining the boundary of the level set; the same reasoning explains why spatially adaptive penalties are so effective for binary classification problems using dyadic decision trees [16]. To see this, note that $[\![L]\!] \asymp j$, while $\widehat{p}'_L(\delta) \asymp 2^{-j}$. This implies that deep nodes contribute less to $\Phi'_n(T)$ than shallow nodes, and so, for two trees with the same number of leafs, $\Phi'_n(T)$ will be smaller for the more unbalanced tree.

Lemma 2 implies that if we were to select a tree which minimized the sum $\widehat{\mathcal{R}}_n(T) + \Phi'_n(T)$, then our estimator would have a low risk with high probability. However, since p_L must be known to compute $\Phi'_n(T)$, such a strategy is not practically feasible. We can, however, produce a computable bound as follows. First, we recall the following lemma, which was proved in [16]. For $\delta \in [0,1]$, define $\widehat{p}'_L(\delta) \equiv 4\max\left(\widehat{p}_L\,,\,\frac{[\![L]\!]\log 2+\log(1/\delta)}{n}\right)$ and $p'_L(\delta) \equiv 4\max\left(p_L\,,\,\frac{[\![L]\!]\log 2+\log(1/\delta)}{2n}\right)$.

Lemma 3 *[16] Let $\delta \in [0,1]$. Then with probability at least $1-\delta$, $p_L \leq \widehat{p}'_L(\delta) \;\; \forall\; L \in \mathcal{L}_M$, and with probability at least $1-\delta$, $\widehat{p}_L \leq p'_L(\delta) \;\; \forall\; L \in \mathcal{L}_M$.*

Using this lemma, we can design a penalty term Φ_n based on Φ'_n which is computable from the data themselves:

$$\Phi_n(T) \equiv \sum_{L \in \pi(T)} \sqrt{\frac{8\left(\log(2/\delta) + [\![L]\!]\log 2\right)\widehat{p}'_L(\delta)}{n}}. \tag{11}$$

This penalty can be understood using the same arguments used to explain the uncomputable penalty (10) from Lemma 2. As we will see in the next section, this penalty leads to near minimax optimal error decay rates for broad classes of level set problems. Note that the form of this penalty is *distinctly different* from penalty terms used to achieve near minimax optimal error decay rates for function estimation and image denoising. For example, [22] demonstrated that the optimal tree-based penalty for denoising applications had the form $\Phi_n^{\text{denoising}}(T) = \sum_{L \in \pi(T)} c$ for some constant $c > 0$; that is, it was proportional to the *number of leafs in the tree*. In contrast, the penalty term we have developed for near minimax optimal level set estimation *does not weigh all leafs equally*; rather, "smaller" leafs (with respect to the underlying probability measure) receive a smaller penalty, allowing us to favor the deep, unbalanced trees most likely to accurately represent the true level set of interest.

We now have the following main result:

Theorem 4 *Let $\Phi_n(T)$ be defined as in (11). Then with probability at least $1 - 2\delta$,*

$$\mathcal{R}(T) \leq \widehat{\mathcal{R}}_n(T) + \Phi_n(T)$$

for all $T \in \mathcal{T}_M$.

Proof of Theorem 4 This follows trivially from Lemmas 2 and 3. ■

The key impact of Theorem 4 is to give us a principled way to choose a good tree-based level set estimate: by choosing an estimate which minimizes $\widehat{\mathcal{R}}_n(T) + \Phi_n(T)$, we are also choosing an estimate which makes the excess risk small with very high probability. In addition, the estimator defined by (9) and (11) is rapidly computable, as described in Section VI. Furthermore, as shown in the following section, the estimator is nearly minimax optimal. This is a key feature which, to the best of our knowledge, is not possible without a spatially adaptive "leaf-wise" penalty as we have here. For example, as a first pass one might have chosen to determine how well $\widehat{\mathcal{R}}_n(T)$ predicts $\mathcal{R}(T)$ is through (the classical, non-relative form of) Hoeffding's inequality. However, this form of concentration inequality is not as tight as the relative form, especially in cells with very low p_L, and so leads to suboptimal estimators.

IV. Performance Analysis

Many of the key points in the following theoretical analysis were derived using the error bounding techniques developed by Scott and Nowak [16] in the context of binary classification. As discussed in the introduction, there are many striking similarities between classification and level set estimation, and we take advantage of these similarities and the extensive literature available on classification to analyze the performance capabilities of the proposed level set estimators.

Not only does the above framework give us a principled way to choose a good level set estimator, but it also allows us to bound the expected risk for a collection of n observations. In particular, we have the following theorem:

Theorem 5 *Let $\widehat{T}_n$ be as in (9) with $\Phi_n(T)$ as in (11), with $\delta = 1/n$. Let*

$$\widetilde{\Phi}_n(T) \equiv \sum_{L \in \pi(T)} \sqrt{32 p'_L \frac{\log(2n) + [\![L]\!] \log 2}{n}}$$

be a data-independent analog of $\Phi_n(T)$. With probability at least $1 - 3/n$,

$$\mathcal{R}(\widehat{T}_n) - \mathcal{R}(S^*) \leq \min_{T \in \mathcal{T}_M} \left\{ \mathcal{R}(T) - \mathcal{R}(S^*) + 2\widetilde{\Phi}_n(T) \right\}.$$

As a consequence,

$$\mathbb{E}^n\left[\mathcal{R}(\widehat{T}_n) - \mathcal{R}(S^*)\right] \leq \min_{T\in\mathcal{T}_M}\left(\mathcal{R}(T) - \mathcal{R}(S^*) + 2\widetilde{\Phi}_n(T)\right) + \frac{3}{n}.$$

Proof of Theorem 5 This can be proved by closely following the proof of Theorem 3 in [16] and using the above lemmas and theorems. ■

Note that in the above theorem we take $\delta = 1/n$ in (11). This setting is optimal in that it leads to near-minimax optimal estimators, as we will see later in this section. The bound on the expected error in Theorem 5 allows us to analyze the proposed method in terms of rates of error convergence. In particular, because this problem has been posed as a generalization of the binary classification problem, we may draw extensively from the results of [16] to highlight several advantageous features of the error convergence of the proposed method. In this analysis, for sequences a_n and b_n let the notation $a_n \preceq b_n$ imply there exists some $C > 0$ such that $a_n \leq Cb_n$ for all n. We will examine rates of convergence for two classes of functions f and distributions $\mathbb{P}_X$ to demonstrate that the proposed method both (a) adapts to the regularity of f in the vicinity of ∂S^* and (b) adapts to the regularity of the curve ∂S^*.

A. Adaptation to the regularity of f

Define $\mathcal{D}_{\mathrm{BOX}}(\kappa, \gamma, c_0, c_1, c_2)$, for $c_0, c_1, c_2 > 0$ and $1 \leq \kappa \leq \infty$ to be the set of all $(f, \mathbb{P}_X)$ pairs such that

1) $\mathbb{P}_X(L) \leq c_0\lambda(L)$ for all measurable $L \subseteq [0,1]^d$;
2) for $S = \{x \in [0,1]^d : f(x) \geq \gamma\}$, ∂S is in a box-counting class; *i.e.* if $[0,1]^d$ is partitioned into m^d equal sized cells, each with sidelength $1/m$ and volume m^{-d}, and $N_S(m)$ denotes the number of such cells intersected by the boundary of S, then $N_S(m) \leq c_1 m^{d-1}$ for all m; and
3) for all dyadic m, there exists some $T'_m \in \mathcal{T}_m$ such that $\mathcal{R}(T'_m) - \mathcal{R}(S^*) \leq c_2/m^\kappa$.

This class is designed to include functions whose level sets are *not* highly-irregular space-filling curves. The role of condition III is to provide a means of exploring how the estimator behaves for various amplitudes of f in the vicinity of ∂S^*. Recall

$$\mathcal{R}(T'_m) - \mathcal{R}(S^*) = \frac{1}{A}\int_{\Delta(T'_m, S)} |\gamma - f(x)| d\mathbb{P}_X.$$

For large κ, the third condition will only hold when f is close to γ in the vicinity of ∂S^* (e.g. when the slope of f is very small near ∂S^*). In this case, it will be very difficult to estimate S accurately, but the estimate could be wrong over a large volume and still only incur a small error, resulting in faster rates of convergence. In contrast, small κ allows a jump in f near ∂S^* (e.g. a very large slope of f) and so is a relatively "easy" problem (i.e. edge detection); however, an estimate which is incorrect on a small volume

will result in a large error, resulting in slower rates of convergence. Condition III allows us to examine the performance of the proposed estimator under these different conditions.

This definition, when combined with Theorem 3 and Lemma 8 of [16] yields the following result:

Theorem 6 *Choose M such that $M \succeq (n/\log n)^{1/d}$. For $d \geq 2$ we have*

$$\sup_{\mathcal{D}_{\mathrm{BOX}}(\kappa,\gamma,c_0,c_1,c_2)} \left\{ \mathbb{E}^n \left[\mathcal{R}(\widehat{T}_n) - \mathcal{R}(S^*) \right] \right\} \preceq \min_m \left\{ m^{-\kappa} + m^{d/2-1} \sqrt{\frac{\log n}{n}} \right\} \preceq \left(\frac{\log n}{n} \right)^{\frac{\kappa}{d+2\kappa-2}}. \quad (12)$$

That is, the expected value of the excess risk for the hardest possible problem in $\mathcal{D}_{\mathrm{BOX}}$ decreases with n at a rate which *adapts* to κ, while the algorithm has no prior knowledge of κ. The derivation on the lower bounds presented in [16] should carry over to the regression level set estimation context with only minor modifications, which would mean that the rate in Theorem 6 is within a log factor of the minimax lower bound and hence near optimal.

This theorem may be proved using an argument very similar to the one used to prove Theorem 6 of [16]. In particular, condition III above leads to the $m^{-\kappa}$ term in (12), and Lemma 8 of [16] leads to the second term in (12). Careful examination of minimax lower bounds derived in the context of binary classification (Theorem 5, [16]) indicates that the same lower bounds should hold for the level set estimation problem posed here. Specifically, recall that the excess risk in this setting is $\mathcal{R}(T) - \mathcal{R}(S^*) = \frac{1}{A} \int_{\Delta(S^*,T)} |f(x) - \gamma| d\mathbb{P}_X$. The excess risk in the classification setting has the same form, except $\gamma = 1/2$, $f \geq 0$, and $\int f(x) d\mathbb{P}_X = 1$.

B. Adaptation to the regularity of ∂S^*

It is also possible to show that the proposed method adapts to the regularity of ∂S^*. In particular, define the "boundary fragment" class $\mathcal{D}_{\mathrm{BF}}(\alpha, \gamma, c_0, c_1)$ for $\alpha < 1$ to be the set of all $(f, \mathbb{P}_X)$ such that

1) $\mathbb{P}_X(L) \leq c_0 \lambda(L)$ for all measurable $L \subseteq [0,1]^d$; and
2) one coordinate of ∂S^* is a function of the others, where the function has Hölder smoothness $\alpha < 1$ and constant c_1.

This class imposes constraints on the regularity of the boundary of the level set S^*. The analysis here would also apply to compositions of members of $\mathcal{D}_{\mathrm{BF}}$, even if the "orientation" (which coordinate is a function of the others) varies member to member. The orientation does not need to be known. Following the argument Theorem 9 in [16], it is straightforward to prove the following theorem:

Theorem 7 *Choose M such that $M \succeq (n/\log n)^{1/(d-1)}$. For $d \geq 2$ and $\alpha < 1$, we have*

$$\sup_{\mathcal{D}_{\mathrm{BF}}(\alpha,\gamma,c_0,c_1)} \left\{ \mathbb{E}^n \left[\mathcal{R}(\widehat{T}_n) - \mathcal{R}(S^*) \right] \right\} \preceq \min_m \left\{ m^{-\alpha} + m^{(d-\alpha-1)/2} \sqrt{\frac{\log n}{n}} \right\} \preceq \left(\frac{\log n}{n} \right)^{\frac{\alpha}{\alpha+d-1}}.$$

That is, the expected excess risk is uniformly bounded above for all boundaries in $\mathcal{D}_{\mathrm{BF}}$ and the upper bound is adaptive to the smoothness of the boundary of the level set even though the proposed method assumes no prior knowledge of this smoothness. As before, an examination of the lower bounds derived in [16] indicates that this rate is within a log factor of the minimax optimal rate. Note that $\alpha < 1$ implies fractal-like boundaries, which are important for a variety of scientific applications and similar to the ones displayed in the simulation section later in this paper.

C. Scaling the penalty

The above theoretical analysis demonstrates that the estimator in (9) can be derived from fundamental statistical concentration inequalities and results in a principled regularization strategy with minimax near-optimal rates of error convergence. We now explore the ramifications of damping the regularization term. In many practical scenarios, the theoretical penalty results in over-regularized (*i.e.* over-smoothed) estimates, and damping the regularization term can result in substantial empirical improvements. Since the theoretical analysis focuses on rates of convergence, multiplication by the empirical risk or penalty should not effect the error bounds, aside from possibly altering the leading constant factors. This insight is formalized in the following result communicated to the authors by Clayton Scott [23].

Let

$$\widehat{T}_n^{\rho} \equiv \operatorname*{arg\,min}_{T \in \mathcal{T}_M} \widehat{\mathcal{R}}_n(T) + \rho \Phi_n(T) \tag{13}$$

for $\rho > 0$ be the damped version of the proposed set estimator; typically we are interested in $0 < \rho < 1$.

Theorem 8 *[23] With probability at least* $1 - 2\delta$

$$\mathcal{R}(\widehat{T}_n^{\rho}) - \mathcal{R}(S^*) \leq \max\left(\frac{1}{\rho}, 1\right) \min_{T \in \mathcal{T}_M} \left\{\mathcal{R}(T) - \mathcal{R}(S^*) + (1+\rho)\Phi_n(T)\right\}.$$

This is proved in Section X. Theorem 8 implies that we may compute an estimate using a damped form of the regularization term $\Phi_n(T)$ without impacting the minimax near-optimal rates derived earlier in this section.

V. Alternative Level Set Estimation Problems

A. Samples on a Regular Lattice

The above analysis assumes that the locations of the observations are random and distributed according to $\mathbb{P}_X$. In many domains, however, the locations of the observations are dictated by the measurement device and lay on a grid (*e.g.* pixels in an image or voxels in a volume). In this case, the analysis above must be modified slightly, as described in this section. Specifically, we will examine two types of sampling: (a)

integration sampling and (b) point sampling. It suffices to consider cases where $n = 2^{dk}$ for some k, so that n points can be distributed uniformly on $[0,1]^d$. Specifically, divide $[0,1]^d$ into n equally sized non-intersecting hypercubes, $C_1, \ldots, C_n$, where the volume of C_i is $1/n$ and its sidelength is $n^{-1/d}$

For the case of integration sampling, let $f_i \equiv \frac{1}{|C_i|}\int_{C_i} f(x)dx$ and $\mathbb{E}[Y_i] = f_i$. The empirical loss and risk metrics remain the same as they were in the random observation location case, and the terms $\frac{A}{2A+B}\widehat{e}_T(X_i, Y_i)$ remain independent and lie within the interval $[0,1]$. Next define

$$e_T(x) \equiv \frac{\gamma - f(x)}{2A}\left[\mathbb{I}_{\{x\in T\}} - \mathbb{I}_{\{X_i \in T^c\}}\right].$$

Note that this is the integrand in the risk function proposed under random sampling conditions. We define the risk in the case of gridded observations to be $\mathcal{R}^G(T) = \int e_T(x)dx$, and the quantity $\mathcal{R}^G(L)$ is defined accordingly. It now becomes necessary to derive a corollary to Lemma 1 and bound the difference $\mathcal{R}^G(L) - \widehat{\mathcal{R}}_n(L)$. Using this setup, we have

$$\begin{aligned}
&\mathcal{R}^G(T) - \widehat{\mathcal{R}}_n(T) \\
&= \left(\mathcal{R}^G(T) - \frac{1}{n}\sum_{i=1}^n e_T(X_i, Y_i)\right) + \left(\frac{1}{n}\sum_{i=1}^n e_T(X_i, Y_i) - \widehat{\mathcal{R}}_n(T)\right) \\
&= \sum_{i=1}^n \left(\int_{C_i} \frac{\gamma - f(x)}{2A}dx - \frac{\gamma - f_i}{2nA}\right)\left[\mathbb{I}_{\{C_i\in T\}} - \mathbb{I}_{\{C_i \in T^c\}}\right] + \left(\frac{1}{n}\sum_{i=1}^n e_T(X_i, Y_i) - \widehat{\mathcal{R}}_n(T)\right) \\
&= \sum_{i=1}^n \int_{C_i} \frac{f(x) - f(x)}{2A}dx + \left(\frac{1}{n}\sum_{i=1}^n e_T(X_i, Y_i) - \widehat{\mathcal{R}}_n(T)\right) \\
&= \left(\frac{1}{n}\sum_{i=1}^n e_T(X_i, Y_i) - \widehat{\mathcal{R}}_n(T)\right).
\end{aligned}$$

For point sampling, we assume $\mathbb{E}[Y_i] = f(X_i)$ as in the random sample location case, except that X_i lies on the i^{th} of n uniformly placed grid points on $[0,1]^d$. Specifically, let X_i be the center of C_i. In the point sampling case, we must assume something about the smoothness of f; specifically, assume that for some constant $L > 0$, f satisfies the Lipschitz condition

$$|f(x_1) - f(x_0)| \leq L\|x_1 - x_0\|_2^1 \quad \forall\, x_0, x_1 \in [0,1]^d. \tag{14}$$

If f meets the criterion in (14), then e_T also satisfies the criterion in the interior of each C_i. Note that

$$\begin{aligned}
\mathcal{R}^G(T) - \widehat{\mathcal{R}}_n(T) &= \left(\mathcal{R}^G(T) - \frac{1}{n}\sum_{i=1}^n e_T(X_i, Y_i)\right) + \left(\frac{1}{n}\sum_{i=1}^n e_T(X_i, Y_i) - \widehat{\mathcal{R}}_n(T)\right) \\
&= \sum_{i=1}^n \int_{C_i} (e_T(x) - e_T(X_i, Y_i))\, dx + \left(\frac{1}{n}\sum_{i=1}^n e_T(X_i, Y_i) - \widehat{\mathcal{R}}_n(T)\right) \\
&\leq \frac{L\sqrt{d}}{2A} n^{-1/d} + \left(\frac{1}{n}\sum_{i=1}^n e_T(X_i, Y_i) - \widehat{\mathcal{R}}_n(T)\right) \\
&= O(n^{-1/d}) + \left(\frac{1}{n}\sum_{i=1}^n e_T(X_i, Y_i) - \widehat{\mathcal{R}}_n(T)\right),
\end{aligned}$$

where the inequality holds because of the condition on f and because each C_i is either completely in T or completely in T^c. Note that, unlike integration sampling, point sampling results in a limit on the rate of risk decay at $O(n^{-1/d})$. This is the minimax optimal rate when $d = 2$ or when $\kappa = 1$ (*i.e.* when the level set boundary corresponds to an edge or boundary in the function), but is slower than the optimal rate otherwise.

In both the point sampling and integration sampling cases, it becomes necessary to bound the expression $\left(\frac{1}{n}\sum_{i=1}^n e_T(X_i, Y_i) - \widehat{\mathcal{R}}_n(T)\right)$. Our analysis from the random sample locations example can easily be used to show the following analog to Theorem 4.

Corollary 9 *Let $\delta \in [0, 1]$, and define*

$$\Phi_n^G(T) \equiv \sum_{L \in \pi(T)} \sqrt{\frac{8\left(\log(2/\delta) + [\![L]\!] \log 2\right)}{n} \lambda(L)},$$

where λ denotes Lebesgue measure. With probability at least $1 - \delta$ for all $T \in \mathcal{T}_M$,

$$\frac{1}{n}\sum_{i=1}^n e_T(X_i, Y_i) - \widehat{\mathcal{R}}_n(T) < \Phi^G(T).$$

B. Simultaneous extraction of multiple level sets

In some applications one may wish to extract a collection of level sets, *e.g.* for the generation of a contour plot. This has proven to be very useful in the context of digital elevation map storage and retrieval [1]. Simultaneous (rather than sequential) extraction of multiple level sets is important when it ensures that the estimated level sets are nested. This is critical because the value of f at a point x cannot both be both greater than γ_{big} and less than γ_{small} when $\gamma_{\text{big}} > \gamma_{\text{small}}$.

Consider the case in which we are interested in a collection of levels, $\{\gamma_k\}_{k=1}^K$, where $-A \leq \gamma_1 < \gamma_2 < \cdots < \gamma_K \leq A$. For each γ_k, let $S_k^* \equiv \{x \in [0, 1]^d : f(x) \geq \gamma_k\}$, so that $[0, 1]^d \supseteq S_1^* \supseteq S_2^* \supseteq \cdots \supseteq S_K^*$, and denote the collection of level sets as $\mathcal{S}^* \equiv \{S_k^*\}_{k=1}^K$. In order to ensure that the estimated level sets are

likewise nested, we will solve the equivalent problem of estimating the function

$$g_{\mathcal{S}^*}(x) \equiv \max\{k \in \{0,1,\ldots,K\} : x \in S_k^*\}. \tag{15}$$

Thus for every $x \in [0,1]^d$, we have that $x \in S_k^*$ for all $k \in \{0,\ldots,g_{\mathcal{S}^*}(x)\}$. Note that there is a bijective relationship between the collection of potential level sets $\mathcal{S}^*$ and the space of functions $g_{\mathcal{S}^*} : [0,1]^d \longrightarrow \{0,1,\ldots,K\}$. In particular, for $S_k^* \in \mathcal{S}^*$, $S_k^* = \{x \in [0,1]^d : g_{\mathcal{S}^*}(x) \geq k\}$. By estimating the function $g_{\mathcal{S}^*}(x)$, we necessarily estimate a nested collection of level sets, as desired. To see this, let $g_S(x)$ denote the estimate. For any $x \in [0,1]^d$, having $g_S(x) = k$ implies $x \in S_k$ ($f(x) \geq \gamma_k$), and hence we must have $x \in S_{k-1}$ ($f(x) \geq \gamma_{k-1}$) because $\gamma_k > \gamma_{k-1}$.

Let $\mathcal{S} \equiv \{S_k\}_{k=1}^K$ be a candidate collection of level sets. For multiple level sets, let the empirical loss be computed as

$$\begin{aligned}
\widehat{e}_{\mathcal{S}}^K(X_i,Y_i) &= \frac{1}{K}\sum_{k=1}^K \widehat{e}_{S_k}(X_i,Y_i) = \frac{1}{2KA}\sum_{k=1}^K (\gamma_k - Y_i)\left[\mathbb{I}_{\{X_i \in S_k\}} - \mathbb{I}_{\{X_i \in S_k^c\}}\right] \\
&= \frac{1}{2KA}\sum_{k=1}^K (\gamma_k - Y_i)\left[2\mathbb{I}_{\{g_{\mathcal{S}}(X_i)\geq k\}} - 1\right].
\end{aligned}$$

Note that this empirical loss is increased for each incorrect level set in which x is placed by $\mathcal{S}$. If we define the risk as before to be $\mathcal{R}^K(\mathcal{S}) \equiv \mathbb{E}\left[\widehat{e}_{\mathcal{S}}^K\right]$, then the excess risk is simply

$$\mathcal{R}^K(\mathcal{S}) - \mathcal{R}^K(\mathcal{S}^*) = \frac{1}{K}\sum_k (\mathcal{R}(S_k) - \mathcal{R}(S_k^*)) = \frac{1}{KA}\sum_{k=1}^K \int_{\Delta(S_k^*,S_k)} |\gamma_k - f(x)| d\mathbb{P}_X.$$

As before, the empirical risk can be computed as $\widehat{\mathcal{R}}^k(\mathcal{S}) = \frac{1}{n}\sum_{i=1}^n \widehat{e}_{\mathcal{S}}^K(X_i,Y_i)$.

In this context, regularization terms can be developed using the same methodology as described above. Let $\ell(L) \in \{0,1,\ldots,K\}$ denote the label of leaf L, which corresponds to the maximum k such that L is estimated to belong to the level set S_k^*, as described above. Also, let

$$\begin{aligned}
\widehat{e}_L^K(X_i,Y_i) &\equiv \sum_{k=1}^K \frac{\gamma_k - Y_i}{2AK}\left[\mathbb{I}_{\{\ell(L)\geq k\}} - \mathbb{I}_{\{\ell(L)<k\}}\right]\mathbb{I}_{\{X_i \in L\}} \\
\mathcal{R}^K(L) &\equiv \mathbb{E}\left[\widehat{e}_L^K\right] = \int \sum_{k=1}^K \frac{\gamma_k - f(x)}{2AK}\left[\mathbb{I}_{\{\ell(L)\geq k\}} - \mathbb{I}_{\{\ell(L)<k\}}\right]\mathbb{I}_{\{x \in L\}} d\mathbb{P}_X \\
\widehat{\mathcal{R}}_n^K(L) &\equiv \frac{1}{n}\sum_{i=1}^n \widehat{e}_L^K(X_i,Y_i).
\end{aligned}$$

These definitions lead to the following corollary:

Corollary 10 *For $K \geq 1$ an integer, let*

$$\Phi_n^K(T) \equiv \sum_{L \in \pi(T)} \sqrt{\frac{8\left(\log(1/\delta) + [\![L]\!] \log 2 + \log(K+1)\right) \widehat{p}'_L(\delta)}{n}}.$$

Then with probability at least $1 - 2\delta$, $\mathcal{R}^K(T) \leq \widehat{\mathcal{R}}_n^K(T) + \Phi_n^K(T)$ for all $T \in \mathcal{T}_M$.

This is proved in Section X. The corollary suggests that $\widehat{T}_n = \arg\min_{T \in \mathcal{T}_M} \widehat{\mathcal{R}}_n^K(T) + \Phi_n^K(T)$ would be an effective estimator in this setting. Each leaf of this tree has a label $k \in \{0, 1, \ldots, K\}$ which denotes the highest (and hence smallest) level set S_k^* to which the corresponding domain of f is estimated to belong. The optimization method used to compute this estimate is detailed in Section VI.

C. Density level set estimation

The above proposed error metric can be altered slightly to become appropriate for density level set estimation. Specifically, given n iid observations of some density $f : [0,1]^d \longrightarrow [0, A]$, $A \geq 1$ we wish to identify the set $S^* \equiv \{x \in [0,1]^d : f(x) \geq \gamma\}$ as before. This problem is important in a variety of applications because of its close ties to anomaly detection. For example, researchers studying computer networks may wish to identify worms or malicious agents by identifying packets with "unusual" routing behavior. By estimating the level set of a density, we estimate patterns of behavior whose likelihoods are below some critical level.

The formulation of this problem is slightly different from the one above because we do not make noisy observations of the amplitude of f. Let $\lambda(S)$ denote the Lebesgue measure of S, and define the empirical loss to be

$$\widehat{e}_S^D(X_i) \equiv \frac{1}{2A}\left[\left(\gamma\lambda(S) - \mathbb{I}_{\{X_i \in S\}}\right) - \left(\gamma\lambda(\bar{S}) - \mathbb{I}_{\{X_i \in \bar{S}\}}\right)\right]. \tag{16}$$

To see why this is an effective metric in this context, first observe that

$$\mathcal{R}^D(S) \equiv \mathbb{E}\left[\widehat{e}_S^D\right] = \int \frac{\gamma - f(x)}{2A}\left[\mathbb{I}_{\{x \in S\}} - \mathbb{I}_{\{x \in \bar{S}\}}\right] dx \tag{17}$$

and the excess risk is $\mathcal{R}^D(S) - \mathcal{R}^D(S^*) = \frac{1}{A}\int_{\Delta(S,S)} |\gamma - f(x)| dx$. As before, define the empirical risk to be

$$\widehat{\mathcal{R}}_n^D(S) \equiv \frac{1}{n}\sum_{i=1}^n \widehat{e}_S^D(X_i). \tag{18}$$

(As in the regression level set estimation framework, the $1/2$ term is used for normalization in the proceeding analysis.) This is precisely the excess risk we explored in the context of regression level set estimation (4). Note that here, just as in regression level set estimation, the empirical loss function $\widehat{e}_S^D$ is bounded, and this will play a critical role in our analysis later in this paper. For simplicity of presentation, we will focus

on regression level set estimation for most of this paper, but note that the key assumption underlying our analysis (that the observations and hence empirical loss functions are bounded) also holds in the context of density level set estimation. This means that the theoretical properties of the proposed regression level set estimator also hold (up to a normalizing constant) for a density level set estimator based on minimizing the risk in (17).

The problem of density level set estimation has been explored in other contexts, notably [15] and [24]. In [24], Scott and Nowak use dyadic trees and build upon the bounding techniques described in this paper to solve the closely related problem of minimum volume set estimation; *i.e.* finding the set S with minimum volume subject to the constraint that $\int_S f(x)dx = \gamma$. Tsybakov [15] addresses the density level set estimation problem described in this subsection. In his formulation, estimators attempt to maximize an *excess mass* metric, which is defined to be $M(S) \equiv \int_S (f(x) - \gamma)\, dx$. Clearly this is closely related to the error metric proposed in this paper, since $\mathcal{R}^D(S) = \left(M(\bar{S}) - M(S)\right)/(2A)$. Tsybakov's estimator exhibits rates of error convergence for estimation of density level sets similar to the ones derived in this paper, and in fact can nearly optimally estimate much smoother boundaries of level sets than those considered here by fitting polynomials to the boundary. However, his problem is formulated in the plane ($d = 2$) and his approach does not exhibit the spatial adaptivity accompanying the proposed tree-based method.

In the context of density level set estimation, regularization terms can be developed using the same methodology as described above. To this end, set

$$\begin{aligned}
\widehat{e}_L^D(X_i) &\equiv \frac{\gamma\lambda(L) - \mathbb{I}_{\{X_i \in L\}}}{2A}\left[\mathbb{I}_{\{\ell(L)=1\}} - \mathbb{I}_{\{\ell(L)=0\}}\right] \\
\mathcal{R}^D(L) &\equiv \mathbb{E}\left[\widehat{e}_L^D\right] = \int_L \frac{\gamma - f(x)}{2A}\left[\mathbb{I}_{\{\ell(L)=1\}} - \mathbb{I}_{\{\ell(L)=0\}}\right] dx \\
\widehat{\mathcal{R}}_n^D(L) &\equiv \frac{1}{n}\sum_{i=1}^n \widehat{e}_L^D(X_i) = \frac{\gamma\lambda(L) - \widehat{p}_L}{2A}\left[\mathbb{I}_{\{\ell(L)=1\}} - \mathbb{I}_{\{\ell(L)=0\}}\right].
\end{aligned}$$

Note that these definitions correspond with the definitions (16), (18) and (17). In particular, $\mathcal{R}^D(T) = \sum_{L\in\pi(T)} \mathcal{R}^D(L)$. These definitions can be used in the same manner as before to arrive at the following corollary:

Corollary 11 *Let $\delta \in [0,1]$, and define*

$$\Phi_n^D(T) \equiv \sum_{L\in\pi(T)} \sqrt{\frac{2\left(\log(2/\delta) + [\![L]\!]\log 2\right)}{nA}\left[\widehat{p}'_L(\delta)\mathbb{I}_{\{\ell(L)=0\}} + \gamma\lambda(L)\mathbb{I}_{\{\ell(L)=1\}}\right]}.$$

With probability at least $1 - 2\delta$, for all $T \in \mathcal{T}_M$, $\mathcal{R}^D(L) - \widehat{\mathcal{R}}_n^D(L) < \Phi^D(T)$.

This is proved in Section X. Similar to the case of regression level set estimation, Corollary 11 suggests

that minimizing $\widehat{\mathcal{R}}_n^D(T) + \Phi^D(T)$ will make $\mathcal{R}^D(T)$ small with very high probability, and hence $\widehat{T}_n = \arg\min_{T \in \mathcal{T}_M} \widehat{\mathcal{R}}_n^D(T) + \Phi_n^D(T)$ would be an effective estimator in this setting.

VI. Computation and Cycle Spinning

In this section we explore computational techniques for computing the estimator in (9) in two different cases: the general case ($d \geq 2$) and the low-dimensional case ($d = 2$ or $d = 3$). We do this because two- or three-dimensional problems occur frequently in many applications, and when the dimensionality of the data is low it is possible to perform a technique we call "exhaustive voting over shifts" in a computationally tractable manner. This technique will be detailed later in this section.

Fast methods for binary classification in arbitrary dimensions using dyadic decision trees are described in [25], [16] and can easily be extended to the level set estimation problem. Specifically, let $L = \log_2 M$ be the maximum number of dyadic refinements along any coordinate used to form a tree T. Then $\widehat{T}_n$ can be computed in $O(ndL^d \log(nL^d))$ operations using a dynamic programming algorithm.

Recall that $\mathcal{L}_M$ is the collection of all dyadic hyperrectangles L such that $L \in \pi(T)$ for some $T \in \mathcal{T}_M$. A label is assigned to a leaf $L \in \pi(T)$ by choosing $\ell(L) \in \{0, 1\}$ to minimize $\widehat{\mathcal{R}}_n(L)$. Specifically,

$$\ell(L) = \begin{cases} 1, & \sum_{i:X_i \in L} (\gamma - Y_i) \leq 0 \\ 0, & \text{otherwise.} \end{cases}$$

In the case of estimating K level sets simultaneously, $\ell(L) \in \{0, 1, \ldots, K\}$ is assigned as

$$\ell(L) = \max \left\{ k \in \{0, 1, \ldots, K\} \,\middle|\, \sum_{i:X_i \in L} (\gamma_k - Y_i) \leq 0 \right\},$$

where $\gamma_0 \equiv -A$, which minimizes $\widehat{\mathcal{R}}_n^K(L)$.

For some $L \in \mathcal{L}_M$, let T_L denote a subtree rooted at L, and let T_L^* be the subtree T_L which minimizes $\widehat{\mathcal{R}}_n(T_L) + \Phi_n(T_L)$, where $\widehat{\mathcal{R}}_n(T_L) = \frac{1}{n} \sum_{i:X_i \in L} \widehat{e}_T(X_i, Y_i)$. Recall that $L^{s,\text{left}}$ and $L^{s,\text{right}}$ denote the hyperrectangles corresponding to splitting L in half in the s coordinate direction, and denote by $\textsc{merge}(L, T_1, T_2)$ the tree rooted at L having T_1 and T_2 as its right and left branches. The estimate in (9) can then be computed by

$$T_L^* = \arg\min \left\{ \widehat{\mathcal{R}}_n(T_L) + \Phi_n(T_L) \;|\; T_L = \{L\} \text{ or } T_L = \textsc{merge}(L, T_{L^{s,1}}^*, T_{L^{s,2}}^*), s = 1, \ldots, d \right\}.$$

That is, the estimate T_L is computed by choosing the best of $d+1$ trees: the tree associated with make L a leaf node and assigning the label as above to minimize the empirical risk, and the d trees associated with merging the two optimal tree estimates for the two halves of L split in the s^{th} coordinate direction. This method is highly analogous to lower-dimensional dynamic programming methods. One of the key observations made

in [25] is that most $L \in \mathcal{L}_M$ will not contain data when d is large; this eliminates the need for computation on many nodes and thus reduces the computational complexity of the method substantially.

To counteract partition boundary artifacts which result from the use of binary trees, it is possible to perform a "voting over shifts" routine similar to the "averaging over shifts" or "cycle spinning" methods commonly used in function estimation. These methods eliminate the arbitrary choice of alignment between RDPs and the function and thus effectively smooth to reduce the partition boundary artifacts. Rather than averaging the results of shifted estimators, we determine whether each of the M initial cells is more often in or out of the level set of interest and make the corresponding set assignment. While observations may or may not be aligned to a grid, we advocate considering shift amounts which are multiples of the smallest possible hyperrectangle sidelength, $1/M$; this means that some cells will appear at multiple shifts which can lead to computational savings. Ideally, one might like to compute an estimate at each of the M potential shifts, which we call *exhaustive* voting over shifts. When d is large, however, M must be large also and the exhaustive technique may be computationally prohibitive. Gains can instead be realized by voting over random shifts.

When d is small (*i.e.* $d = 2$ or $d = 3$), exhaustive voting over shifts can be very computationally efficient. For this low-dimensional problem, we may use quadtrees or 8-ary trees instead of the binary trees emphasized in much of this paper because presumably $n \gg 2^d$. This allows us to use a bottom-up tree-pruning minimization routine in place of the dyadic programming method described above. The key to making exhaustive voting over shifts computationally efficient is the observation that many tree-pruning decisions occur at multiple different shifts; we can exploit this redundancy as described below.

For the case where M^d is the maximum number of leaf nodes allowed in the tree-based estimate, a tree-pruning algorithm requires computing $\widehat{\mathcal{R}}_n(T_L)$ on $O(M^d)$ different dyadic hypercubes. Since there are M^d different shifts to consider, naively voting over shifts can require computation of $\widehat{\mathcal{R}}_n(T_L)$ on as many as $O(M^{2d})$ hypercubes. However, each hypercube corresponds to a node in several different shifted trees, which means there are $M^d \log_2(M^d)$ ***unique*** pruning decisions to be made on $M^d \log_2 M^d$ ***unique*** hypercubes.

Once these pruning decisions have been made, the next task to convert the resulting leaf nodes and their labels to an estimate. In the case where only one level set is being extracted, this can be accomplished by mapping the pruning decisions and labels to Haar wavelet transform coefficients, computing the inverse wavelet transform, and thresholding the result at one-half. The thresholding step is necessary because the inverse wavelet transform essentially averages the labels over shifts; when the average is greater than one-half, the majority vote must be for a label of one. Both of these operations can be done in $O(M^d \log_2 M^d)$ time. The relationship between the pruning decisions and the wavelet coefficients is delicate and does not correspond to traditional hard or soft thresholding schemes. In the case of multiscale penalized likelihood

estimation, wavelet coefficients are scaled depending on their ancestors' pruning decisions. That is, each wavelet coefficient is weighted by the percentage of different shifts in which the corresponding node was not pruned.

It is important to note that this scaling can be done one level at a time for increased efficiency. Specifically, once all $M^d \log_2 M^d$ pruning decisions have been made, the estimate can be calculated with a breadth-first traversal of the tree associated with the wavelet coefficients. The coefficient weight is initialized to one for each node in the tree. Then, at the top of the tree, each coefficient weight is multiplied by a zero or a one, depending on the pruning decision. Next, the 2^d children coefficients have their weight reduced by a $1/2^d$ of the total weight loss of the parent. This way, the grandchildren and other descendants will be appropriately weighted when the breadth-first traversal reaches their scales, and a nested loop for updating these weights is unnecessary.

In the case where multiple level sets are being extracted simultaneously, application of the inverse wavelet transform introduces a small error. In particular, each pruned node is assigned a label in $\{0, 1, \ldots, K\}$; the majority vote at a given location would correspond to the mode of the labels over all shifts, not the thresholded mean label we can calculate using the inverse wavelet transform. This means that the process of converting the leaf nodes and their labels to an estimate is necessarily more computationally demanding under the onus of multiple level sets. That said, averaging labels over shifts is not always a poor approximation for voting over shifts in practice, particularly when the boundaries of the different level sets are far from each other. When the boundary of level set S_k is relatively isolated from the others, the label associated with points near the boundary will be typically be $k-1$ or k at most shifts, and so computing the mean label and thresholding may result in only a small reduction in accuracy in return for a large increase in computational efficiency. We find that this is an effective approach in our simulations. Note that assigning a label to each partition cell in the final estimate ultimately produces an estimate of $g_S^*(x)$, and so the estimated level sets remain nested.

VII. Simulation results

To test the practical effectiveness of the proposed method, we simulated observations of the elevation of New Orleans, where the true elevations were obtained in DTED Level 0 files from the National Geospatial-Intelligence Agency website and are displayed in Figure 3(a). Organizations such as the US Geological Survey are often interested in identifying flood plains, which shift as a result of erosion and natural disasters. The true flood plain (the level set of interest) is displayed in Figure 3(b); note that it encompasses low-lying regions outside the river, distinguishing this problem from an edge detection problem. Our goal is to extract the flood plain from the set of noisy observations displayed in Figure 3(c). The noisy observations were

obtained by adding zero-mean beta-distributed noise with a variance of $3,333$ to the true image. (Note that this implies that the X_i's are deterministic.) As shown in Figure 3(d), simply thresholding the observations to obtain the level set $\widehat{S}_{thresh}$ is highly insufficient in the presence of noise.

In contrast, the application of the proposed method to this data results in an accurate estimate of the level set, $\widehat{T}_n^{\rho}$, as displayed in Figure 3(e). Because the number of dimensions in this example (two) is small relative to the number of observations, we employ a quad-tree structure instead of the binary tree described above. While this would not be feasible for much larger dimensions, in two dimensions it is both easy to implement and subject to all the performance characterizations derived earlier in this paper. For this simulation, we selected the value of $\rho = 0.0124$ by searching over a range of values to minimize the risk (as defined in (3)); the risk was calculated clairvoyantly using the true function f. We present this clairvoyant estimate to objectively highlight the difference between the proposed approach and a "plug-in" approach (which was given the same advantage, as described below). Furthermore, we employed "voting over shifts", a process analogous to averaging over shifts or using an undecimated wavelet transform. Careful thought reveals that voting over shifts can be accomplished in $O(n \log n)$ time, as described in Section VI.

Compare this result with the result of a more indirect approach: namely, performing wavelet denoising and thresholding the denoised image to obtain a level set estimate, $\widehat{S}_{wavelet}$, to produce to image in Figure 3(f). We used undecimated Haar wavelet denoising [26], and hard threshold the wavelet coefficients at a level 3.39 times the noise variance; as above, this value was selected by searching over a range of values to minimize the risk, which was calculated clairvoyantly using the true function f.

We observed the following mean risks over one hundred noise realizations: $\mathcal{R}(\widehat{S}_{thresh}) - \mathcal{R}(S^*) = 0.0944$, $\mathcal{R}(\widehat{S}_{wavelet}) - \mathcal{R}(S^*) = 0.00450$, and $\mathcal{R}(\widehat{T}_n) - \mathcal{R}(S^*) = 0.00377$. Roughly speaking, wavelet denoising is analogous to choosing a partition with a penalty proportional to the size of the tree or partition, as opposed to the spatially adaptive penalty employed in this method. This example demonstrates that, as expected, the spatially adaptive penalty results in a partition which drills down on the location of the boundary; the wavelet-based approach, in contract, appears to oversmooth the boundary. Furthermore, since the level set of interest does not correspond to an edge in the image, we would not expect curvelets or wedgelets to significantly outperform wavelets in this context.

We see similar performance in the case to multiple simultaneous level set extraction, as displayed in Figure 4.

VIII. Conclusions

This paper demonstrates that tree-based partitioning approaches to level set estimation exhibit near minimax optimal performance and can be computed rapidly to produce effective and practical estimates. This

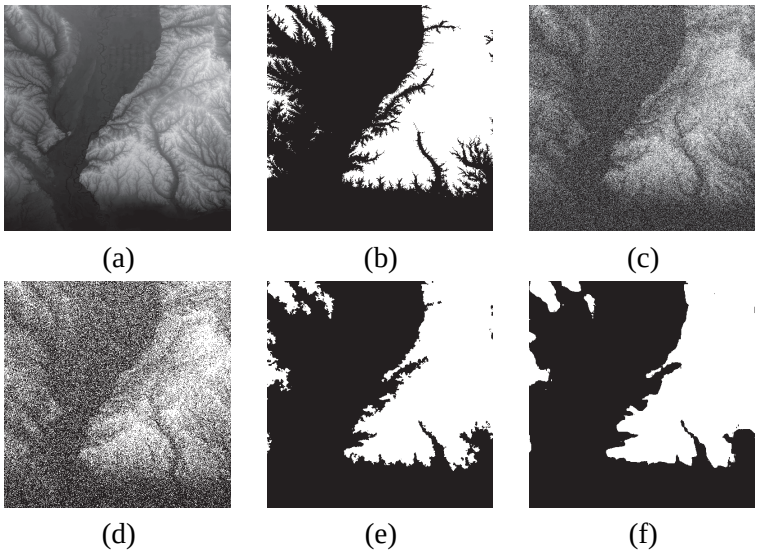


Fig. 3. Simulation results. (a) True function $f : [0,1]^2 \to [-99.5, 99.5]$. DTED values were in the range $[0, 199]$; recentering the elevations to lie in $[-99.5, 99.5]$ does not impact the result. (b) Level set $S^* = \{x \in [0,1]^2 : f(x) > -29.5\}$. (c) Noisy observations, $Y_i \in [-200, 200]$, $i = 1, \ldots, 512^2$. (d) Level set of observations $\widehat{S}_{thresh} = \{X_i : Y_i - 29.5\}$. $\mathcal{R}(\widehat{S}_{thresh}) - \mathcal{R}(S^*) = 0.0944$. (e) Level set estimated with the proposed method. $\mathcal{R}(\widehat{T}_n^\rho) - \mathcal{R}(S^*) = 0.00365$. (f) Level set estimated by TI Haar wavelet denoising followed by thresholding. $\mathcal{R}(\widehat{S}_{wavelet}) - \mathcal{R}(S^*) = 0.00463$.

new estimator is especially promising in image processing applications such as Digital Terrain Elevation Data (DTED) analysis, microarray image analysis, and spatial environmental analysis. Extensions to the simultaneous estimation of multiple level sets and to general regression level set and density level set estimation are also presented. The introduction of a new error metric allows us to bound the weighted symmetric difference between the true level set and the estimate using the relative form of Hoeffding's inequality. This has proven to yield much more accurate results than those achievable with a plug-in method, such as denoising followed by computing the level set from the denoised data. In particular, we demonstrated that, while classical near minimax optimal tree-based image denoising methods penalize the estimator based on the number of leafs in the tree, a substantially different penalization method is necessary for near minimax optimal level set estimation. The new penalty term weighs tree leafs according to their size (with respect to the underlying probability measure), thereby favoring deep, unbalanced trees most likely to accurately represent the true level set of interest.

While the analysis presented in this paper results in an effective tool for level set estimation, there are several areas of ongoing research. One such area is the incorporation of additional information about the noise process (such as noise variance) into the estimation procedure. This may result in an useful estimation technique for a broader spectrum of potential applications. In addition, the proposed method has interesting ties with classical set estimation methods based on snakes or active contours. In particular, the

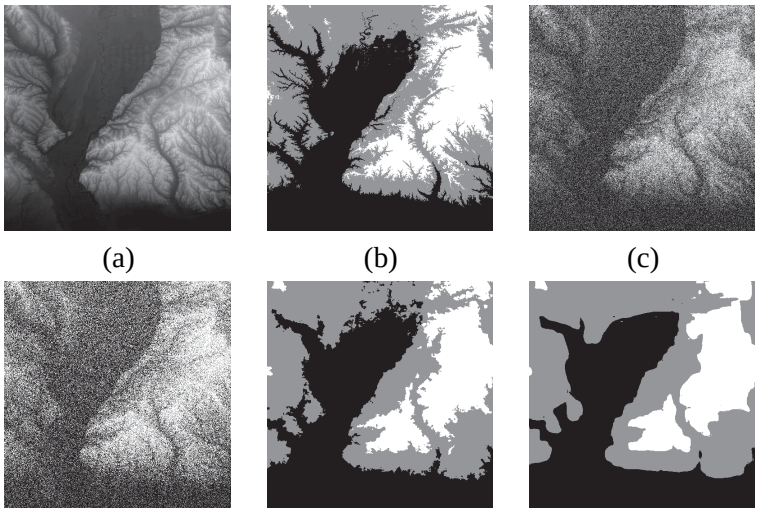


Fig. 4. Simulation results. (a) True function $f : [0,1]^2 \to [-99.5, 99.5]$. (b) Level sets with $\gamma_1 = -59.5$ and $\gamma_2 = 0.5$. (c) Noisy observations, $Y_i \in [-200, 200]$, $i = 1, \ldots, 512^2$. (d) Level sets of observations. $\mathcal{R}(\widehat{S}_{thresh}) - \mathcal{R}(S^*) = 0.0791$. (e) Level sets estimated with the proposed method. $\mathcal{R}(\widehat{T}_n^{\rho}) - \mathcal{R}(S^*) = 0.00325$. (f) Level sets estimated by TI Haar wavelet denoising followed by thresholding. $\mathcal{R}(\widehat{S}_{wavelet}) - \mathcal{R}(S^*) = 0.00402$.

optimization criterion we develop is similar to criteria employed in active contour methods (in that our risk function can be interpreted as an "image force" and our penalty as a measure of active contour energy due to bending [4]), yet was derived from fundamental probabilistic concentration inequalities and admits a thorough theoretical analysis of performance bounds. Nevertheless, the level set estimation method presented in this paper cannot yet address the scope of problems solved using active contours. Ongoing work includes the further development of this analysis in order to better design and characterize set estimation methods for a wider variety of problems.

IX. Acknowledgements

The authors would like to thank Clayton Scott for his insights into our performance analysis when the penalty term is multiplied by a scalar damping factor.

X. Proofs of Lemmas and Theorems

Proof of Lemma 1 First, recall the *relative* form of Hoeffding's inequality:

Theorem 12 *[21] Let the random variables $U_1, U_2, \ldots, U_n$ be independent, with $0 \leq U_i \leq 1$ for each i. Let $S_n = \sum_i U_i$ and let $\mu = \mathbb{E}^n[S_n]$. Then for any $\epsilon > 0$,*

$$\mathbb{P}^n \left[S_n \leq (1-\epsilon)\mu \right] = \mathbb{P}^n \left[\mu - S_n \geq \epsilon \mu \right] \leq e^{-\epsilon^2 \mu / 2}.$$

Now let $U_i \equiv (\widehat{e}_L(X_i, Y_i) + \mathbb{I}_{\{X_i \in L\}})/2$; note that since $\gamma - Y_i \in [-2A, 2A]$, $U_i \in [0, 1]$, as required by Theorem 12. This means that $S_n = (n\widehat{\mathcal{R}}_n(L) + n\widehat{p}_L)/2$ and $\mu = (n\mathcal{R}(L) + np_L)/2$. From these definitions, we find that the following statements are equivalent:

$$\begin{aligned}
\mathbb{P}^n\left[\mu - S_n \geq \sqrt{2\log(1/\delta_L)\mu}\right] &\leq \delta_L \\
\mathbb{P}^n\left[\frac{n}{2}\left(\mathcal{R}(L) - \widehat{\mathcal{R}}_n(L)\right) + \frac{n}{2}\left(p_L - \widehat{p}_L\right) \geq \sqrt{\log(1/\delta_L)\left(n\mathcal{R}(L) + np_L\right)}\right] &\leq \delta_L \\
\mathbb{P}^n\left[\mathcal{R}(L) - \widehat{\mathcal{R}}_n(L) \geq \sqrt{\frac{4\log(1/\delta_L)\left(\mathcal{R}(L) + p_L\right)}{n}} - \left(p_L - \widehat{p}_L\right)\right] &\leq \delta_L.
\end{aligned}$$

Next, note that

$$\mathcal{R}(L) + p_L = \int_L \left[\frac{\gamma - f(x)}{2A}\left[\mathbb{I}_{\{\ell(L)=1\}} - \mathbb{I}_{\{\ell(L)=0\}}\right] + 1\right] d\mathbb{P}_X \leq 2p_L$$

where the inequality follows from the fact that $-1 \leq \frac{\gamma - f(x)}{2A} \leq 1$ since both γ and $f(x) \in [-A, A]$. This yields

$$\mathbb{P}^n\left[\mathcal{R}(L) - \widehat{\mathcal{R}}_n(L) \geq \sqrt{\frac{8\log(1/\delta_L)p_L}{n}} - \left(p_L - \widehat{p}_L\right)\right] \leq \delta_L.$$

■

Proof of Lemma 2 (This closely follows the proof of Theorem 2 in [16], but since the risk cannot be expressed in terms of probability measures in this context, some technical details are different.) Applying Lemma 1, we have that for a particular L, with probability not exceeding $\delta 2^{-(\llbracket L \rrbracket + 1)}$,

$$\begin{aligned}
\mathcal{R}(L) - \widehat{\mathcal{R}}_n(L) &< \sqrt{\frac{8\left(\left(\llbracket L \rrbracket + 1\right)\log 2 + \log(1/\delta)\right)p_L}{n}} - \left(p_L - \widehat{p}_L\right) \\
&\leq \sqrt{\frac{8\left(\llbracket L \rrbracket \log 2 + \log(2/\delta)\right)p_L}{n}} - \left(p_L - \widehat{p}_L\right).
\end{aligned}$$

We now wish to show that a similar bound holds uniformly for all $T \in \mathcal{T}_M$ and all $L \in \pi(T)$ by applying the union bound. However, we want to avoid summing over redundant leaf-label pairs since this will introduce slack into the bound. Note the for a given $L \in \mathcal{L}_M$, T will have the sole effect of assigning a label to L, indicating whether it is estimated to be in S. Thus we can sum over all $L \in \mathcal{L}_M$ and, for each L, over just two trees which assign different labels to L. Applying the union bound in this manner, and letting E_L denote the event that

$$\mathcal{R}(L) - \widehat{\mathcal{R}}_n(L) < \sqrt{\frac{8\left(\llbracket L \rrbracket \log 2 + \log(2/\delta)\right)p_L}{n}} - \left(p_L - \widehat{p}_L\right),$$

we have

$$\mathbb{P}^n\left[\bigcup_{T\in\mathcal{T}_M}\bigcup_{L\in\pi(T)} E_L\right] \leq \sum_{\substack{L\in\pi(T)\\ \text{label = 0 or 1}}} \delta 2^{-([\![L]\!]+1)} \leq \sum_{L\in\mathcal{L}_M} \delta 2^{-[\![L]\!]} \leq \delta$$

where the last inequality follows from the Kraft inequality, which is applicable since $[\![L]\!]$ is a prefix codelength.

Thus, with probability at least $1-\delta$, $\mathcal{R}(T) - \widehat{\mathcal{R}}_n(T) \leq \Phi'_n(T) - \sum_{L\in T}(p_L - \widehat{p}_L)$ for all $T \in \mathcal{T}_M$. The statement of the lemma follows from the fact that $\sum_{L\in T} p_L = \sum_{L\in T}\widehat{p}_L = 1$.

■

Proof of Theorem 8 First note that the collection of potential set estimates $\mathcal{T}_M$ is closed with respect to complimentation. Furthermore, $\mathcal{R}(T) = -\mathcal{R}(T^c)$, $\widehat{\mathcal{R}}_n(T) = -\widehat{\mathcal{R}}_n(T^c)$, and $\Phi_n(T) = \Phi_n(T^c)$, which means

$$\widehat{\mathcal{R}}_n(T) - \mathcal{R}(T) = \mathcal{R}(T^c) - \widehat{\mathcal{R}}_n(T^c) \leq \Phi_n(T^c) = \Phi_n(T)$$

and so $|\mathcal{R}(T) - \widehat{\mathcal{R}}_n(T)| \leq \Phi_n(T)$. Now, assume the n observations $\{(X_i, Y_i)\}_{i=1}^n$ are such that $\mathcal{R}(T) \leq \widehat{\mathcal{R}}_n(T) + \Phi_n(T)$ holds for all $T \in \mathcal{T}_M$; Theorem 4 states that the observations meet this criterion with probability at least $1-2\delta$. We then have

$$\begin{aligned}
\mathcal{R}(\widehat{T}_n^\rho) &\leq \widehat{\mathcal{R}}_n(\widehat{T}_n^\rho) + \Phi_n(\widehat{T}_n^\rho) \leq \max\left(\frac{1}{\rho}, 1\right)\left(\widehat{\mathcal{R}}_n(\widehat{T}_n^\rho) + \rho\Phi(\widehat{T}_n^\rho)\right) \\
&= \max\left(\frac{1}{\rho}, 1\right)\min_{T\in\mathcal{T}_M}\left\{\widehat{\mathcal{R}}_n(T) + \rho\Phi(T)\right\} \leq \max\left(\frac{1}{\rho}, 1\right)\min_{T\in\mathcal{T}_M}\left\{\mathcal{R}(T) + (1+\rho)\Phi(T)\right\}.
\end{aligned}$$

Subtracting $\mathcal{R}(S^*)$ from both sides, we arrive at the statement of the theorem. ■

Proof of Corollary 10 Since each $\gamma_k \in [-A, A]$, we can define U_i as before to be $\frac{\widehat{e}_L^K(X_i,Y_i)}{2} + \frac{\mathbb{I}_{\{X_i\in L\}}}{2}$ and will have $U_i \in [0,1]$. This allows us to apply the relative form of Hoeffding's inequality as before to derive the penalty. The $\log(K+1)$ term is used because, while before only two labels were considered during the union bound in the proof of Lemma 2, we must now consider $K+1$ different possible labels. ■

Proof of Corollary 11 Let $U_i \equiv \widehat{e}_L^D(X_i) + \mathbb{I}_{\{X_i\in L\}}/(2A) + \gamma\lambda(L)/(2A)$; note that since $\gamma \in [0, A]$, $A \geq 1$, and $\lambda(L) \in [0,1]$, $U_i \in [0,1]$, as required by Theorem 12. This means that $S_n = n\widehat{\mathcal{R}}_n^D(L) + n\widehat{p}_L/(2A) + n\gamma\lambda(L)/(2A)$ and $\mu = n\mathcal{R}^D(L) + np_L/(2A) + n\gamma\lambda(L)/(2A)$. From these definitions, we find

that the following statements are equivalent for some $\delta_L \in [0,1]$:

$$\begin{aligned}
\mathbb{P}^n\left[\mu - S_n \geq \sqrt{2\log(1/\delta_L)\mu}\right] &\leq \delta_L \\
\mathbb{P}^n\left[n\left(\mathcal{R}^D(L) - \widehat{\mathcal{R}}_n^D(L)\right) + \frac{n}{2A}\left(p_L - \widehat{p}_L\right) \geq \sqrt{2n\log(1/\delta_L)\left(\mathcal{R}^D(L) + (p_L + \gamma\lambda(L))/(2A)\right)}\right] &\leq \delta_L \\
\mathbb{P}^n\left[\mathcal{R}^D(L) - \widehat{\mathcal{R}}_n^D(L) \geq \sqrt{\frac{2\log(1/\delta_L)}{n}(\mathcal{R}^D(L) + (p_L + \gamma\lambda(L))/(2A))} - \frac{p_L - \widehat{p}_L}{2A}\right] &\leq \delta_L.
\end{aligned}$$

Next note that

$$\mathcal{R}^D(L) + \frac{\gamma\lambda(L)}{2A} + \frac{p_L}{2A} = \begin{cases} \gamma\lambda(L)/A, & \ell(L) = 1 \\ p_L/A, & \ell(L) = 0 \end{cases}.$$

Following the proof of Lemma 2 and applying Lemma 3, we arrive at the desired bound. ■

REFERENCES

[1] A. Sole, V. Caselles, G. Sapiro, and F. Arandiga, "Morse description and geometric encoding of digital elevation maps," *IEEE Trans. on Im. Proc.*, vol. 13, no. 9, pp. 1245–1262, 2004.

[2] 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.

[3] 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.

[4] M. Kass, A. Witkin, and D. Terzopoulos, "Snakes: Active contour models," in *Proceedings, First International Conference on Computer Vision*, London, England, 1987, pp. 259–268.

[5] A. Blake and M. Isard, *Active Contours*. Springer-Verlag, 1998.

[6] X. Han, C. Xu, and J. L. Prince, "A topology preserving level set method for gemetric deformable models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, no. 6, pp. 755–768, 2003.

[7] R. Willett and R. Nowak, "Platelets: a multiscale approach for recovering edges and surfaces in photon-limited medical imaging," *IEEE Transactions on Medical Imaging*, vol. 22, no. 3, pp. 332–350, 2003.

[8] R. Castro, R. Willett, and R. Nowak, "Coarse-to-fine manifold learning," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing — ICASSP '04*, 17-21 May, Montreal, CA, 2004.

[9] E. Candès and D. Donoho, "Curvelets: A surprisingly effective nonadaptive representation for objects with edges," To appear in Curves and Surfaces, L. L. Schumaker et al. (eds), Vanderbilt University Press, Nashville, TN.

[10] A. P. Korostelev and A. B. Tsybakov, *Minimax theory of image reconstruction*. New York: Springer-Verlag, 1993.

[11] Y. Yang, "Minimax nonparametric classification-part i: Rates of convergence," *IEEE. Trans. Inform. Theory*, vol. 45, no. 7, pp. 2271–2284, 1979.

[12] J. S. Marron, "Optimal rates of convergence to bayes risk in nonparametric discrimination," *Annals of Statistics*, vol. 11, no. 4, pp. 1142–1155, 1983.

[13] E. Mammem and A. Tsybakov, "Asymptotical minimax recovery of sets with smooth boundaries," *Annals of Statistics*, vol. 23, no. 2, pp. 502–524, 1995.

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

[15] A. Tsybakov, "On nonparametric estimation of density level sets," *Annals of Statistics*, vol. 25, no. 3, pp. 948–969, 1997.

[16] C. Scott and R. Nowak, "Minimax-optimal classification with dyadic decision trees," to appear in *IEEE Trans. on Info. Th.*, April 2006.

[17] J. Klemela, "Complexity penalized support estimation," *J. Multivariate Anal.*, vol. 88, pp. 247–297, 2004.

[18] C. Scott, "Regression level set estimation reduces to cost-sensitive classification," submitted to *IEEE Trans. Sig. Proc.*, available at http://www.stat.rice.edu/∼cscott/.

[19] R. Willett and R. Nowak, "Minimax optimal level set estimation," Duke University, Tech. Rep., 2006, available at http://www.ee.duke.edu/∼willett/papers/WillettLevelSetTechReport.pdf.

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

[21] C. McDiarmid, "Concentration," in *Probabilistic Methods for Algorithmic Discrete Mathematics*. Berlin: Springer, 1998, pp. 195–248.

[22] E. Kolaczyk and R. Nowak, "Multiscale likelihood analysis and complexity penalized estimation," *Annals of Stat.*, vol. 32, pp. 500–527, 2004.

[23] Personal communication with Clayton Scott.

[24] C. Scott and R. Nowak, "Learning minimum volume sets," accepted for publication in *Journal of Machine Learning Research*, available at http://www.stat.rice.edu/∼cscott/pubs.html.

[25] G. Blanchard, C. Schäfer, and Y. Rozenholc, "Oracle bounds and exact algorithm for dyadic classification trees," in *Proceedings of COLT: The Annual Workshop on Learning Theory*, Banff, Canada, 2004, pp. 378–392.

[26] R. R. Coifman and D. L. Donoho, *Wavelets and Statistics, Lecture Notes in Statistics 103*. Springer-Verlag, 1995, ch. Translation-Invariant De-Noising.