LEVEL SET ESTIMATION VIA TREES

Rebecca Willett and Robert Nowak

Department of Electrical and Computer Engineering
University of Wisconsin, 1415 Engineering Drive, Madison, WI 53706, USA

ABSTRACT

Tree-structured partitions provide a natural framework for rapid and accurate extraction of the level sets of a multivariate function f from noisy data. In general, a level set is the set S on which f exceeds some critical value (e.g., $S = \{x : f(x) \geq \gamma\}$). Boundaries of level sets typically constitute manifolds embedded in the high-dimensional observation space. The identification of these boundaries is an important theoretical problem with applications for digital elevation maps, medical imaging, and pattern recognition. Because level set identification is intrinsically simpler than field denoising or estimation, explicit level set extraction methods can achieve higher accuracy than more indirect approaches (such as extracting a level set from an estimate of the function). The trees underlying our method are constructed by minimizing a complexity regularized data-fitting term over a family of dyadic partitions. Our method automatically adapts to spatially varying regularity of both the level set and the field underlying the data. Level set extraction using multiresolution trees can be implemented in near linear time and specifically aims to minimize an error metric sensitive to both the error in the location of the level set and the associated field estimation error.

1. INTRODUCTION

Set estimation is an important theoretical problem with a large number of important and diverse applications. Set estimation, which is the recovery of regions in the signal's domain in which the signal satisfies some criterion, arises in the extraction of lines of iso-height from digital elevation maps, feature identification, and pattern recognition. In these and other applications, set estimation can be daunting but necessary task for effective data analysis.

The capabilities of set extraction methods are typically governed by the existence of boundaries and edges in the data. Existing work has tried to address this by using specialized basis functions and representations (such as wedgelets [1], curvelets [2], or platelets [3]), but these solve the sometimes tangential problem of denoising objects separated by boundaries, rather than identifying the boundaries themselves. Boundary estimation arises in several data analysis problems where neither field estimation nor traditional classification is appropriate. The boundaries of such sets typically constitute lower-dimensional manifolds embedded in a higher-dimensional observation space. In this paper, we address the problem of recovering sets from noisy multi-dimensional data.

Set estimation may be more desirable than complete signal reconstruction for a variety of reasons. In many applications, the location of boundaries or level sets is of principal importance, while the amplitude of the function away from any boundary is secondary, if not irrelevant. For example, doctors may wish to identify regions where uptake of a pharmaceutical exceeds some critical level, or mapping agencies may want to extract networks of roads from satellite images. Signal estimates with very low mean error over the entire image may not be suitably accurate near the level set boundaries. Because set estimation is intrinsically simpler than field estimation, explicit level set extraction methods can potentially achieve higher accuracy than more indirect approaches (such as estimating a set from an estimate of the function).

This distinction manifests itself in the selection of tree-pruning criteria. Tree-based signal estimation frameworks (e.g. [1, 3, 4]) determine whether nodes should be pruned from the tree based on the size of the tree. Recent work in tree-based methods for binary classification [5], however, has revealed that optimal pruning decisions must exhibit a spatial adaptivity not possible using traditional, tree size based criteria. This difference implies that one critical aspect of tree-based set estimation must be the selection of the tree-pruning rule most appropriate for set estimation.

1.1. Relationship to previous work

There exist a number of straightforward approaches to level set estimation, including thresholding observations, denoising observations and thresholding the result, or performing a classification routine on thresholded observations. The difficulty with these approaches, however, is that performance analysis is very difficult, if not intractable, in the presence of noise. Related work was conducted by Mammen and Tsybakov in [6], but this work focused on estimation of a boundary between a black and a white region from binary observations, 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 method proposed in this paper is that it is capable of utilizing additional information available from non-binary observations.

1.2. Notation

Let $f : [0,1]^d \longrightarrow [C_\ell, C_u]$ be a field of bounded amplitude, and let

$$S \equiv \{x \in [0,1]^d : f(x) \geq \gamma\}$$

for some $\gamma \in [C_\ell, C_u]$. Given n noisy observations $(x_i, y_i) \in [0,1]^d \times [C_\ell, C_u]$, $i = 1, \ldots, n$, where $\mathbb{E}[y_i] = f(x_i)$, our goal is to learn an estimate of S. Let $\mathbb{P}_X$ be the probability measure for X, which determines the marginal distribution of observations on the support of f, and let $p_A \equiv \int_A d\mathbb{P}_X$ for $A \subseteq [0,1]^d$. Given two sets, S and F, let

$$\Delta(S, F) \equiv \{x : x \in (S \cap F^c) \cup (F \cap S^c)\}$$

denote the symmetric difference, where S^c denotes the complement of S.

Supported by the National Science Foundation, grants CCR-0310889 and ANI-0099148, and the Office of Naval Research, grant N00014-00-1-0390

2. ERROR METRICS FOR SET ESTIMATION

Careful selection of an error metric is the first step in designing a level set estimator. The goal of signal estimation is typically to minimize the mean squared error between the true signal and the estimate, and the goal of binary classification is usually to minimize the probability of misclassification, which leads to minimizing the symmetric difference between the decision sets of the Bayes' and the learned classifiers. In level set estimation, however, it is more appropriate to minimize the symmetric difference between the true set of interest and its estimate weighted by severity of the error over the symmetric difference.

An appropriate error metric can be designed as follows. For a given set F, we define the local error function to be

$$e_F(x) = \frac{\gamma - f(x)}{C_u - C_\ell} \left[\mathbb{I}_{\{x \in F\}} - \mathbb{I}_{\{x \in F^c\}} \right] - \frac{C_\ell}{C_u - C_\ell}$$

where $\mathbb{I}$ is the indicator function. The normalization ensures that $e_F \in [0, 1]$. From here, we define risk function as

$$\mathcal{R}(F) = \int e_F(x) d\mathbb{P}_X;$$

this measures the distance between the signal, f, and the threshold, γ, and weights the distance at each location x by plus or minus one according to whether $x \in F$. Thus regions where $x \in F$ but $f(x) < \gamma$ (that is, $x \in S^c$) will contribute positively to the risk function. Given the risk function, we can define the "excess risk" as $\mathcal{R}(F) - \mathcal{R}(S)$, which measures the difference between the risk of an estimate and the risk of the true level set, S. Using the definition of the risk, the excess risk can be written as

$$\mathcal{R}(F) - \mathcal{R}(S) = \frac{2}{C_u - C_\ell} \int_{\Delta(S,F)} |\gamma - f(x)| d\mathbb{P}_X, \quad (1)$$

which gives a weighted measure of the symmetric difference between S and F, as desired. Note that minimizing (1) is equivalent to minimizing $\mathcal{R}(F)$ since $\mathcal{R}(S)$ is a constant.

The effect of such a metric is demonstrated in Figure 1. On the left is drawn a contour outlining the true level set S. The center and rightmost figures show the boundary of two different candidate level set estimates. 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.

An additional advantage of the proposed metric is that it is simple to define an empirical error metric for a candidate level set estimate F as

$$\widehat{e}_F(x_i) = \frac{\gamma - y_i}{C_u - C_\ell} \left[\mathbb{I}_{\{x_i \in F\}} - \mathbb{I}_{\{x_i \in F^c\}} \right] - \frac{C_\ell}{C_u - C_\ell}$$

resulting in the empirical risk function

$$\widehat{\mathcal{R}}_n(F) = \frac{1}{n} \sum_{i=1}^{n} \widehat{e}_F(x_i),$$

which is both computable and constructed so that $\mathbb{E}[\widehat{\mathcal{R}}_n(F) - \mathcal{R}(S)] = \mathcal{R}(F) - \mathcal{R}(S)$.

This metric is distinctly different from the L_p norms typically encountered in signal estimation or the "unweighted" symmetric difference metrics arising in classification.

(a) (b) (c)

Fig. 1. Behavior of level set error metric. (a) Field f and true 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.

3. LEVEL SET ESTIMATION PROCEDURE

We propose to estimate the level set of a function from noisy observations by using a tree-pruning method akin to CART [4] or dyadic decision trees [5]. Let $\pi(T)$ denote the partition induced on $[0, 1]^d$ by the binary tree T. This can be used to represent an estimate of a level set by assigning a zero or a one to each leaf (i.e. each $A \in \pi(T)$) to indicate whether that cell of the partition is in $\widehat{S}$.

Let $\widehat{p}_A = (1/n) \sum_i \mathbb{I}_{\{x_i \in A\}}$ be the empirical estimate of p_A, and

$$\widehat{p}'_A(\delta) \equiv 4 \max \left(\widehat{p}_A, \frac{\log(1/\delta) + q_A \log 2}{n} \right).$$

Let q_A denote the number of bits required to encode A. Specifically, consider the prefix code proposed in [5] for $A \in \pi(T)$. If A is at level j in the binary tree T, then $j + 1$ bits must be used to describe the depth of A, j bits must be used to describe whether each branch is a left or right branch, $j \log_2 d$ bits must be used to describe the coordinate direction of each of the j branches, and one bit must be used to assign a label (inside or outside of $\widehat{S}$) to A. This results in a total of $2j + j \log_2 d + 2$ bits, and this expression is denoted as q_A.

Next set the penalty associated with a tree-based estimator to be

$$\Phi_n(T) = \sum_{A \in \pi(T)} \sqrt{\frac{2\widehat{p}'_A(\delta)}{n} (\log(2/\delta) + q_A \log 2)}; \quad (2)$$

the next section contains a detailed discussion of the origin of this penalty. Intuitively, this penalty is designed to favor unbalanced trees which hone in on the location of the manifold defining the level set. To see this, note that $q_A \asymp j$, while $\widehat{p}'_A(\delta) \asymp 2^{-j}$. 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, as displayed in Figure 2.

Define the level set estimator to be

$$\widehat{T}_n \equiv \arg \min_{T \in \mathcal{T}_M} \widehat{\mathcal{R}}_n(T) + \Phi_n(T), \quad (3)$$

where $\mathcal{T}_M$ is the set of all dyadic trees which partition $[0, 1]^d$ into rectangular cells with sidelengths no longer than $1/M$. As shown

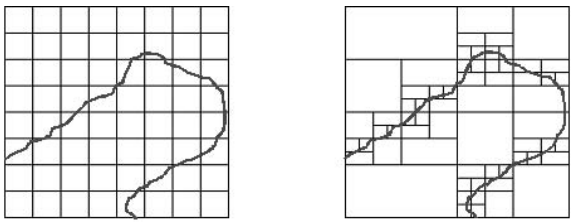

Fig. 2. Balanced and unbalanced partitions of the function support. Both partitions have the same number of leaves, but the partition on the right is better adapted to the boundary and has a smaller $\Phi_n(T)$.

in the following section, this estimator is nearly minimax optimal. Furthermore, the estimator is rapidly computable, as described in [5, 7]. 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.

4. PERFORMANCE ANALYSIS

Many of the key points in the following theoretical analysis were derived using the error bounding techniques developed by Scott and Nowak [5] in the context of binary classification. The proposed level set estimation method hinges on the following main result:

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

$$\mathcal{R}(T) \leq \widehat{\mathcal{R}}_n(T) + \Phi_n(T) \tag{4}$$

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

A sketch proof of this result follows: For a partition cell $A \in \pi(T)$, let $\mathcal{R}(T, A) = \int_A e_T(x) d\mathbb{P}_X$ and $\widehat{\mathcal{R}}_n(T, A) = (1/n) \sum_i \widehat{e}_T(x_i) \mathbb{I}_{\{x_i \in A\}}$. Applying the "relative" form of Hoeffding's inequality (Theorem 2.3(c) in [8]), we have that, with probability at least $1 - \delta$,

$$\mathcal{R}(T, A) - \widehat{\mathcal{R}}_n(T, A) < \sqrt{\frac{2\mathcal{R}(T, A) \log(1/\delta)}{n}}. \tag{5}$$

Next note that (5) implies

$$\mathcal{R}(T, A) - \widehat{\mathcal{R}}_n(T, A) < \sqrt{\frac{2\mathcal{R}(T, A) \left((q_A + 1) \log 2 + \log(1/\delta)\right)}{n}}.$$

with probability not exceeding $\delta 2^{-(q_A+1)}$. Summing over all $A \in \pi(T)$, we have

$$\mathcal{R}(T) - \widehat{\mathcal{R}}_n(T) < \sum_{A \in \pi(T)} \sqrt{\frac{2\mathcal{R}(T, A) \left((q_A + 1) \log 2 + \log(1/\delta)\right)}{n}} \tag{6}$$

except on a set of probability not exceeding

$$\sum_{\substack{A \in \pi(T) \\ \text{label = 0 or 1}}} \delta 2^{-(q_A+1)} = \sum_{A \in \pi(T)} \delta 2^{-q_A} \leq \delta,$$

where the last inequality follows from the Kraft inequality, which is applicable since q_A is a prefix codelength.

While the bound in (6) is valid, it is not computable because it depends on $\mathcal{R}(T, A)$. However, note that $\mathcal{R}(T, A) \leq p_A$ for all A. As shown in [5], $p_A \leq \widehat{p}'_A$ with probability at least $1 - \delta$, leading to the result in Theorem 1.

Not only does this framework give us a principled way to choose a good level set estimator (i.e. the estimator in (3), but it also allows us to bound the expected risk for a collection of n observations. In particular, we have the following theorem:

Theorem 2 *Let $\widehat{T}_n$ be as in (3) with $\Phi_n(T)$ as in (2). With probability at least $1 - 4/n$ over the training sample,*

$$\mathcal{R}(\widehat{T}_n) - \mathcal{R}(S) \leq \min_{T \in \mathcal{T}_M} \left(\mathcal{R}(T) - \mathcal{R}(S) + \Phi_n(T)\right) + \sqrt{\frac{\log n}{2n}}.$$

As a consequence,

$$\mathbb{E}\left[\mathcal{R}(\widehat{T}_n) - \mathcal{R}(S)\right] \leq \min_{T \in \mathcal{T}_M} \left(\mathcal{R}(T) - \mathcal{R}(S) + \Phi(T)\right) + \sqrt{\frac{\log n}{2n}} + \frac{4}{n}.$$

The proof closely follows that of Theorem 7 in [5] and is omitted here for brevity. This theorem decomposes the expected error into two main components: (a) $\mathcal{R}(T) - \mathcal{R}(S)$, the error associated with approximating S with a tree-based partition, and (b) $\Phi_n(T)$, which can be viewed as a bound on the estimation error.

The bound on the expected error in Theorem 2 allows us to analyze the proposed method in terms of rates of error convergence. Assume that the boundary of S has box-counting dimension one; i.e., if $[0, 1]^d$ is partitioned into M^d equal sized cells, then the boundary of S intersects no more than CM^{d-1} of those cells for all M and some constant C. Then $\mathcal{R}(T) - \mathcal{R}(S) \leq C_1/M$ and $\Phi(T) \leq C_2 M^{d-1}(\log n)/n$ for some constants C_1 and C_2. Optimizing over M, we find that

Theorem 3 *If the boundary of S has box-counting dimension one, then*

$$\mathbb{E}\left[\mathcal{R}(\widehat{T}_n) - \mathcal{R}(S)\right] \leq C_3 (n/\log n)^{-1/d}$$

for some constant C_3.

Because level set estimation can be viewed as a generalization of the binary classification problem, we have that the minimax lower bound on the error convergence rate for this problem is $n^{-1/d}$, which implies that the proposed method performs within a logarithmic factor of the optimal rate.

5. SIMULATION RESULTS

To test the practical effectiveness of the proposed method, we simulated observations of the elevation of St. Louis, where the true elevations, normalized to lie between zero and 255 were obtained from the U.S. Geological Survey website and are displayed in Figure 3(a). Organizations such as the USGS are often interested in identifying flood plains, which shift as a result of plate tectonics. 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 uniform noise with a variance of three thousand to the true image. (Note that this implies that the x_i's are deterministic, not random as assumed in the theoretical analysis. The extension of the above analysis to deterministic sample locations is part of our ongoing research.) 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, $\widehat{T}_n$, as displayed in Figure 3(e). This estimate was formed by weighting the penalty $\Phi_n(T)$ to minimize the average empirical error and objectively highlight the difference between the proposed approach and a wavelet-based approach (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. 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, and set the hard threshold to maximize the level set estimation accuracy. After empirically selecting an appropriate weight on $\Phi_n(T)$ and wavelet threshold, we observed the following mean errors over one hundred noise realizations:

$$\begin{aligned} \mathcal{R}(\widehat{S}_{thresh}) - \mathcal{R}(S) &= 8,595 \\ \mathcal{R}(\widehat{S}_{wavelet}) - \mathcal{R}(S) &= 1,469 \\ \mathcal{R}(\widehat{T}_n) - \mathcal{R}(S) &= 1,101. \end{aligned}$$

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.

6. CONCLUSIONS AND FUTURE WORK

We have demonstrated that tree-pruning based approaches to level set estimation result in nearly optimal estimates and can be computed rapidly to produce effective practical estimates. 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. The extension of this method to the estimation of a collection of level sets simultaneously is an area of ongoing investigation.

Because of the variety of applications for which level set extraction may be useful, understanding the optimal tree pruning strategy and associated performance bounds is important for a variety of observation models, including Gaussian and Poisson noise models. We plan to extend the above framework to these cases in future work. We also plan to more thoroughly characterize how these performance bounds compare with two alternative, more implicit, level set extraction methods: (a) signal estimation followed by set extraction from the signal estimate, and (b) thresholding the observations according to whether they meet the given set criterion followed by binary classification.

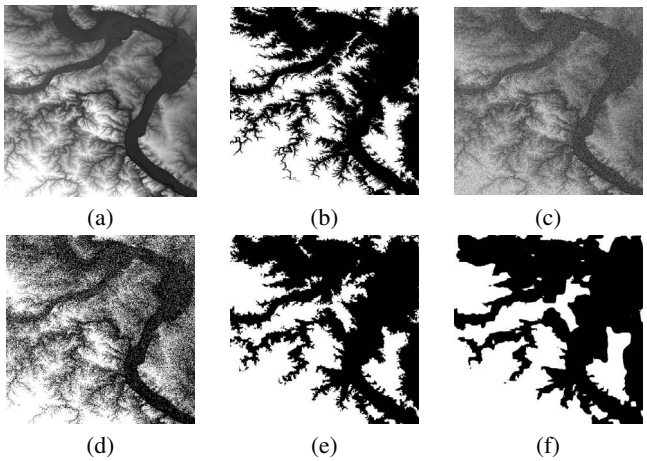

Fig. 3. Simulation results. (a) True function $f : [0,1]^2 \rightarrow [0,255]$. (b) True level set $S = \{x \in [0,1]^2 : f(x) > 120\}$. (c) Noisy observations, $y_i \in [-95, 350]$, $i = 1, \ldots, 512^2$. (d) Level set of observations $\widehat{S}_{thresh} = \{x_i : y_i > 120\}$. $\mathcal{R}(\widehat{S}_{thresh}) - \mathcal{R}(S) = 8,598$. (e) Level set estimated with the proposed method. $\mathcal{R}(\widehat{T}_n) - \mathcal{R}(S) = 1,104$. (f) Level set estimated by TI Haar wavelet denoising followed by thresholding. $\mathcal{R}(\widehat{S}_{wavelet}) - \mathcal{R}(S) = 1,472$.

7. REFERENCES

[1] D. Donoho, "Wedgelets: Nearly minimax estimation of edges," *Ann. Statist.*, vol. 27, pp. 859 – 897, 1999.

[2] 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.

[3] 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, 2003.

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

[5] C. Scott, *Dyadic Decision Trees*, Ph.D. thesis, Rice University, 2004.

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

[7] 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.

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