SIAM J. IMAGING SCIENCES
Vol. 6, No. 4, pp. 2047–2074
© 2013 Society for Industrial and Applied Mathematics

Level Set Estimation from Projection Measurements: Performance Guarantees and Fast Computation*

Kalyani Krishnamurthy†, Waheed U. Bajwa‡, and Rebecca Willett†

Abstract. Estimation of the level set of a function (i.e., regions where the function exceeds some value) is an important problem with applications in digital elevation mapping, medical imaging, astronomy, etc. In many applications, the function of interest is not observed directly. Rather, it is acquired through (linear) projection measurements, such as tomographic projections, interferometric measurements, coded-aperture measurements, and random projections associated with compressed sensing. This paper describes a new methodology for rapid and accurate estimation of the level set from such projection measurements. The key defining characteristic of the proposed method, called the projective level set estimator, is its ability to estimate the level set from projection measurements without an intermediate reconstruction step. This leads to significantly faster computation relative to heuristic "plug-in" methods that first estimate the function, typically with an iterative algorithm, and then threshold the result. The paper also includes a rigorous theoretical analysis of the proposed method, which utilizes results from the literature on concentration of measure and characterizes the estimator's performance in terms of geometry of the measurement operator and ℓ_1-norm of the discretized function.

Key words. level set estimation, projection measurements, compressive sensing, performance guarantees

AMS subject classifications. 15A29, 15A60, 68Q17, 94A08, 97M50

DOI. 10.1137/120891927

1. Introduction.

Level set estimation is the process of using indirect observations of a function f defined on the unit hypercube $[0,1]^d$ to estimate the region(s) where f exceeds some critical value γ; i.e., $S^* \triangleq \{x \in [0,1]^d : f(x) > \gamma\}$. Accurate and efficient level set estimation plays a crucial role in a variety of scientific and engineering tasks, including the localization of "hot spots" signifying tumors in medical imaging [30, 18], significant photon sources in astronomy [24], and strong reflectors in remote sensing [2, 34].

In this paper, we consider making observations of the form $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{f} + \boldsymbol{n}$, where $\boldsymbol{f}$ is a discretized version of f, $\boldsymbol{A}$ is a (discrete) linear operator that may not be invertible, and $\boldsymbol{n}$ is additive noise that corrupts our observations. For instance, $\boldsymbol{y}$ might correspond to tomographic projections in tomography [21, 29, 23], interferometric measurements in radar interferometry [38], multiple blurred, low-resolution, dithered snapshots in astronomy [36], or random projections in compressed sensing systems [1, 8, 9, 12, 45]. Our goal in this $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{f} + \boldsymbol{n}$

*Received by the editors September 19, 2012; accepted for publication (in revised form) July 10, 2013; published electronically October 22, 2013. Part of this work appeared previously as [26].
http://www.siam.org/journals/siims/6-4/89192.html
†Department of Electrical and Computer Engineering, Duke University, Durham, NC 27708 (kk63@duke.edu, willett@duke.edu). The work of these authors was supported by NGA Award HM1582-10-1-0002 SUB #1-P3130108 and by AFRL grant FA8650-07-D-1221.
‡Department of Electrical and Computer Engineering, Rutgers University, Piscataway, NJ 08854 (waheed.bajwa@rutgers.edu). This author's work was supported in part by the NSF under grant CCF-1218942.

setting is to perform level set estimation of the continuous-domain function f *without* an intermediate step involving time-consuming reconstruction of $\boldsymbol{f}$. There are two reasons for this. First, level set estimation without reconstruction of $\boldsymbol{f}$ would allow design of sequential measurement schemes optimally adapted to the function of interest. For instance, in tomography we would like to estimate the level set, S^*, quickly from an initial set of observations so that additional observations focused on S^* can be collected immediately, resulting in an overall low radiation dose [27, 33, 32]. Some recent works [20, 19] have provided theoretical characterizations of the significant benefits associated with certain sequential measurement schemes; the method proposed in this paper may facilitate the use of such schemes in time-sensitive or computational-resource-limited applications. Second, "plug-in" approaches that estimate $\boldsymbol{f}$ and threshold the estimate $\widehat{\boldsymbol{f}}$ to extract S^* are notoriously difficult to characterize; their performances hinge upon the statistics of the estimation error $\widehat{\boldsymbol{f}} - \boldsymbol{f}$, which for most reconstruction methods are unknown (with the possible exception of the first moment). More generally, reconstruction methods 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 boundary. Finally, the Vapnik principle [50] states that one should never solve a complex problem as an intermediate step towards solving a simple problem.

1.1. Problem formulation. In this work, we observe samples of a function f supported on $[0,1]^d$ of the form

$$\boldsymbol{y} = \boldsymbol{A}\boldsymbol{f} + \boldsymbol{n} \in \mathbb{R}^K, \tag{1.1}$$

where

- $\boldsymbol{A} \in \mathbb{R}^{K \times N}$ is a linear operator that is assumed to be known with K often less than N,
- $\boldsymbol{f} \in \mathbb{R}^N$ corresponds to integration samples of f; i.e.,

$$f_i = \frac{1}{\operatorname{vol}(C_i)} \int_{C_i} f(x)dx \tag{1.2}$$

  for $i = 0, 1, \ldots, N-1$, where the cells C_i are obtained by partitioning $[0,1]^d$ into nonoverlapping hypercubes such that each C_i has sidelength $N^{-1/d}$ and volume $1/N$, and
- $\boldsymbol{n} \in \mathbb{R}^K$ denotes the additive measurement noise, which is assumed to be zero-mean, sub-Gaussian white noise in our case; i.e., $n_i \overset{\text{i.i.d.}}{\sim} \operatorname{Sub}(c_s)$ is a zero-mean, sub-Gaussian random variable, defined by the condition $\left(\mathbb{E}\left[|n_i|\right]^p\right)^{1/p} \leq c_s\sqrt{p}$ for $p \geq 1$.[1]

We assume without loss of generality that the columns of $\boldsymbol{A}$ have unit ℓ_2-norms, and consider N to be dyadic (power of two). A γ-level set in this discrete setting can be written as $S_N^* = \{i : f_i > \gamma\}$, where the subscript N signifies that the discrete-domain level set is a

[1]Note that the sub-Gaussian noise assumption subsumes the usual assumption of Gaussian noise; in particular, Gaussian random variables and bounded random variables fall under the category of sub-Gaussian random variables [39].

function of the N-dimensional discrete signal $\boldsymbol{f}$.[2] Throughout this paper, the dependencies of the continuous-domain level set S^* and the discrete-domain level set S_N^* on γ are implicit.

Our main goal is to estimate the continuous-domain level set S^* from discrete measurements $\boldsymbol{y}$ *without reconstructing* the underlying signal $\boldsymbol{f}$. In the discussion that follows, we propose a level set estimation method to estimate the discrete-domain level set S_N^* directly from $\boldsymbol{y}$ and show that $S_N^* \longrightarrow S^*$ as $N \longrightarrow \infty$ in section 3. Similar to [53], the error metric used to measure the closeness between S_N^* and a candidate estimate S is defined as

$$\varepsilon_N(S, S_N^*) = \frac{1}{N} \sum_{i \in \Delta(S_N^*, S)} |\gamma - f_i|, \tag{1.3}$$

where $\Delta(S_N^*, S) \triangleq \{i \in (S_N^* \setminus S) \cup (S \setminus S_N^*)\}$ denotes the symmetric set difference between S and S_N^*. Note that (1.3) can be interpreted as an empirical, weighted probability of error under the counting measure where the weights depend on the amplitude of the signal relative to the level set threshold γ. Our error metric penalizes (a) the symmetric difference between a level set estimate S and the true level set S_N^*, and (b) the errors along regions of the level set boundary corresponding to abrupt intensity variations more than the regions where the intensity varies smoothly. This performance measure is ideally suited for the level set estimation problem since in many applications, such as localizing the hot-spots signifying tumors in biomedical imaging, it is more desirable for an algorithm to accurately localize regions with sharp intensity variations.

Instead of working directly with the error metric, we make use of the *risk* of a candidate set S, defined as

$$R_N(S) \triangleq \frac{1}{N} \sum_i \ell_i(S), \tag{1.4}$$

where

$$\ell_i(S) \triangleq (\gamma - f_i) \left[\mathbb{I}_{\{i \in S\}} - \mathbb{I}_{\{i \notin S\}}\right] \tag{1.5}$$

is the *loss* function and $\mathbb{I}_{\{E\}} = 1$ if event E is true and 0 otherwise. The loss function in (1.5) measures the distance between the signal value at location i, f_i, and the threshold, γ, and weights this distance by -1 or 1 according to whether $i \in S$ or not. The loss function $\ell_i(S_N)$ is positive if $i \in \Delta(S_N^*, S)$ and is negative otherwise. To see this, observe that for all $i \in S_N^* \setminus S$, $(\gamma - f_i) \leq 0$ and $\left[\mathbb{I}_{\{i \in S\}} - \mathbb{I}_{\{i \notin S\}}\right] = -1$. A similar explanation holds for all $i \in S \setminus S_N^*$ as well. Note that the risk is related to the error metric defined in (1.3) by virtue of the fact that

$$\begin{aligned} R_N(S) - R_N(S_N^*) &= \frac{1}{N} \sum_i (\gamma - f_i) \left(\left[\mathbb{I}_{\{i \in S\}} - \mathbb{I}_{\{i \notin S\}}\right] - \left[\mathbb{I}_{\{i \in S_N^*\}} - \mathbb{I}_{\{i \notin S_N^*\}}\right] \right) \\ &= \frac{2}{N} \sum_{i \in \Delta(S_N^*, S)} |\gamma - f_i| = 2\varepsilon_N(S, S_N^*). \end{aligned} \tag{1.6}$$

[2] In this work, we adopt the terminology of "function" for the continuous-domain f, and "signal" for its discrete counterpart $\boldsymbol{f}$.

Finding an estimator that minimizes the *excess risk error* $\varepsilon_N(S, S_N^*)$ is thus equivalent to finding an estimator that minimizes $R_N(S)$, since $R_N(S_N^*)$ is simply a constant with respect to S.

This paper presents an optimization problem for choosing an estimate of S_N^* from the data $\boldsymbol{y}$ and theoretical characterization of $\varepsilon_N(S_N, S_N^*)$ when $\boldsymbol{f}$ consists of samples of a piecewise smooth function.

2. Our contribution and relation with previous work. In this work, we demonstrate that, subject to certain conditions on $\boldsymbol{A}$ and the ℓ_1-norm of $\boldsymbol{f}$, the level set S^* can be estimated quickly and accurately via S_N^* without first reconstructing $\boldsymbol{f}$. For $\boldsymbol{A} = \boldsymbol{I}$, [53] provides minimax optimal, tree-based level set estimation techniques to extract S^* from noisy observations $\boldsymbol{y} = \boldsymbol{f} + \boldsymbol{n} \in \mathbb{R}^N$ without estimating $\boldsymbol{f}$. We cannot directly apply those results to our problem since $\boldsymbol{A} \neq \boldsymbol{I}$. Instead, we draw on the key idea of constructing *proxy observations*,

$$\boldsymbol{z} = \boldsymbol{A}^T\boldsymbol{y} = \boldsymbol{f} + \underbrace{\left(\boldsymbol{A}^T\boldsymbol{A} - \boldsymbol{I}\right)\boldsymbol{f} + \boldsymbol{A}^T\boldsymbol{n}}_{\boldsymbol{n}'}, \tag{2.1}$$

from the literature on support detection of sparse signals (see, e.g., [3, 15, 16]) and then exploit some of the important insights from [53] to address our problem. A part of this work was previously published in [26]. This work, however, significantly expands on the previous work and presents new and tighter theoretical bounds and extensive simulation experiments.

Before we present our estimation method, we discuss prior work on level set estimation and sparse support detection.

2.1. Previous work on level set estimation. Large volumes of research have been dedicated to the problem of estimating level sets of an unknown density or a regression function f from its noisy measurements by either using plug-in estimators that find level sets of estimates of f [5, 11, 37, 44, 35] or direct methods that do not involve an intermediate reconstruction step [48, 43, 53, 41, 42]. Plug-in methods are easy to implement and in some cases lead to theoretical results on consistency and convergence based on some smoothness assumptions on the function of interest. For instance, [5, 11, 37] propose plug-in methods based on kernel estimators and show that they exhibit fast rates of convergence. Mason and Polonik [35] derive the asymptotic normality of the symmetric difference between a true level set and an estimate derived using a kernel density–based plug-in estimator. Singh, Scott, and Nowak [44] propose a plug-in method based on a regular histogram partition that minimizes the Hausdorff distance between the true and the estimated level sets. They also demonstrate that the proposed method adapts to unknown regularity parameters and achieves near minimax optimality on a wide variety of density function classes.

In the specific $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{f} + \boldsymbol{n}$ case studied in this paper, a number of plug-in methods can be proposed by exploiting the vast literature on ill-posed inverse problems [25]. Two popular and computationally simple methods in this regard are the truncated singular value decomposition (TSVD) (also known as the pseudoinverse solution) and Tikhonov regularization. While both these methods lead to fast plug-in approaches to level set estimation, essentially involving first an estimation of $\boldsymbol{f}$ from $\boldsymbol{y}$ and then thresholding of the resulting estimate, we do not expect these approaches to perform well in practice. This is because both TSVD and

Tikhonov regularization focus on "minimum-energy solutions," which effectively involves projecting $\boldsymbol{y}$ onto the principal subspace of $\boldsymbol{A}$. In the case of underdetermined $\boldsymbol{A}$, however, sparse signal processing research in the last decade or so has established the suboptimal nature of such "subspace approaches" to ill-posed inverse problems [31]. Instead, the state-of-the-art approach to ill-posed linear inverse problems with an underdetermined $\boldsymbol{A}$ involves projecting $\boldsymbol{y}$ onto a "union of subspaces" [14], accomplished through the use of either total-variation (TV) regularization [52, 40] or ℓ_1 regularization [6].

While the aforementioned plug-in approaches to level set estimation seem attractive, they solve a much harder problem as an intermediate step to solving a set estimation problem—a problem that is simpler than function estimation. Vapnik's principle stated earlier, together with the minimax convergence results shown in the context of classification problems in [54], tells us that plug-in methods are often suboptimal to direct estimation methods. As a result, in our work we focus on direct set estimation strategies.

Several researchers have considered direct set estimation methods for the case $\boldsymbol{A} = \boldsymbol{I}$. In [48], Tsybakov proposes a direct density level set estimation method that finds piecewise polynomial estimators of the true level set and achieves optimal minimax rates of convergence. The estimation method in [48] is hard to compute and cannot be directly extended to our problem where $\boldsymbol{A} \neq \boldsymbol{I}$. In [41], the authors show the theoretical and practical advantages of reducing a regression level set estimation problem to a cost-sensitive classification problem. Previous work by one of the coauthors [53] draws on the relationship between classification and level set estimation frameworks and proposes a set estimation method based on dyadic decision trees by exploiting some of the ideas from [43]. A closely related work is the estimation of minimum volume sets such that their masses are at least greater than some specified γ [42]. In that work, the authors discuss tree-based techniques and provide universal consistency results and rates of convergence.

We briefly review the basic idea in [53] upon which our set estimation strategy is built. The goal in that work was to design an estimator of the form

$$\widehat{S} = \arg\min_{S \in \mathcal{S}_M} \widehat{R}_N(S) + \mathrm{pen}(S),$$

where $\mathcal{S}_M$ is a class of candidate estimates, $\widehat{R}_N$ is an empirical measure of the estimator risk based on N noisy observations of the signal $\boldsymbol{f}$, and $\mathrm{pen}(\cdot)$ is a regularization term which penalizes improbable level sets. That work described choices for $\widehat{R}_N$, $\mathrm{pen}(\cdot)$, and $\mathcal{S}_M$ that made $\widehat{S}$ rapidly computable and minimax optimal for a large class of level set problems. Specifically, it derived a regularizer $\mathrm{pen}(\cdot)$ using Hoeffding's inequality for bounded random variables [22] and developed a dyadic tree-based framework for obtaining $\widehat{S}$. Trees were utilized for a couple of reasons. First, they both restricted and structured the space of potential estimators in a way that allowed the global optimum to be both rapidly computable and very close to the best possible (not necessarily tree-based) estimator. Second, they allowed the estimator selection criterion to be spatially adaptive, which was critical for the formation of provably optimal estimators. Note that while we intend to build upon the insights developed in [53], an extension of those techniques to the case of proxy observations in (2.1) is made nontrivial for two reasons. First, the *effective noise* $\boldsymbol{n}'$ is nonzero-mean because of the presence of $\left(\boldsymbol{A}^T\boldsymbol{A} - \boldsymbol{I}\right)\boldsymbol{f}$. Second, and most importantly, $\boldsymbol{n}'$ is *correlated* due to the nonunitary nature

of $\boldsymbol{A}$, which prohibits the use of canonical Hoeffding's inequality [22] for characterization of the penalty term.

2.2. Relationship with previous work on sparse support detection. Sparse support detection is the problem of detecting a set of locations $S_N^* = \{i : f_i \neq 0\}$ corresponding to a discrete signal $\boldsymbol{f} \in \mathbb{R}^N$, given observations of the form in (1.1). This is a special case of level set estimation, and the two are equivalent if $\boldsymbol{f}$ is nonnegative and $\gamma = 0$. The idea of constructing proxy observations $\boldsymbol{z}$ to deduce certain properties of the underlying $\boldsymbol{f}$ has been successfully employed in the recent compressed sensing and statistics literature to solve the problem of support detection of a discrete $\boldsymbol{f}$ having no more than m nonzero entries; see, e.g., [3, 15, 16, 19]. Specifically, it is established in [3] that the support of an m-sparse $\boldsymbol{f}$ can be reliably and quickly detected from appropriately thresholded proxy observations with overwhelming probability as long as $\boldsymbol{A}$ satisfies a certain, easily verifiable coherence property. The success of this thresholding method stems primarily from the sparsity assumption on $\boldsymbol{f}$. However, when $\boldsymbol{f}$ is not sparse, as is the case in level set estimation, simply thresholding the proxy observations will result in numerous false positives and misses, as discussed in detail in the numerical experiments in section 7; see Figures 2(a), 2(b), and Figures 3(a)–3(c). These results clearly suggest that we cannot simply use a support detection algorithm and an optimally chosen threshold to achieve an accurate level set estimation. In contrast, our methodology relies on a novel two-step approach that enables us to work with proxy observations without requiring $\boldsymbol{f}$ to be sparse.

3. Fast level set estimation from projection measurements. In order to extract the γ-level set of $\boldsymbol{f}$ from $\boldsymbol{y}$, we propose a novel two-step procedure. First, we construct a proxy of $\boldsymbol{f}$ according to (2.1), which allows us to arrive at the canonical signal plus noise observation model. Next, we perform level set estimation on the proxy observations $\boldsymbol{z}$, rather than on $\boldsymbol{y}$, using a method similar to the one derived in [53].[3] We refer to the resulting estimator as the *projective level set estimator*. Note that for any unitary $\boldsymbol{A}$, $\boldsymbol{z}$ in (2.1) reduces to $\boldsymbol{y} = \boldsymbol{f} + \tilde{\boldsymbol{n}}$ with $\tilde{\boldsymbol{n}}$ having independent, zero-mean entries. However, for nonunitary $\boldsymbol{A}$, the proxy defined in (2.1) creates a signal-dependent interference term $\left(\boldsymbol{A}^T\boldsymbol{A} - \boldsymbol{I}\right)\boldsymbol{f}$ and a zero-mean *correlated* noise term $\boldsymbol{A}^T\boldsymbol{n}$.

Intuitively, if we tried to make a decision about each z_i independently, then we would be vulnerable to noise (see, e.g., Figures 3(b) and 3(c)). On the other hand, if we consider *patches* p_{j_s} of z_i's, defined as groups of s proxy measurements (z_i's) centered around z_j for $s \in \{1, \ldots, N\}$, and force each patch to be wholly inside or outside the level set estimate, then we increase our robustness to noise but also increase our bias. Ideally, we want spatially adaptive patches that allow us to balance between an accurate approximation of the true level set boundary and estimator variance. It is in this vein that we theoretically analyze the impact of $\boldsymbol{n}'$ on the level set estimation problem and use our analysis to develop a spatially adaptive, dyadic, tree-based level set estimation approach that adapts to both the interference and the correlated noise term.

[3] There is another equivalent understanding of our approach to level set estimation, which helps connect it to the classical literature on inverse problems. The proxy observations $\boldsymbol{z}$ can be thought of as setting up the *normal equations* $\boldsymbol{A}^T\boldsymbol{y} = \boldsymbol{A}^T\boldsymbol{A}\widehat{\boldsymbol{f}}$. Instead of first solving the normal equations for one of infinitely many $\widehat{\boldsymbol{f}}$'s, arising due to the underdetermined nature of $\boldsymbol{A}$, our approach can be construed as estimating the level set directly from the normal equations.

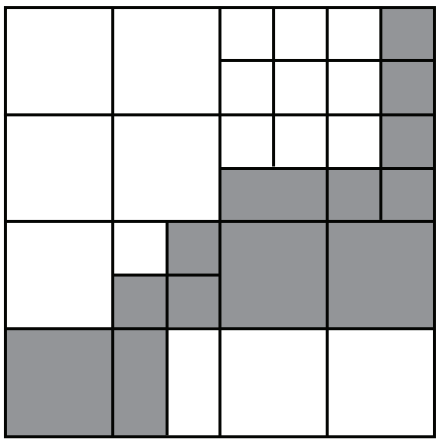


Figure 1. *An example level set estimate $S \in \mathcal{S}_M$, where the domain of the underlying signal is $[0,1]^2$. Shaded regions are estimated to be outside the level set.*

The algorithm that we propose basically works by using $\boldsymbol{z}$ to find a partition of $\boldsymbol{f}$ into a collection of disjoint sets of "pixels." For each set, we determine whether it is inside or outside the level set with a simple voting procedure—i.e., we determine whether the majority of the z_i's in the set are greater than gamma. Thus searching for the optimal level set estimate amounts to searching for a good partition of $\boldsymbol{f}$ and then performing empirical risk minimization, defined in (3.2) below, on that partition. We restrict our attention to partitions defined using binary trees because they yield tractable algorithms and, in the case where $\boldsymbol{A} = \boldsymbol{I}$, minimax optimality [53].

Specifically, let $\mathcal{S}_M$ be a collection of candidate level set estimates for a dyadic M (i.e., $M = 2^q$ for some positive integer q), where each $S \in \mathcal{S}_M$ is obtained by recursively partitioning the domain of $\boldsymbol{f}$ in dyadic intervals. The number of dyadic intervals along different coordinate directions is not required to be the same. In other words, each cell in the partition can potentially have different sidelengths, and the sidelength of the smallest cell is $1/M$. An estimate $S \in \mathcal{S}_M$ is obtained by assigning each cell in the partition to be inside or outside of the level set. Figure 1 shows one such estimate in two dimensions, where the shaded regions are the partition cells that are estimated to be outside the level set. Though we do not specify M in terms of N here, we derive an upper bound on M as a function of N that achieves a certain expected excess risk in Theorem 3.2.

Given $\boldsymbol{z}$, our goal is to find a level set estimate

$$\widetilde{S}_N = \operatorname*{arg\,min}_{S \in \mathcal{S}_M} R_N(S) - R_N(S_N^*) = \operatorname*{arg\,min}_{S \in \mathcal{S}_M} R_N(S), \tag{3.1}$$

where $R_N(\cdot)$ is defined in (1.4) and the second equality follows since $R_N(S_N^*)$ is a constant. (Note that $\widetilde{S}_N = S_N^*$ if $S_N^* \in \mathcal{S}_M$.) Since $\boldsymbol{f}$ is unknown, $R_N(S)$ cannot be computed; instead, let us consider an empirical risk of the form

$$\widehat{R}_N(S) = \frac{1}{N} \sum_{i=1}^{N} (\gamma - z_i) \left[\mathbb{I}_{\{i \in S\}} - \mathbb{I}_{\{i \notin S\}} \right]. \tag{3.2}$$

We show that finding an estimate that minimizes a penalized empirical risk results in an estimate that asymptotically approaches $\widetilde{S}_N$. Specifically, we find

$$\widehat{S}_N = \operatorname*{arg\,min}_{S \in \mathcal{S}_M} \widehat{R}_N(S) + \mathrm{pen}_N(S), \tag{3.3}$$

where $\mathrm{pen}_N(S)$ is an interference-dependent penalty term that yields $\left|R_N(\widehat{S}_N) - R_N(\widetilde{S}_N)\right| \overset{N \longrightarrow \infty}{\longrightarrow} 0$ subject to certain conditions on $\boldsymbol{A}$, which generally require $K \longrightarrow \infty$ as $N \longrightarrow \infty$. The penalty term plays a major role in our estimation strategy and is crucial in finding estimates that hone in on the boundary of the level set S_N^*. We thus focus on designing a spatially adaptive penalty $\mathrm{pen}_N(S)$ that promotes well-localized level sets with potentially nonsmooth boundaries. Let $\pi(S)$ be the partition induced by an estimate $S \in \mathcal{S}_M$; i.e., $\pi(S)$ is the collection of all leaves in the estimate $S \in \mathcal{S}_M$. Figure 1 shows a partition induced by one of the estimates $S \in \mathcal{S}_M$, where every white or gray shaded block is a leaf. We assign a label $\ell(L)$ to each leaf L depending on whether L is in the level set ($\ell(L) = 1$) or otherwise ($\ell(L) = 0$). Then the risk of S in each of its leaf $L \in \pi(S)$ is given by

$$R_N(L) \triangleq \frac{1}{N} \sum_{i=1}^{N} (\gamma - f_i) \left[\mathbb{I}_{\{\ell(L)=1\}} - \mathbb{I}_{\{\ell(L)=0\}}\right] \mathbb{I}_{\{i \in L\}}.$$

Note that $R_N(S) = \sum_{L \in \pi(S)} R_N(L)$. We design a spatially adaptive penalty term by analyzing $R_N(L) - \widehat{R}_N(L)$ within each leaf separately. To facilitate our analysis, let us define

$$\widetilde{R}_N(L) \triangleq \frac{1}{N} \sum_{i=1}^{N} (\gamma - \mathbb{E}[z_i]) \left[\mathbb{I}_{\{\ell(L)=1\}} - \mathbb{I}_{\{\ell(L)=0\}}\right] \mathbb{I}_{\{i \in L\}}.$$

Then

$$\begin{aligned}
\left|R_N(L) - \widehat{R}_N(L)\right| &= \left|R_N(L) - \widetilde{R}_N(L) + \widetilde{R}_N(L) - \widehat{R}_N(L)\right| \\
&= \left|\frac{1}{N} \sum_{i=1}^{N} \left[(\mathbb{E}[z_i] - f_i) + (z_i - \mathbb{E}[z_i])\right] \left[\mathbb{I}_{\{\ell(L)=1\}} - \mathbb{I}_{\{\ell(L)=0\}}\right] \mathbb{I}_{\{i \in L\}}\right| \\
&\leq \underbrace{\left|\frac{1}{N} \sum_{i=1}^{N} (\mathbb{E}[z_i] - f_i) \left[\mathbb{I}_{\{\ell(L)=1\}} - \mathbb{I}_{\{\ell(L)=0\}}\right] \mathbb{I}_{\{i \in L\}}\right|}_{T_1} \\
&\quad + \underbrace{\left|\frac{1}{N} \sum_{i=1}^{N} (z_i - \mathbb{E}[z_i]) \left[\mathbb{I}_{\{\ell(L)=1\}} - \mathbb{I}_{\{\ell(L)=0\}}\right] \mathbb{I}_{\{i \in L\}}\right|}_{T_2}.
\end{aligned} \tag{3.4}$$

Note that while T_1 is a measure of the bias in $\boldsymbol{z}$, T_2 is a measure of the concentration of $\boldsymbol{z}$ about its mean. Since the columns of $\boldsymbol{A}$ are assumed to have unit ℓ_2-norms, one can easily see from (2.1) that

$$z_i = f_i + \sum_{j=1, j \neq i}^{N} f_j \left\langle \boldsymbol{A}^{(i)}, \boldsymbol{A}^{(j)} \right\rangle + \left\langle \boldsymbol{A}^{(i)}, \boldsymbol{n} \right\rangle, \tag{3.5}$$

where $\boldsymbol{A}^{(i)}$ denotes the ith column of $\boldsymbol{A}$ and $\langle \cdot, \cdot \rangle$ denotes the usual innerproduct. Since $\boldsymbol{A}$ is given and $\boldsymbol{n}$ is zero-mean, the term

$$\mathbb{E}[z_i] - f_i = \sum_{j=1, j\neq i}^{N} f_j \left\langle \boldsymbol{A}^{(i)}, \boldsymbol{A}^{(j)} \right\rangle \tag{3.6}$$

in T_1 is the signal-dependent interference term at the ith location due to the signal energies at other locations. We upper-bound T_1 by the ℓ_1-norm of $\boldsymbol{f}$ and the worst-case coherence of $\boldsymbol{A}$ (defined in the statement of Theorem 3.1), bound T_2 using a Hoeffding-like inequality for a weighted sum of independent sub-Gaussian random variables [39], and sum the risk in each leaf of the estimate S to arrive at the following result.

Theorem 3.1 (concentration of risk around the empirical risk). *Suppose that the entries of noise $\boldsymbol{n}$ are sub-Gaussian distributed with parameter c_s. Then, for $\delta \in [0, 1/2]$ and $c > 0$, with probability at least $1 - 2\delta$, the following holds for all $S \in \mathcal{S}_M$:*

$$\left| R_N(S) - \widehat{R}_N(S) \right| \leq \left(\frac{N-1}{N} \right) \mu(\boldsymbol{A}) \|\boldsymbol{f}\|_1 + \mathrm{pen}_N(S), \tag{3.7}$$

where $\|\boldsymbol{f}\|_1 = \sum_i |f_i|$ is the ℓ_1-norm of $\boldsymbol{f}$,

$$\mathrm{pen}_N(S) \triangleq \sum_{L \in \pi(S)} \frac{1}{N} \sqrt{\frac{[\log(2/\delta) + [\![L]\!] \log 2]\, c_s^2 \sum_{i,j \in L} \left\langle \boldsymbol{A}^{(i)}, \boldsymbol{A}^{(j)} \right\rangle}{2c}} \tag{3.8}$$

is the penalty term,

$$\mu(\boldsymbol{A}) \triangleq \max_{i,j \in \{1,\ldots,N\}, i \neq j} \left| \left\langle \boldsymbol{A}^{(i)}, \boldsymbol{A}^{(j)} \right\rangle \right|$$

is the worst-case coherence of $\boldsymbol{A}$, and $[\![L]\!]$ is the number of bits in a prefix code used to uniquely encode the position of a leaf L in the tree.

The proof of this theorem is provided in section 5.1. The above bound holds for *any* prefix code $[\![L]\!]$. In order to achieve the error rates in Theorem 3.2, we use a certain prefix code, which is discussed before the statement of Theorem 3.2. Note that the bounds in (3.7) and (3.8) depend on (a) the signal-dependent interference term in (2.1) through $\|\boldsymbol{f}\|_1$, (b) the noise statistics through c_s, (c) the choice of $\boldsymbol{A}$ through $\mu(\boldsymbol{A})$, (d) the depth of each leaf through $[\![L]\!]$, (e) the size of each leaf through $\sum_{i,j \in L} \left\langle \boldsymbol{A}^{(i)}, \boldsymbol{A}^{(j)} \right\rangle$, and (f) the parameter δ. Ideally we would like to minimize $R_N(S)$ to obtain $\widetilde{S}_N$ in (3.1). Since $R_N(S)$ is bounded by $\widehat{R}_N(S) + \mathrm{pen}_N(S)$, minimizing the bound (3.7) will ensure that our estimate $\widehat{S}_N$ in (3.3) is as close to $\widetilde{S}_N$ in (3.1) as possible. In order to minimize the risk difference in (3.7), one needs to choose an estimate $S \in \mathcal{S}_M$ that has the least $\mathrm{pen}_N(S)$ in (3.8). The penalty term in (3.8) is directly proportional to the number of leaves in the partition $\pi(S)$ and the size of each leaf through the term $\sum_{i,j \in L} \left\langle \boldsymbol{A}^{(i)}, \boldsymbol{A}^{(j)} \right\rangle$. As a result, searching for an estimate $S \in \mathcal{S}_M$ that minimizes (3.3) will favor estimates with few, deep leaves that hone in on the boundary of the level set.

The theoretical analysis of our method is significantly different from the analysis in [53] because of the statistics of the noise term $\boldsymbol{n}'$ in our problem. However, this only changes

the way the penalty is defined in our setup. As a result, we can adapt the computational techniques discussed in [53] to compute our estimator in an efficient way. Our method is computationally efficient since the proxy computation needs at most $O(KN)$ operations (fewer if $\boldsymbol{A}$ is structured; e.g., $\boldsymbol{A}$ is a Toeplitz matrix), and the level set estimation method needs $O(N \log N)$ operations, as noted in [53].

3.1. Performance analysis. As discussed earlier in section 1.1, our eventual goal is to estimate the continuous-domain level set S^* from discrete measurements $\boldsymbol{y}$. In this section, we show that estimating the discrete-domain level set S_N^* helps us achieve this goal by (i) establishing that $S_N^* \longrightarrow S^*$ as $N \longrightarrow \infty$ and (ii) providing conditions, as a function of problem parameters, under which the discrete-domain level set estimate obtained according to (3.3) approaches S^*. To this end, we can utilize the results of Theorem 3.1 to upper-bound the expected excess risk $\mathbb{E}\big[R(\widehat{S}_N) - R(S^*)\big]$, taken with respect to the noise distribution, in terms of the problem parameters, where

$$R(S) = \int_{[0,1]^d} (\gamma - f(x)) \big[\mathbb{I}_{\{x \in S\}} - \mathbb{I}_{\{x \notin S\}}\big] dx$$

is the definition of risk in the continuous domain. The expected excess risk is a measure of the effectiveness of our level set estimator. Before studying it, we make certain assumptions about the smoothness of f in the vicinity of the level set boundary. Let ∂S^* represent the level set boundary corresponding to $S^* = \{x : f(x) > \gamma\}$. We assume that f is in a box-counting function class $\mathcal{D}_{\mathrm{BOX}}(\kappa, \gamma, c_1, c_2)$ for $c_1, c_2 > 0$ and $1 \leq \kappa \leq \infty$ [53] such that the following hold:

(a) If we partition $[0,1]^d$ to m^d equisized cells for $m \leq M$, with each of them having a sidelength of $1/m$ and volume m^{-d}, then the number of such cells intersected by the level set boundary $N_{S^*}(m) \leq c_1 m^{d-1}$. This ensures that ∂S^* varies smoothly and is not an irregular, space-filling curve.

(b) For all dyadic m, let

$$S_m^* = \operatorname*{arg\,min}_{S \in \mathcal{S}_m} \lambda(\Delta(S, S^*)) \tag{3.9}$$

be a candidate in $\mathcal{S}_m$ that minimizes the symmetric difference between any $S \in \mathcal{S}_m$ and the true level set S^* in terms of the Lebesgue measure λ. For this S_m^*, the excess risk in the continuous domain follows

$$\varepsilon(S_m^*, S^*) = R(S_m^*) - R(S^*) \triangleq \int_{\Delta(S^*, S_m^*)} |\gamma - f(x)| dx \leq c_2 m^{-\kappa}. \tag{3.10}$$

Parameter κ and the assumption on the excess risk in (3.10) allow us to study the fluctuations of f around ∂S^* and thus examine the behavior of $\varepsilon(S_m^*, S^*)$ in the vicinity of the level set boundary. If f exhibits a very small fluctuation around ∂S^*, then $\varepsilon(S_m^*, S^*)$ is small even if the symmetric difference $\Delta(S^*, S_m^*)$ is very large, since the excess risk is weighted by how close f is to γ. In other words, a high value of κ indicates that f varies very smoothly around the level set boundary, and a low value of κ means that there is a jump in f around ∂S^*.

Recall that S_N^* is obtained by partitioning the space $[0,1]^d$ to N equisized cells of side-lengths N^{-d} and assigning each cell to be inside or outside of the level set. From (3.10), for $m = N^{1/d}$,

$$R(S_N^*) - R(S^*) \leq c_2 N^{-\kappa/d}.$$

Thus $S_N^* \longrightarrow S^*$ as $N \longrightarrow \infty$, and estimation of S^* via S_N^* is reasonable.

To achieve the results of Theorem 3.2 stated below, we adapt the prefix code proposed in [53, 43]. According to [53, 43], a leaf L of a level set at depth j of the tree can be uniquely encoded using a total of $j(\log_2 d + 2) + 1$ bits. Specifically, one needs $j+1$ bits to encode the depth of the leaf, j bits to encode whether each of its ancestors corresponded to a left or a right branch of the tree, and $j \log_2 d$ bits to encode the orientation of each of the j branches.

Before we state our main theorem, let us clarify the notation used in the following. For a given set of sequences a_n and b_n, $a_n \preccurlyeq b_n$ implies that there exists a constant $C > 0$ such that $a_n \leq C b_n$ for all n and $a_n \asymp b_n$ implies that there exists constants C_1 and C_2 such that $C_1 a_n \leq b_n \leq C_2 a_n$ for all n.

Theorem 3.2 (upper bound on the expected excess risk). *If $f \in \mathcal{D}_{\mathrm{BOX}}(\kappa, \gamma, c_1, c_2)$ is discretized according to (1.2), $-B \leq f(x) \leq B$ for $x \in [0,1]^d$, $-B \leq \gamma \leq B$, and the estimate $\widehat{S}_N$ is chosen according to (3.3) with $\mathrm{pen}_N(\widehat{S}_N)$ defined according to Theorem 3.1, then, for a given $\boldsymbol{A}$, $d \geq 2$, and for $M \succcurlyeq \left(\frac{N}{\|\boldsymbol{A}\|_2^2 \log N}\right)^{1/d}$,*

$$\mathbb{E}\Big[R(\widehat{S}_N) - R(S^*)\Big] \preccurlyeq \left(\frac{\|\boldsymbol{A}\|_2^2 \log N}{N}\right)^{\frac{\kappa}{2\kappa+d-2}} + \mu(\boldsymbol{A})\|\boldsymbol{f}\|_1, \tag{3.11}$$

where the expectation is with respect to the noise distribution, $\|\boldsymbol{A}\|_2$ is the spectral norm of $\boldsymbol{A}$, $\|\boldsymbol{A}\|_2 \triangleq \sqrt{\lambda_{\max}(\boldsymbol{A}^T\boldsymbol{A})}$, and $\mu(\boldsymbol{A})$ is the worst-case coherence of $\boldsymbol{A}$.

The proof of this theorem is given in section 5.2. This theorem tells us how the expected excess risk scales with the dimensionality N of the underlying signal $\boldsymbol{f}$, the ℓ_1-norm of $\boldsymbol{f}$, the choice of $\boldsymbol{A}$, and the smoothness of the underlying function around the level set boundary through the parameter κ. For a unitary matrix $\boldsymbol{A}$, $\|\boldsymbol{A}\|_2 = 1$ since its singular values are all equal to 1, $\mu(\boldsymbol{A}) = 0$ and $\mathbb{E}\big[R(\widehat{S}_N) - -R(S^*)\big] \preccurlyeq \left(\frac{\log N}{N}\right)^{\frac{\kappa}{2\kappa+d-2}}$, which is the minimax optimal rate derived in [53] without the projection matrix $\boldsymbol{A}$. Since in practice $\boldsymbol{A}$ is dictated by the physics of the measurement system, it is not always unitary. In such cases, the above theorem tells us how any given $\boldsymbol{A}$ increases this bound. For some $\boldsymbol{A}$, such as the one discussed in the following section, $\mu(\boldsymbol{A}) \longrightarrow 0$ as N and K go to ∞ since $\boldsymbol{A}^T\boldsymbol{A} \overset{N,K\longrightarrow\infty}{\longrightarrow} \boldsymbol{I}$. Note that (3.11) can be specified in terms of the continuous-domain function f by noting that $\|\boldsymbol{f}\|_1 \leq N\|f\|_{L_1}$, although it is a loose bound when the function f is not completely positive (or negative).

Corollary 3.3 (performance with random projections). *If the entries of $\boldsymbol{A} \in \mathbb{R}^{K\times N}$ are drawn from $\mathcal{N}(0, 1/K)$, and the columns of $\boldsymbol{A}$ are normalized to have unit ℓ_2-norm, then*

$$\mathbb{E}\Big[R(\widehat{S}_N) - R(S^*)\Big] \preccurlyeq \left(\frac{\log N}{N}\right)^{\frac{\kappa}{2\kappa+d-2}} \left[\frac{\sqrt{K}+\sqrt{N}}{\sqrt{K-\sqrt{12K\log N}}}\right]^{\frac{2\kappa}{2\kappa+d-2}} + \frac{\sqrt{15\log N}}{\sqrt{K}-\sqrt{12\log N}}\|\boldsymbol{f}\|_1 \tag{3.12}$$

holds with probability at least $1 - \left[e^{-c(K+N)} + N^{-2} + 11N^{-1}\right]$ *as long as* $60 \log N \leq K \leq \frac{N-1}{4 \log N}$.

The proof of this corollary is provided in section 5.3. The result above yields an upper bound on the expected excess risk as a function of the dimensions of the projection operator $\boldsymbol{A}$ and $\|\boldsymbol{f}\|_1$. In words, this corollary states that the expected excess risk in the case of random Gaussian projections is minimized if the number of measurements K scales linearly with N and increases if K scales sublinearly with N. Dependence of the estimator's performance on the ℓ_1-norm of $\boldsymbol{f}$ is due to the interference term $\left(\boldsymbol{A}^T\boldsymbol{A} - \boldsymbol{I}\right)\boldsymbol{f}$ that arises during the proxy construction. The foregoing results provide key insights into ways in which we can minimize the expected excess risk and improve performance, as discussed in detail in the following section.

We conclude our discussion of Theorem 3.2 by pointing out that practically meaningful lower bounds for this problem are unknown at this time, but would be the subject of a future investigation. In addition, note that our focus in here has been on very fast, easily implementable methods for real-time estimation of level sets. While significantly slower methods could conceivably be developed to potentially provide lower errors, such methods would not be able to compete with our proposed approach in terms of the computational costs (see, e.g., section 7).

4. Performance improvement via projected median subtraction. So far we have shown that the signal-dependent interference term in (2.1) leads to a penalty term proportional to $\|\boldsymbol{f}\|_1$ in (3.7). This implies that the interference in $\boldsymbol{z}$ and thus the performance of our method may worsen with the increase in $\|\boldsymbol{f}\|_1$, which is indeed confirmed by the experimental results in section 7. To find a way to minimize the signal-dependent interference, let us write $\boldsymbol{f} = \widetilde{\boldsymbol{f}} + \lambda\mathbb{1}$, where λ is a constant DC offset such that

$$\left\|\widetilde{\boldsymbol{f}}\right\|_1 \leq \|\boldsymbol{f}\|_1. \tag{4.1}$$

If we have access to an estimate $\widehat{\lambda}$ of λ, then we can minimize the signal-dependent interference by subtracting a projection of this constant offset to obtain

$$\begin{aligned}\widetilde{\boldsymbol{y}} = \boldsymbol{y} - \boldsymbol{A}\widehat{\lambda}\mathbb{1} &= \boldsymbol{A}\left(\widetilde{\boldsymbol{f}} + \lambda\mathbb{1}\right) + \boldsymbol{n} - \boldsymbol{A}\widehat{\lambda}\mathbb{1} \\ &= \boldsymbol{A}\left(\widetilde{\boldsymbol{f}} + \left(\lambda - \widehat{\lambda}\right)\mathbb{1}\right) + \boldsymbol{n} \approx \boldsymbol{A}\widetilde{\boldsymbol{f}} + \boldsymbol{n},\end{aligned}$$

assuming that $\widehat{\lambda} \approx \lambda$. The proxy observations in this case reduce to

$$\widetilde{\boldsymbol{z}} = \boldsymbol{A}^T\widetilde{\boldsymbol{y}} \approx \boldsymbol{A}^T\boldsymbol{A}\widetilde{\boldsymbol{f}} + \boldsymbol{A}^T\boldsymbol{A}\boldsymbol{n} = \widetilde{\boldsymbol{f}} + \left(\boldsymbol{A}^T\boldsymbol{A} - \boldsymbol{I}\right)\widetilde{\boldsymbol{f}} + \boldsymbol{A}^T\boldsymbol{A}\boldsymbol{n}.$$

Since $S_N^* = \{i : f_i > \gamma\} \triangleq \{i : \widetilde{f}_i > \widetilde{\gamma}\}$, where $\widetilde{\gamma} = \gamma - \lambda$, we can estimate S_N^* from $\widetilde{\boldsymbol{z}}$ using our level set estimation method discussed in the previous section.

If we let λ be the median of $\boldsymbol{f}$, then we can easily show that (4.1) holds for this particular choice of λ. Note that if λ is the median of $\boldsymbol{f}$, then half the pixel values of $\boldsymbol{f}$ are below the median and half of the pixel values are above the median. Let $\mathcal{G} = \{i : f_i > \lambda\}$ and

$\mathcal{G}^c = \{i : f_i < \lambda\}$. The cardinality of $\mathcal{G}$ is $|\mathcal{G}| = N/2$ for N even.[4] By the definition of median, $|\mathcal{G}| = |\mathcal{G}^c|$. Then

$$\begin{aligned}\left\|\widetilde{\boldsymbol{f}}\right\|_1 &= \|\boldsymbol{f} - \lambda\mathbb{1}\|_1 = \sum_{i\in\mathcal{G}} |f_i - \lambda| + \sum_{i\in\mathcal{G}^c} |f_i - \lambda| \\ &= \sum_{i\in\mathcal{G}} (f_i - \lambda) + \sum_{i\in\mathcal{G}^c} (\lambda - f_i) = \sum_{i\in\mathcal{G}} f_i + \sum_{i\in\mathcal{G}^c} -f_i - |\mathcal{G}|\lambda + |\mathcal{G}^c|\lambda \\ &= \sum_{i\in\mathcal{G}} f_i + \sum_{i\in\mathcal{G}^c} -f_i \\ &\le \sum_{i\in\mathcal{G}} |f_i| + \sum_{i\in\mathcal{G}^c} |f_i| = \|\boldsymbol{f}\|_1.\end{aligned}$$

In practice, however, estimation of the median of $\boldsymbol{f}$ from $\boldsymbol{y}$ might be hard, though the estimation of the mean of $\boldsymbol{f}$ might be tractable. For instance, if we construct $\boldsymbol{A}' = \left[\begin{smallmatrix}\mathbb{1}^T \\ \boldsymbol{A}\end{smallmatrix}\right]$ (i.e., the first row of $\boldsymbol{A}'$ is $\mathbb{1}^T$), then $\boldsymbol{y}' = \boldsymbol{A}'\boldsymbol{f} + \boldsymbol{n} = \left[\begin{smallmatrix}y_1' \\ \boldsymbol{y}\end{smallmatrix}\right]$ and $\widehat{\lambda} = y_1'/N = (\sum_i f_i + n_1)/N = \lambda + n_1/N$. If the observation noise is negligible or if N is large, then $\widehat{\lambda} \approx \lambda$, and we can perform projected mean subtraction, instead of a projected median subtraction, to reduce the signal-dependent interference. While (4.1) does not always hold if λ is the mean of $\boldsymbol{f}$, simulation results in section 7 suggest that projected mean subtraction can result in significant improvement in performance.

5. Proofs of theorems and corollaries.

This section presents the proofs of all the theorems and corollaries stated before.

5.1. Proof of Theorem 3.1 (concentration of risk).

Let us begin by bounding T_1 and T_2 in (3.4) separately. Let $\widehat{p}_L = \sum_{i\in L} \frac{1}{N}$ be the ratio of the number of observations in leaf L to the total number of observations N. From the statistics of z, we can bound T_1 as follows:

$$\begin{aligned}T_1 &\le \frac{1}{N} \sum_{i,j:j\neq i} |f_j| \left|\left\langle \boldsymbol{A}^{(i)}, \boldsymbol{A}^{(j)} \right\rangle\right| \left|\left[\mathbb{I}_{\{\ell(L)=1\}} - \mathbb{I}_{\{\ell(L)=0\}}\right]\right| \mathbb{I}_{\{i\in L\}} \\ &\le \frac{\mu(\boldsymbol{A})}{N} \sum_{i\in L} \sum_{j=1:j\neq i}^{N} |f_j| = \frac{\mu(\boldsymbol{A})}{N} \sum_{i\in L} \left(\sum_{j=1}^{N} |f_j| - |f_i| \right) \\ &\le \mu(\boldsymbol{A})\widehat{p}_L \|\boldsymbol{f}\|_1 - \frac{\mu(\boldsymbol{A})}{N} \sum_{i\in L} |f_i|, \qquad (5.1)\end{aligned}$$

where the second inequality is due to the fact that $|[\mathbb{I}_{\{\ell(L)=1\}} - \mathbb{I}_{\{\ell(L)=0\}}]| = 1$ and $|\langle \boldsymbol{A}^{(i)}, \boldsymbol{A}^{(j)}\rangle| \le \mu(\boldsymbol{A})$ for all $j \neq i$.

Rewriting T_2 in terms of (3.5) and (3.6), we have

$$T_2 = \frac{1}{N} \sum_{i\in L} \left(\sum_{k=1}^{K} a_{k,i} n_k \right) \left[\mathbb{I}_{\{\ell(L)=1\}} - \mathbb{I}_{\{\ell(L)=0\}}\right] = \sum_{k=1}^{K} b_k n_k,$$

[4] We do not consider N to be odd since our recursive dyadic partitions require N to be in powers of two.

where $b_k = \frac{1}{N}\sum_{i\in L} a_{k,i}\left[\mathbb{I}_{\{\ell(L)=1\}} - \mathbb{I}_{\{\ell(L)=0\}}\right]$. Observe that T_2 is a weighted sum of K independent, zero-mean, sub-Gaussian random variables. It then follows from a Hoeffding-like inequality for a weighted sum of independent, zero-mean sub-Gaussian random variables [39, Theorem 3.3] that

$$\mathbb{P}\left(\left|\sum_{k=1}^{K} b_k n_k\right| \geq \epsilon\right) \leq 2\exp\left(\frac{-c\epsilon^2}{c_s^2\sum_{k=1}^{K} b_k^2}\right) \tag{5.2}$$

for $\epsilon > 0$, where $c > 0$ is an absolute numerical constant. Let us now evaluate the term $\sum_{k=1}^{K} b_k^2$ in the above expression as follows:

$$\begin{aligned}\sum_{k=1}^{K} b_k^2 &= \sum_{k=1}^{K}\left(\frac{1}{N}\sum_{i\in L} a_{k,i}\left[\mathbb{I}_{\{\ell(L)=1\}} - \mathbb{I}_{\{\ell(L)=0\}}\right]\right)^2 \\ &= \frac{1}{N^2}\sum_{k=1}^{K}\sum_{i\in L} a_{k,i}\left[\mathbb{I}_{\{\ell(L)=1\}} - \mathbb{I}_{\{\ell(L)=0\}}\right]\sum_{j\in L} a_{k,j}\left[\mathbb{I}_{\{\ell(L)=1\}} - \mathbb{I}_{\{\ell(L)=0\}}\right] \\ &= \frac{1}{N^2}\sum_{k=1}^{K}\sum_{i\in L}\sum_{j\in L} a_{k,i}a_{k,j} = \frac{1}{N^2}\sum_{i\in L}\sum_{j\in L}\left\langle \boldsymbol{A}^{(i)}, \boldsymbol{A}^{(j)}\right\rangle, \end{aligned} \tag{5.3}$$

where the above equation is due to the fact that $\left[\mathbb{I}_{\{\ell(L)=1\}} - \mathbb{I}_{\{\ell(L)=0\}}\right]^2 = 1$. By substituting (5.3) into (5.2) and by setting the right-hand side of (5.2) equal to $\delta_L \in (0, 1/2)$ and solving for ϵ, we can show that, with probability at least $1 - 2\delta_L$,

$$T_2 \leq \sqrt{\frac{\log(1/\delta_L)c_s^2\sum_{i,j\in L}\left\langle \boldsymbol{A}^{(i)}, \boldsymbol{A}^{(j)}\right\rangle}{2cN^2}}. \tag{5.4}$$

Applying the bounds in (5.1) and (5.4) to (3.4), we can see that with probability at least $1 - 2\delta_L$ the following holds:

$$\begin{aligned}\left|R_N(L) - \widehat{R}_N(L)\right| \leq &\left(\mu(\boldsymbol{A})\widehat{p}_L\|\boldsymbol{f}\|_1 - \frac{\mu(\boldsymbol{A})}{N}\sum_{i\in L}|f_i|\right) \\ &+ \sqrt{\frac{\log(1/\delta_L)c_s^2\sum_{i,j\in L}\left\langle \boldsymbol{A}^{(i)}, \boldsymbol{A}^{(j)}\right\rangle}{2N^2}}.\end{aligned}$$

Thus for a given $S \in \mathcal{S}_M$ the risk difference $|R_N(S) - \widehat{R}_N(S)|$ is upper-bounded by summing the bound corresponding to each leaf separately. Since $\sum_{L\in\pi(S)}\widehat{p}_L = 1$ and $\sum_{L\in\pi(S)}\sum_{i\in L}|f_i| = \|\boldsymbol{f}\|_1$ we have

$$\left|R_N(S) - \widehat{R}_N(S)\right| \leq \mu(\boldsymbol{A})\left(\frac{N-1}{N}\right)\|\boldsymbol{f}\|_1 + \sum_{L\in\pi(S)}\sqrt{\frac{\log(1/\delta_L)c_s^2\sum_{i,j\in L}\left\langle \boldsymbol{A}^{(i)}, \boldsymbol{A}^{(j)}\right\rangle}{2N^2}}$$

with high probability. If we let $\delta_L = \delta 2^{-([\![L]\!]+1)}$, where $[\![L]\!]$ is the number of bits required to uniquely encode the position of leaf L, then it is straightforward to follow the proof of Lemma 2 in [53] to show that the bound above holds for every $S \in \mathcal{S}_M$, which leads to the result of Theorem 3.1. ■

5.2. Proof of Theorem 3.2 (performance analysis). In order to analyze the performance of our estimator, we will draw upon the proof techniques and the associated performance analyses in previous works on classification and level set estimation [43, 53]. Note that some of the steps in our analysis that are adapted from [43, 53] are repeated here for readability.

The proof of this theorem follows by relating the continuous-domain risk of a level set $S \in \mathcal{S}_M$ to its discrete counterpart and exploiting the results from Theorem 3.1. By expanding $R(S)$ for any $S \in \mathcal{S}_M$ in terms of the discretization of f in (1.2), we have

$$\begin{aligned} R(S) &= \int_x (\gamma - f(x))\left[\mathbb{I}_{\{x\in S\}} - \mathbb{I}_{\{x\notin S\}}\right] dx \\ &= \sum_{i=1}^{N} \int_{C_i} (\gamma - f(x))\left[\mathbb{I}_{\{C_i\in S\}} - \mathbb{I}_{\{C_i\notin S\}}\right] dx \\ &= \sum_{i=1}^{N} \left(\gamma \mathrm{vol}\,(C_i) - \mathrm{vol}\,(C_i)\, f_i\right)\left[\mathbb{I}_{\{C_i\in S\}} - \mathbb{I}_{\{C_i\notin S\}}\right] \\ &= \sum_{i=1}^{N} \left(\frac{\gamma}{N} - \frac{f_i}{N}\right)\left[\mathbb{I}_{\{i\in S\}} - \mathbb{I}_{\{i\notin S\}}\right] \equiv R_N(S), \end{aligned}$$

where the second equality holds since C_i is contained either in S or in the complement of S. Since $\widehat{S}_N \in \mathcal{S}_M$, $R(\widehat{S}_N) = R_N(\widehat{S}_N)$. Let us consider some $S'_N \in \mathcal{S}_M$ that minimizes the penalized excess risk between any $S \in \mathcal{S}_M$ and the true level set S^*, i.e.,

$$S'_N = \min_{S\in\mathcal{S}_M} [R(S) - R(S^*) + 2\mathrm{pen}_N(S)].$$

From the definitions of $\widehat{S}_N$ in (3.3) and S'_N, and the results of Theorem 3.1, the following holds with probability at least $1 - 2\delta$ for $\delta \in [0, 1/2]$:

$$R(\widehat{S}_N) - R(S^*) = R_N(\widehat{S}_N) - R(S^*) \le \min_{S\in\mathcal{S}_M} [R(S) - R(S^*) + 2\mathrm{pen}_N(S)]. \tag{5.5}$$

Let Ω denote the event that (3.7) from Theorem 3.1 holds for all proxy observations $\boldsymbol{z}$. Since $-B \le f(x) \le B$ for all $x \in [0,1]^d$ and $-B \le \gamma \le B$, for $\delta = 1/N$

$$\begin{aligned} \mathbb{E}\Big[R(\widehat{S}_N) - R(S^*)\Big] &= \mathbb{E}\Big[R_N(\widehat{S}_N) - R(S^*)\Big] \\ &= \mathbb{E}\Big[R_N(\widehat{S}_N) - R(S^*)|\Omega\Big]\mathbb{P}(\Omega) + \mathbb{E}\Big[R_N(\widehat{S}_N) - R(S^*)|\Omega^c\Big]\mathbb{P}(\Omega^c) \\ &\le \mathbb{E}\Big[R_N(\widehat{S}_N) - R(S^*)|\Omega\Big] + \mathbb{E}\Big[R_N(\widehat{S}_N) - R(S^*)|\Omega^c\Big]\frac{2}{N} \\ &\le \min_{S\in\mathcal{S}_M} [R(S) - R(S^*) + 2\mathrm{pen}_N(S)] + 4B \times \frac{2}{N}, \end{aligned} \tag{5.6}$$

where the first term in (5.6) is due to (5.5) and the second term is due to the boundedness assumption on $f(x)$ and γ. Specifically,

$$\begin{aligned} R_N(\widehat{S}_N) - R(S^*) &\equiv R(\widehat{S}_N) - R(S^*) \\ &= \int_x (\gamma - f(x)) \Big[\mathbb{I}_{\{x \in \widehat{S}_N\}} - \mathbb{I}_{\{x \notin \widehat{S}_N\}} - \mathbb{I}_{\{x \in S^*\}} + \mathbb{I}_{\{x \notin S^*\}} \Big] dx \\ &\leq \int_x 4B dx = 4B \end{aligned}$$

since $\gamma - f(x) \leq 2B$ and $\mathbb{I}_{\{x \in S\}} - \mathbb{I}_{\{x \notin S\}} \leq 1$. Rewriting (5.6), we have

$$\mathbb{E}\Big[R(\widehat{S}_N) - R(S^*)\Big] \leq \min_{S \in \mathcal{S}_M} [R(S) - R(S^*) + 2\mathrm{pen}_N(S)] + \frac{8}{N} \tag{5.7}$$

$$\leq \min_{1 \leq m \leq M} \min_{S \in \mathcal{S}_m} [R(S) - R(S^*) + 2\mathrm{pen}_N(S)] + \frac{8}{N} \tag{5.8}$$

$$\leq \min_{1 \leq m \leq M} R(S_m^*) - R(S^*) + 2\mathrm{pen}_N(S_m^*) + \frac{8}{N} \tag{5.9}$$

$$\leq \min_{1 \leq m \leq M} m^{-\kappa} + 2\mathrm{pen}_N(S_m^*) + \frac{8}{N}, \tag{5.10}$$

where S_m^* in (5.9) is defined in (3.9) and (5.10) is due to (3.10).

Let us now bound $\mathrm{pen}_N(S_m^*)$ given in (3.8). To this end, let us rewrite

$$\mathrm{pen}_N(S_m^*) = \left(\frac{N-1}{N}\right) \mu(\boldsymbol{A}) \|\boldsymbol{f}\|_1 + \mathrm{pen}'_N(S_m^*), \tag{5.11}$$

where

$$\mathrm{pen}'_N(S_m^*) = \sum_{L \in \pi(S_m^*)} \frac{1}{N} \sqrt{\frac{[\log(2N) + [\![L]\!] \log 2] \, |c_u - c_\ell|^2 \sum_{i,j \in L} \left\langle \boldsymbol{A}^{(i)}, \boldsymbol{A}^{(j)} \right\rangle}{2}}$$

and bound $\mathrm{pen}'_N(S_m^*)$.

To keep the notation simple, let $|L| = \left(\sum_{i \in L} 1\right)$ be the number of pixels in leaf L, and $\widetilde{\boldsymbol{A}}_L$ be a $K \times |L|$ matrix formed by collecting the columns of $\boldsymbol{A}$ corresponding to the indices $i \in L$. Note that $|L| = \sum_{i \in L} 1 = N\widehat{p}_L$. Let

$$q_L = [\log(2N) + [\![L]\!] \log 2] \left(|c_u - c_\ell|^2/2\right) \widehat{p}_L. \tag{5.12}$$

Using this notation, we can write

$$\mathrm{pen}'_N(S_m^*) = \sum_{L\in\pi(S_m^*)} \sqrt{\frac{[\log(2N) + [\![L]\!]\log 2]\,|c_u - c_\ell|^2 \left[\mathbb{1}^T_{(|L|\times 1)}\left(\widetilde{\boldsymbol{A}}_L^T\widetilde{\boldsymbol{A}}_L\right)\mathbb{1}_{(|L|\times 1)}\right]}{2N^2}}$$

$$= \sum_{L\in\pi(S_m^*)} \sqrt{\frac{q_L}{N}\left[\frac{\mathbb{1}^T_{(|L|\times 1)}\left(\widetilde{\boldsymbol{A}}_L^T\widetilde{\boldsymbol{A}}_L\right)\mathbb{1}_{(|L|\times 1)}}{N\widehat{p}_L}\right]} = \sum_{L\in\pi(S_m^*)} \sqrt{\frac{q_L}{N}\left[\frac{\left\|\widetilde{\boldsymbol{A}}_L\mathbb{1}_{(|L|\times 1)}\right\|_2^2}{|L|}\right]}$$

$$= \sum_{L\in\pi(S_m^*)} \sqrt{\frac{q_L}{N}}\frac{\left\|\widetilde{\boldsymbol{A}}_L\mathbb{1}_{(|L|\times 1)}\right\|_2}{\sqrt{|L|}} \le \sum_{L\in\pi(S_m^*)} \sqrt{\frac{q_L}{N}}\frac{\left\|\widetilde{\boldsymbol{A}}_L\right\|_2\left\|\mathbb{1}_{(|L|\times 1)}\right\|_2}{\sqrt{|L|}} \tag{5.13}$$

$$= \sum_{L\in\pi(S_m^*)} \sqrt{\frac{q_L}{N}}\frac{\left\|\widetilde{\boldsymbol{A}}_L\right\|_2\sqrt{|L|}}{\sqrt{|L|}} = \sum_{L\in\pi(S_m^*)} \sqrt{\frac{q_L}{N}}\left\|\widetilde{\boldsymbol{A}}_L\right\|_2 \le \|\boldsymbol{A}\|_2 \sum_{L\in\pi(S_m^*)} \sqrt{\frac{q_L}{N}}, \tag{5.14}$$

where the inequality in (5.13) follows from the definition of the spectral norm of $\widetilde{\boldsymbol{A}}_L$ given below:

$$\left\|\widetilde{\boldsymbol{A}}_L\right\|_2 = \max_{\boldsymbol{x}\neq\boldsymbol{0}} \frac{\left\|\widetilde{\boldsymbol{A}}_L\boldsymbol{x}\right\|_2}{\|\boldsymbol{x}\|_2} \ge \frac{\left\|\widetilde{\boldsymbol{A}}_L\mathbb{1}_{(|L|\times 1)}\right\|_2}{\left\|\mathbb{1}_{(|L|\times 1)}\right\|_2}.$$

The term $\sum_{L\in\pi(S_m^*)}\sqrt{q_L/N}$ in (5.14) can now be bounded from above by using the proof techniques in [43, 53]. Previous work [43] showed that for a binary tree with N leaves at its finest level, $[\![L]\!] \preccurlyeq \log N$. Note that $\widehat{p}_L = \sum_{i\in L}\frac{1}{N} \triangleq \sum_{i\in L}\int_{C_i} dx \triangleq \int_L dx = p_L = 2^{-j(L)}$, where $j(L)$ is the depth corresponding to leaf L of the tree. By substituting these results into (5.12), we have

$$\sum_{L\in\pi(S_m^*)} \sqrt{\frac{q_L}{N}} \preccurlyeq \sum_{L\in\pi(S_m^*)} \sqrt{\frac{[\log(2N) + \log N\log 2]\left(|c_u - c_\ell|^2/2\right)2^{-j(L)}}{N}}$$

$$= \sqrt{\frac{[\log(2N) + \log N\log 2]\left(|c_u - c_\ell|^2/2\right)}{N}} \sum_{L\in\pi(S_m^*)} 2^{-j(L)/2}$$

$$\le \sqrt{\frac{[\log(2N) + \log N\log 2]\left(|c_u - c_\ell|^2/2\right)}{N}} \sum_{j=1}^{J} T_j\sqrt{2^{-j}}$$

$$\le \sqrt{\frac{\log N}{N}} c m^{d/2-1}, \tag{5.15}$$

where $J = \log_2 N$ is the deepest level of the binary tree, T_j is the number of leaves at depth j of the tree, c is a constant that is a function of the upper and lower bounds, c_u and c_ℓ, on noise,

and (5.15) follows straightforwardly from the proof of Theorem 6 in [43]. By substituting (5.15) into (5.14), we have the following:

$$\text{pen}'_N(S_m^*) \preccurlyeq m^{d/2-1}\sqrt{\frac{\log N}{N}}\|\boldsymbol{A}\|_2. \tag{5.16}$$

From (5.10), (5.11), and (5.16),

$$\begin{aligned}\mathbb{E}\Big[R(\widehat{S}_N) - R(S^*)\Big] &\preccurlyeq \min_{1\le m\le M}\left\{m^{-\kappa} + m^{d/2-1}\sqrt{\frac{\log N}{N}}\|\boldsymbol{A}\|_2 + \left(\frac{N-1}{N}\right)\mu(\boldsymbol{A})\|\boldsymbol{f}\|_1 + \frac{8B}{N}\right\}\\ &\preccurlyeq \min_{1\le m\le M}\left\{m^{-\kappa} + m^{d/2-1}\sqrt{\frac{\log N}{N}}\|\boldsymbol{A}\|_2 + \mu(\boldsymbol{A})\|\boldsymbol{f}\|_1 + \frac{8B}{N}\right\}.\end{aligned}$$

We can easily show that $m \asymp \left(\frac{N}{\|\boldsymbol{A}\|_2^2 \log N}\right)^{\frac{1}{2\kappa+d-2}}$ minimizes the expression above. Since $1 \le \kappa \le \infty$, the bound on m is largest for $\kappa = 1$. Exploiting this result and the fact that $m \le M$, we have that for $M \succcurlyeq \left(\frac{N}{\|\boldsymbol{A}\|_2^2 \log N}\right)^{\frac{1}{d}}$

$$\mathbb{E}\Big[R(\widehat{S}_N) - R(S^*)\Big] \preccurlyeq \left(\frac{\|\boldsymbol{A}\|_2^2 \log N}{N}\right)^{\frac{\kappa}{2\kappa+d-2}} + \mu(\boldsymbol{A})\|\boldsymbol{f}\|_1. \tag{5.17}$$

■

5.3. Proof of Corollary 3.3 (performance with random projections). The proof of this corollary is obtained by bounding the spectral norm of $\boldsymbol{A}$ and the worst-case coherence of $\boldsymbol{A}$ with high probability. Let $\widetilde{\boldsymbol{A}} \in \mathbb{R}^{K\times N}$ be a matrix whose entries are independent and identically distributed draws from $\mathcal{N}(0, 1/K)$. Each column of $\boldsymbol{A}$ is then simply obtained by normalizing the columns of $\widetilde{\boldsymbol{A}}$; that is, $\boldsymbol{A}^{(i)} = \frac{\widetilde{\boldsymbol{A}}^{(i)}}{\|\widetilde{\boldsymbol{A}}^{(i)}\|_2}$ for $i \in \{1, \dots, N\}$. The bound on $\|\boldsymbol{A}\|_2$ is obtained by first showing that

$$\|\boldsymbol{A}\|_2^2 \le q\left\|\widetilde{\boldsymbol{A}}\right\|_2^2 \tag{5.18}$$

for some constant q and then bounding $\|\widetilde{\boldsymbol{A}}\|_2$ using the results from random matrix theory. In particular, [51] states that the spectral norm of an $K \times N$ sub-Gaussian matrix $\boldsymbol{M}$ is upper-bounded by $\|\boldsymbol{M}\|_2 \le c\big(\sqrt{K} + \sqrt{N}\big)$ with probability $1 - \exp(-c(K+N))$. This result can be straightforwardly extended to show that

$$\left\|\widetilde{\boldsymbol{A}}\right\|_2^2 \le c^2\Big(\sqrt{N/K} + 1\Big)^2 \tag{5.19}$$

with probability $1-\exp(-c(K+N))$. We show that (5.18) holds with high probability by taking the following approach:

$$\|\boldsymbol{A}\|_2^2 = \max_{\boldsymbol{x}:\|\boldsymbol{x}\|_2=1} \|\boldsymbol{A}\boldsymbol{x}\|_2^2 = \max_{\boldsymbol{x}:\|\boldsymbol{x}\|_2=1} \sum_i \left|\sum_j a_{i,j}x_j\right|^2 = \max_{\boldsymbol{x}:\|\boldsymbol{x}\|_2=1} \sum_i \left|\sum_j \frac{\widetilde{a}_{i,j}}{\left\|\widetilde{\boldsymbol{A}}^{(j)}\right\|_2} x_j\right|^2$$
$$= \max_{\boldsymbol{p}:\sum_j p_j^2\|\widetilde{\boldsymbol{A}}^{(j)}\|_2^2=1} \sum_i \left|\sum_j \widetilde{a}_{i,j}p_j\right|^2 \triangleq \max_{\boldsymbol{p}\neq\boldsymbol{0}} \frac{\sum_i \left|\sum_j \widetilde{a}_{i,j}p_j\right|^2}{\sum_j p_j^2\left\|\widetilde{\boldsymbol{A}}^{(j)}\right\|_2^2},$$

where $p_j = \frac{x_j}{\|\widetilde{\boldsymbol{A}}^{(j)}\|_2}$ and $\boldsymbol{p} = [p_1\ p_2\ \cdots\ p_N]^T$. Following the proofs of Lemma 1 in [28] and Theorem 8 in [4], we can easily show that $\|\widetilde{\boldsymbol{A}}^{(j)}\|_2^2 \geq 1 - \frac{\sqrt{12\log N}}{\sqrt{K}}$ with probability at least $1-N^{-3}$ for any $j \in \{1,\ldots,N\}$. Applying the union bound over every possible $j \in \{1,\ldots,N\}$, $\|\widetilde{\boldsymbol{A}}^{(j)}\|_2^2 \geq 1 - \frac{\sqrt{12\log N}}{\sqrt{K}}$ with probability at least $1-N^{-2}$. Using this result in the above equation, we have

$$\|\boldsymbol{A}\|_2^2 \leq \max_{\boldsymbol{p}} \frac{\sum_i \left|\sum_j \widetilde{a}_{i,j}p_j\right|^2}{\sum_j p_j^2\left(1-\frac{\sqrt{12\log N}}{\sqrt{K}}\right)} \triangleq \frac{1}{1-\frac{\sqrt{12\log N}}{\sqrt{K}}}\left\|\widetilde{\boldsymbol{A}}\right\|_2^2 \tag{5.20}$$

with probability exceeding $1-N^{-2}$. By substituting (5.19) into (5.20) and applying the union bound, the following holds with probability exceeding $1-\exp(-c(K+N))-N^{-2}$:

$$\|\boldsymbol{A}\|_2 \leq c\frac{\sqrt{N/K}+1}{\sqrt{1-\frac{\sqrt{12\log N}}{\sqrt{K}}}} = c\frac{\sqrt{K}+\sqrt{N}}{\sqrt{K-\sqrt{12K\log N}}}. \tag{5.21}$$

The rest of the proof follows straight from Theorem 8 of [4], which states that

$$\mu(\boldsymbol{A}) \leq \frac{\sqrt{15\log N}}{\sqrt{K}-\sqrt{12\log N}} \tag{5.22}$$

with probability exceeding $1-11N^{-1}$ as long as $60\log N \leq K \leq \frac{N-1}{4\log N}$. The bound in (5.22) together with the bound in (5.21) and the result of Theorem 3.2 yields the result of Corollary 3.3. ■

6. Relationship with plug-in methods. The success of wavelet-based methods in estimating a piecewise smooth function from noisy measurements suggests a potential extension of such methods to the problem of level set estimation [13]. For instance, one possible approach for level set estimation from projection measurements is to first estimate the underlying signal $\boldsymbol{f}$ from proxy measurements $\boldsymbol{z}$ using wavelet-based denoising methods and then threshold the resulting estimate at level γ. Estimating $\boldsymbol{f}$ from $\boldsymbol{y}$ through an intermediate proxy construction step is similar to the iterative hard thresholding method in compressive sensing literature with just one iteration [7]. While such plug-in estimation techniques using wavelet-based methods

offer practical solutions to the level set estimation problem, their estimation performances are not yet understood.

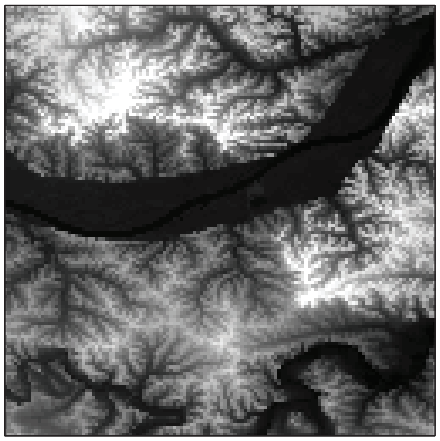
The proposed multiscale, partition-based set estimation method with proxy measurements can be thought of as a combination of an iterative hard thresholding method with just one iteration and wavelet-based denoising ideas. Specifically, our partition-based method is similar in spirit to the wavelet-based denoising ideas using the unnormalized Haar wavelet transform. Both wavelet-based methods and our method rely on the spatial homogeneity of the underlying signal $\boldsymbol{f}$ to perform level set estimation. The difference between the two methods stems from the way in which the wavelet coefficients are thresholded in each case. While the threshold in the wavelet-based method is chosen to minimize the mean squared error, our method thresholds the coefficients at levels that are tailored to the level set estimation problem. Since the proposed method shares similar ideas with wavelet-based methods, the proof techniques presented in this paper could potentially be extended to wavelet-based methods in order to characterize their estimation performances.

Compressive sensing theory presents a variety of algorithms such as iterative hard thresholding [7], basis pursuit [10], orthogonal matching pursuit [47], LASSO [46], and TV-based methods [6] to reliably estimate $\boldsymbol{f}$ from $\boldsymbol{y}$. One can readily use such algorithms to first estimate $\boldsymbol{f}$ and then threshold it, or use the method in [44] to estimate the level set. However, there are a couple of issues in using these plug-in methods to perform level set estimation. First, these approaches aim to minimize the mean squared error over the entire image. This, however, does not guarantee minimization of errors close to the level set boundaries, which is critical to the characterization of level set estimation performance. Second, the iterative nature of these algorithms makes them computationally intensive and time-consuming.

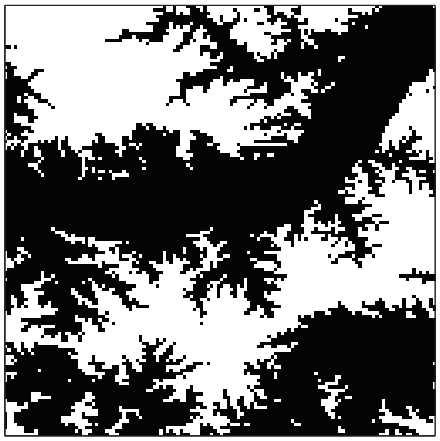
7. Experimental results. Due to the lack of a theoretical performance comparison between plug-in methods and our method, we present an empirical comparison of these methods in this section by conducting experiments on a test image. Simulation results discussed below demonstrate that the proposed partition-based, multiscale method using proxy observations has the following advantages: (a) it is a powerful tool to perform direct level set estimation from projection measurements, (b) it allows us to exploit the spatial homogeneity of the underlying function to perform set estimation, (c) it performs an order of magnitude better than thresholding methods that obtain level set estimates by simply thresholding the proxy observations at level γ, and (d) it yields results that are comparable to the results obtained using wavelet-based thresholding approaches.

In order to test the effectiveness of our projective level set estimator, we conduct experiments on a test image of size 128×128, shown in Figure 2(a). In these experiments, we are interested in estimating a γ-level set of this test image, shown in Figure 2(b), from noisy projection measurements of the form $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{f} + \boldsymbol{n} \in \mathbb{R}^K$ for $K < N = 128 \times 128$, without reconstructing $\boldsymbol{f}$ from $\boldsymbol{y}$. The entries of the projection operator in these experiments are drawn from $\mathcal{N}(0, 1/K)$, and the noise is distributed as $\boldsymbol{n} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$. We compare the performance of our method with the performances of the following approaches using the excess risk error metric defined in (1.3):

(a) *Thresholding method*, where the estimate $\widehat{S}_\gamma$ is simply obtained by thresholding the proxy observations $\boldsymbol{z}$ at level γ; that is, $\widehat{S}_\gamma = \{i : z_i \geq \gamma\}$.



(a) True signal $\boldsymbol{f} \in \mathbb{R}^{128\times128}$ such that $f_i \in [44, 239]$. We measure $K = 8192$ Gaussian random projections of this image.



(b) Level set $S_N^* = \{i : f_i > 125\}$ (white pixels) such that $|S_N^*| \approx 0.4285N$, where $N = 128 \times 128$.

Figure 2. *Snapshots of the true signal and its desired level set.*

(b) *Risk-optimal thresholding method*, where the estimate $\widehat{S}_{\widehat{\gamma}}$ is obtained by thresholding $\boldsymbol{z}$ at a level $\widehat{\gamma}$ that minimizes the excess risk; that is, $\widehat{S}_{\widehat{\gamma}} = \{i : z_i \geq \widehat{\gamma}\}$, where $\widehat{\gamma} = \arg\min_\gamma \varepsilon_N\big(\widehat{S}_\gamma, S_N^*\big)$.

(c) *Noniterative wavelet-based plug-in method*, where the estimate $\widehat{S}_w$ is obtained by first estimating $\boldsymbol{f}$ from $\boldsymbol{z}$ using translation invariant wavelet denoising, and then thresholding the resulting estimate $\widehat{\boldsymbol{f}}$ at level γ; that is, $\widehat{S}_w = \{i : \widehat{f}_i \geq \gamma\}$. In these experiments we perform wavelet denoising using Daubechies-4 wavelets and soft thresholding, where the threshold is chosen to minimize the excess risk.

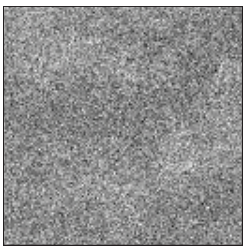
(d) *Total-variation (TV)-based plug-in method*, where the estimate $\widehat{S}_{TV}$ is obtained according to $\widehat{S}_{TV} = \{i : \widehat{f}_i^{(TV)} \geq \gamma\}$. The estimate $\widehat{\boldsymbol{f}}^{(TV)}$ of the input image $\boldsymbol{f}$ is obtained from $\boldsymbol{y}$ by solving

$$\widehat{\boldsymbol{f}}^{(TV)} = \arg\min_{\widetilde{\boldsymbol{f}}} \left\|\boldsymbol{y} - \boldsymbol{A}\widetilde{\boldsymbol{f}}\right\|_2^2 + \tau\left\|\widetilde{\boldsymbol{f}}\right\|_{TV},$$

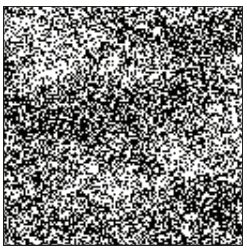
where $\|\widetilde{\boldsymbol{f}}\|_{TV}$ is the TV norm of $\widetilde{\boldsymbol{f}}$ and τ is a user-defined parameter that balances the log-likelihood term and the regularization term. Algorithms such as the two-step iterative shrinkage and thresholding (TwIST) method provide a way to efficiently solve for the above optimization problem [6]. In our experiments, τ is chosen to minimize the excess risk.

In these simulation experiments, we compute the excess risk clairvoyantly based on the knowledge of $\boldsymbol{f}$. We obtain the estimate $\widehat{S}$ using our projective level set estimator according to $\widehat{S}_N = \arg\min_{S\in\mathcal{S}_M} \widehat{R}(S) + \tau\mathrm{pen}(S)$ with a scaling factor τ, which is chosen to minimize $\varepsilon\big(\widehat{S}_N, S_N^*\big)$. In these experiments, we use $M = N$.

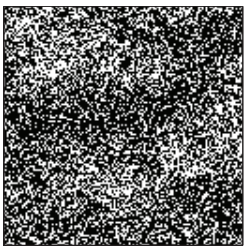
We evaluate the performance of all the competing algorithms discussed above, with and



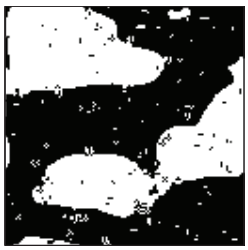
(a) Proxy observations.



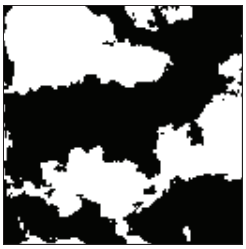
(b) Estimate obtained using the thresholding method; $\varepsilon_N = 15.21$.



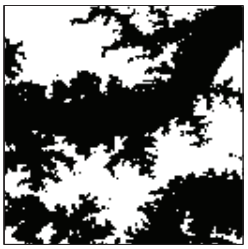
(c) Estimate obtained using the risk-optimal thresholding method; $\varepsilon_N = 14.83$.



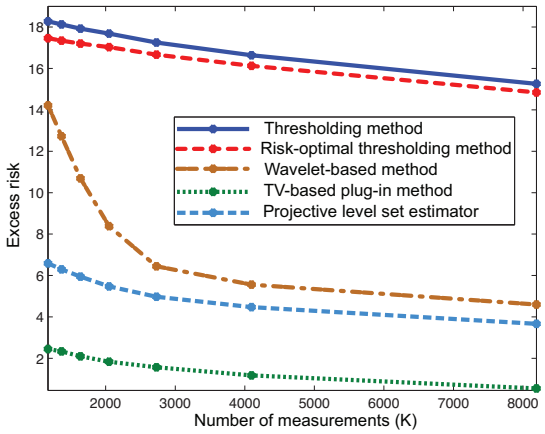
(d) Estimate obtained using the wavelet-based method; $\varepsilon_N = 4.824$.



(e) Estimate obtained using the projective level set estimator; $\varepsilon_N = 3.593$.



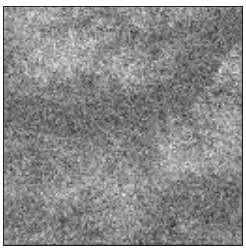
(f) Estimate obtained using the TV-based plug-in estimator; $\varepsilon_N = 0.5596$.



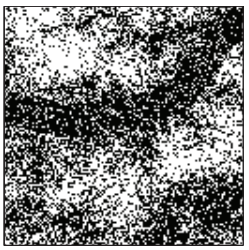
(g) Plot of excess risk as a function of $K < N = 16384$ without performing the projected mean subtraction.

Figure 3. *Snapshots of the simulation results obtained (without performing the projected mean subtraction) from observations of the form in* (1.1).

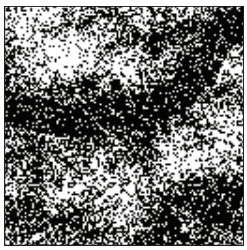
without projected mean subtraction discussed in section 4. The number of observations used in these experiments is $K = N/2 = 8192$. Figure 3(a) shows the proxy observations obtained without mean subtraction. Figure 3(b) shows the level set estimate obtained by simply thresholding the proxy observations at level γ, and Figure 3(c) shows the estimate obtained by performing the risk-optimal thresholding method. These results demonstrate that thresholding noisy proxy observations results in several false positives and misses. Though the wavelet-based plug-in method yields better results in comparison, as shown in Figure 3(d), the estimate is still severely oversmoothed and noisy. The estimate obtained using our projective level set estimator is shown in Figure 3(e). This approach yields lower excess risk compared to the other three approaches discussed above, preserves some of the fine details, but still performs some oversmoothing. Figure 3(f) shows the results obtained using the TV-based plug-in method. This method yields the best results compared to the other approaches and yields the smallest excess risk, at the expense of first estimating the signal. Figure 3(g) plots excess risk as a function of the number of measurements $K < N = 16384$ for all com-



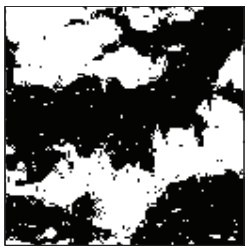
(a) Proxy observations after projected mean subtraction.



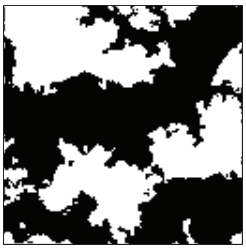
(b) Estimate obtained using the thresholding method; $\varepsilon_N = 8.562$.



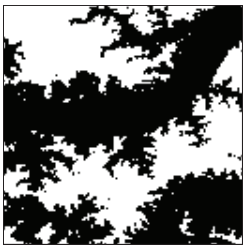
(c) Estimate obtained using the risk-optimal thresholding method; $\varepsilon_N = 8.231$.



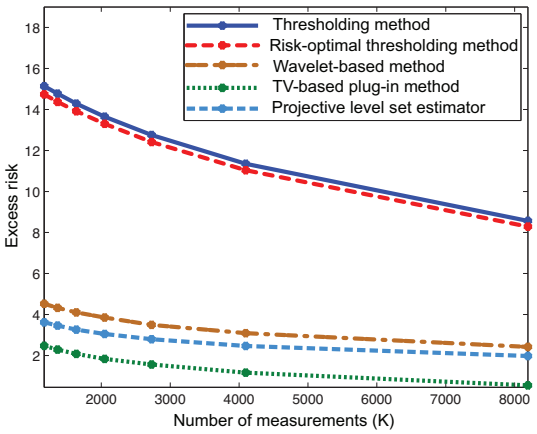
(d) Estimate obtained using the wavelet-based method; $\varepsilon_N = 2.393$.



(e) Estimate obtained using the projective level set estimator; $\varepsilon_N = 1.924$.



(f) Estimate obtained using the TV-based plug-in estimator; $\varepsilon_N = 0.5593$.

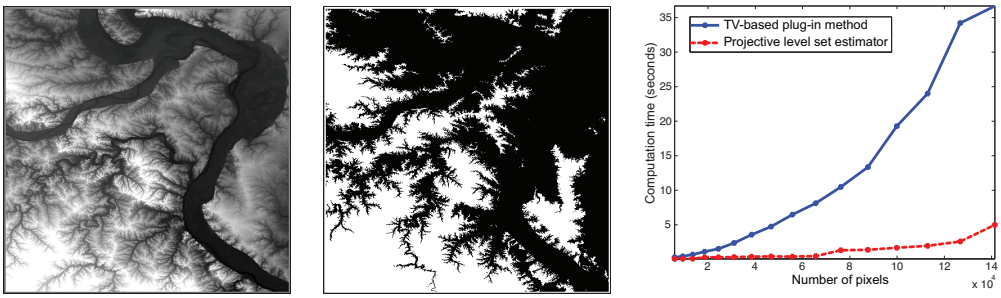


(g) Plot of excess risk as a function of $K < N = 16384$ after performing mean subtraction.

Figure 4. *Snapshots of the simulation results obtained (after performing projected mean subtraction) from observations of the form in* (1.1).

peting methods. These plots are obtained by averaging the results obtained over 200 different noise and projection matrix realizations.

Figures 4(a)–(g) show the improvements in results obtained because of the projected mean subtraction. The improvements stem from the fact that the proxy measurements are less "noisy" after the projected mean subtraction. This subtraction operation lowers the excess risk of the estimates obtained using every method discussed above, except for the TV-based plug-in method, which performs very well in practice irrespective of mean subtraction. TV-based reconstruction is in general implemented using iterative algorithms where convergence is achieved if the mean squared error between estimates obtained in successive iterations does not change beyond a user-specified tolerance value. The TwIST algorithm used in our simulation study uses the proxy measurements to initialize the iterative process and stops iterating when convergence is achieved. As a result, only the number of iterations to achieve a specified convergence will change, depending on the quality of the proxy observations, and not the final estimate. This explains why the TV-based results are insensitive to projected mean subtraction.



(a) (b) (c)

Figure 5. *Comparison of the computation times of the TV-based plug-in method and our projective level set estimator.* (a) *True signal* $\boldsymbol{f} \in \mathbb{R}^{512^2}$ *such that* $f_i \in [0, 239]$, (b) *level set* $S_N^* = \{i : f_i > 125\}$, *and* (c) *a plot of computation time as a function of problem size to approximately achieve the same excess risk. Images of different sizes (*76×76 *to* 376×376*) cropped from* (a) *were used in this experiment. The plots in* (c) *indicate that the time it takes for the TV-based plug-in method to achieve the same excess risk obtained by the projective level set estimator increases dramatically with problem size.*

The TV-based method seems to outperform our projective level set estimator since we evaluate the performance of these methods based solely on the excess risk and not on the computational resources required to achieve that excess risk. In that sense, this comparison is somewhat unfair. A more meaningful comparison would be to either evaluate the excess risk obtained within some unit time or compare the time taken by different approaches to achieve a desired excess risk as the problem size N changes. To make the comparison fair, we ran our projective level set estimator for different problem sizes, used $K \approx N/3$ observations to get our estimates, recorded the excess risk obtained in each case, and ran the TV-based plug-in method to achieve the same excess risk in each case. In other words, instead of using the conventional convergence strategy in the TV-based reconstruction algorithm, we stop iterating if the excess risk is less than or equal to that obtained using our method. We compare the computational time required for both these methods as a function of problem size. Figures 5(a) and 5(b) show a 512×512 image and its corresponding level set, respectively. Note that the image used in the above experiments is a cropped version of the image in Figure 5(a). We cropped this image in order to get images of different sizes; in particular, we used images of size $\ell \times \ell$, where $\ell = 76, 96, 116, \ldots, 376$. Figure 5(c) shows the time-gap between these two methods in achieving similar excess risks, as a function of the number of pixels in the input image. These plots show that the computational time taken by a TV-based plug-in method dramatically increases with problem size, whereas the computation time required by our projective level set estimator increases much more gracefully with problem size.

Before concluding, it is also important to comment on the performance of our approach in relation to that of faster plug-in methods, such as those based on the SVD of $\boldsymbol{A}$. As noted in section 2.1, we do not expect such methods to perform well in the underdetermined ($K < N$) setting for reasons outlined earlier. We have also verified this intuition through numerical experiments (not fully reported here for space reasons). Consider, for example, estimating the level set in Figure 2(b) by thresholding either TSVD or the Tikhonov regularized solution for the case of $K \approx N/2$. In this setting, the excess risks obtained using TSVD and Tikhonov

regularization-based plug-in methods are 14.26 and 14.31, respectively, whereas the excess risk using our proposed method is 3.593. This rather poor performance of SVD-based plug-in methods should not be too surprising. Such methods operate on the assumption that signals lie near a subspace, but a union-of-subspaces model is known to be a better model for real-world signals [31]. Our method performs better than SVD-based approaches since the family S_m over which we search for an estimate of the level set can be construed as a union-of-subspaces, with each subspace in the union being formed by a set of indices corresponding to dyadic, tree-based basis functions.

In conclusion, the experimental results indicate that estimating the underlying signal using TV regularization-based plug-in methods yields more accurate level set estimates compared to those obtained using our projective level set estimator. However, the real strengths of our method are two-fold. First, we can reliably perform *real-time* level set estimation, in contrast to plug-in methods, as shown by the time-gap versus problem size plot in Figure 5(c). Second, we can use our level set estimate to discard regions where the levels of interest are not present and design adaptive measurement schemes to hone in on the regions of interest. Such an adaptive measurement scheme is especially helpful in very high-dimensional settings, where the cost of collecting measurements and performing reconstruction tends to be extremely high.

8. Conclusion. This work proposes a theoretically sound and computationally efficient tree-based approach for extracting level sets of a function from projection measurements without reconstructing the underlying function. The simulation results presented in section 7 suggest that the proposed method may facilitate fast and accurate level set estimates from tomographic projections in medical imaging, Fourier projections in interferometry, or coded projections in compressive optical systems. One of the key advantages of our approach is that many of the operations on the proxy data are easily parallelizable. For instance, in problems where the domain of the signal of interest is very large, we can compute the proxy observations, partition the proxy data into different patches, run our estimation algorithm on each patch separately, and merge the results to identify the regions that correspond to the level set. In applications such as medical imaging, the time saved by collecting fewer projection measurements and parallelization can be significant and crucial.

Empirically, the accuracy of the projective level set estimate is comparable to that of a similar scheme based on wavelet thresholding or an iterative method with TV regularization. Currently, however, there is no theoretical support for these alternatives. Recent work studying the performance of so-called analysis regularization[49, 17] may lead to an improved understanding of theoretical performance bounds for the TV approach, but as we show here this iterative solution requires significantly more computational resources. Our approach is much more similar in spirit to the wavelet-based approach, and the theoretical techniques employed in our analysis may lead to an improved understanding of this and other fast, non-iterative approaches. Furthermore, adaptive sampling schemes such as the one discussed in [19] suggest a potential extension of our method. Specifically, [19] proposes collecting noisy measurements of a sparse signal, estimating its support, and collecting more measurements based on the estimated support to adaptively focus the computational resources on regions of interest. The underlying assumption in such "distilled sensing" [20] schemes is sparsity. Since our level set estimation method offers a way to estimate the level set of a function without re-

quiring sparsity, we expect it to facilitate the development of new adaptive sampling routines that perform better than the ones proposed in earlier works.

REFERENCES

[1] *Special Issue on Compressive Sampling*, IEEE Signal Process. Mag., 25 (2008).

[2] I. Ayed, A. Mitiche, and Z. Belhadj, *Multiregion level-set partitioning of synthetic aperture radar images*, IEEE Trans. Pattern Anal. Machine Intell., 27 (2005), pp. 793–800.

[3] W. U. Bajwa, R. Calderbank, and S. Jafarpour, *Why Gabor frames? Two fundamental measures of coherence and their role in model selection*, J. Commun. Netw., 12 (2010), pp. 289–307.

[4] W. U. Bajwa, R. Calderbank, and D. G. Mixon, *Two are better than one: Fundamental parameters of frame coherence*, Appl. Comput. Harmon. Anal., 33 (2012), pp. 58–78.

[5] A. Baillo, *Total error in a plug-in estimator of level sets*, Stat. Probab. Lett., 65 (2003), pp. 411–417.

[6] J. Bioucas-Dias and M. Figueiredo, *A New TwIST: Two-Step Iterative Shrinkage/Thresholding Algorithms for Image Restoration*, IEEE Trans. Image Process., 16 (2007), pp. 2992–3004.

[7] T. Blumensath and M. Davies, *Iterative hard thresholding for compressed sensing*, Appl. Comput. Harmon. Anal., 27 (2009), pp. 265–274.

[8] E. Candès, J. Romberg, and T. Tao, *Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information*, IEEE Trans. Inform. Theory, 52 (2006), pp. 489–509.

[9] E. Candés and T. Tao, *Near optimal signal recovery from random projections: Universal encoding strategies*, IEEE Trans. Inform. Theory, 52 (2006), pp. 5406–5425.

[10] S. Chen and D. Donoho, *Basis pursuit*, in Proceedings of the Conference Record of the Twenty-Eighth Asilomar Conference on Signals, Systems and Computers, IEEE Press, Piscataway, NJ, 1994, pp. 41–44.

[11] A. Cuevas, W. González-Manteiga, and A. Rodríguez-Casal, *Plug-in estimation of general level sets*, Aust. N. Z. J. Stat., 48 (2006), pp. 7–19.

[12] D. Donoho, *Compressed sensing*, IEEE Trans. Inform. Theory, 52 (2006), pp. 1289–1306.

[13] D. Donoho and I. Johnstone, *Ideal spatial adaptation by wavelet shrinkage*, Biometrika, 81 (1994), pp. 425–455.

[14] M. F. Duarte and Y. C. Eldar, *Structured compressed sensing: From theory to applications*, IEEE Trans. Signal Process., 59 (2011), pp. 4053–4085.

[15] A. Fletcher, S. Rangan, and V. Goyal, *Necessary and sufficient conditions for sparsity pattern recovery*, IEEE Trans. Inform. Theory, 55 (2009), pp. 5758–5772.

[16] C. Genovese, J. Jin, and L. Wasserman, *Revisiting Marginal Regression*, arXiv preprint arXiv:0911.4080, 2009.

[17] R. Giryes and M. Elad, *RIP-based near-oracle performance guarantees for SP, CoSaMP, and IHT*, IEEE Trans. Signal Process., 60 (2012), pp. 1465–1468.

[18] Z. Harmany, R. Willett, A. Singh, and R. Nowak, *Controlling the error in FMRI: Hypothesis testing or set estimation?*, in Proceedings of the 5th IEEE International Symposium on Biomedical Imaging: From Nano to Macro, IEEE Press, Piscataway, NJ, 2008, pp. 552–555.

[19] J. Haupt, R. Baraniuk, R. Castro, and R. Nowak, *Compressive distilled sensing: Sparse recovery using adaptivity in compressive measurements*, in Proceedings of the Forty-Third IEEE Asilomar Conference on Signals, Systems and Computers, 2009, pp. 1551–1555.

[20] J. Haupt, R. Castro, and R. Nowak, *Distilled sensing: Selective sampling for sparse signal recovery*, in Proceedings of the 12th International Conference on Artificial Intelligence and Statistics (AISTATS), J. Mach. Learning Res., 2009, pp. 216–223.

[21] G. Herman, *Image Reconstruction from Projections, The Fundamentals of Computerized Tomography*, New York Academic Press, New York, 1980.

[22] W. Hoeffding, *Probability inequalities for sums of bounded random variables*, J. Amer. Statist. Assoc., 58 (1963), pp. 713–721.

[23] D. Huang, E. A. Swanson, C. P. Lin, J. S. Schuman, W. G. Stinson, W. Chang, M. R. Hee, T. Flotte, K. Gregory, C. A. Puliafito, et al., *Optical coherence tomography*, Science, 254 (1991), pp. 1178–1181.

[24] W. Jang and M. Hendry, *Cluster analysis of massive datasets in astronomy*, Statist. Comput., 17 (2007), pp. 253–262.
[25] J. Kaipio and E. Somersalo, *Statistical and Computational Inverse Problems*, Springer, New York, 2005.
[26] K. Krishnamurthy, W. U. Bajwa, R. Willett, and R. Calderbank, *Fast level set estimation from projection measurements*, in Proceedings of the IEEE Statistical Signal Processing Workshop (SSP), IEEE Press, Piscataway, NJ, 2011, pp. 585–588.
[27] A. Kyrieleis, V. Titarenko, M. Ibison, T. Connolley, and P. Withers, *Region-of-interest tomography using filtered backprojection: Assessing the practical limits*, J. Microscopy, 241 (2010), pp. 69–82.
[28] B. Laurent and P. Massart, *Adaptive estimation of a quadratic functional by model selection*, Ann. Statist., 28 (2000), pp. 1302–1338.
[29] R. M Lewitt, *Processing of incomplete measurement data in computed tomography*, Med. Phys., 6 (1979), pp. 412–417.
[30] C. Li, C. Xu, K. Konwar, and M. Fox, *Fast distance preserving level set evolution for medical image segmentation*, in Proceedings of the 9th International Conference on Control, Automation, Robotics and Vision, IEEE Press, Piscataway, NJ, 2006, pp. 1–7.
[31] Y. M. Lu and M. N. Do, *A theory for sampling signals from a union of subspaces*, IEEE Trans. Signal Process., 56 (2008), pp. 2334–2345.
[32] J. Ma, *Iterative region of interest reconstruction in emission tomography*, in Proceedings of the IEEE International Symposium on Biomedical Imaging: From Nano to Macro, 2010, pp. 604 –607.
[33] C. Maass, M. Knaup, and M. Kachelriess, *New approaches to region of interest computed tomography*, Med. Phys., 38 (2011), pp. 2868–2878.
[34] R. Marques, F. De Medeiros, and D. Ushizima, *Target detection in SAR images based on a level set approach*, IEEE Trans. Systems Man Cybernet. Part C Appl. Rev., 39 (2009), pp. 214–222.
[35] D. Mason and W. Polonik, *Asymptotic normality of plug-in level set estimates*, Ann. Appl. Probab., 19 (2009), pp. 1108–1142.
[36] R. Puetter, T. Gosnell, and A. Yahil, *Digital image reconstruction: Deblurring and denoising*, Annu. Rev. Astron. Astrophys., 43 (2005), pp. 139–194.
[37] P. Rigollet and R. Vert, *Fast rates for plug-in estimators of density level sets*, Bernoulli, 15 (2009), pp. 1154–1178.
[38] P. Rosen, S. Hensley, I. Joughin, F. Li, S. Madsen, E. Rodriguez, and R. Goldstein, *Synthetic aperture radar interferometry*, Proc. IEEE, 88 (2000), pp. 333–382.
[39] M. Rudelson, *Lecture Notes on Non-asymptotic Theory of Random Matrices*, arXiv preprint arXiv:1301.2382, 2013.
[40] L. I. Rudin, S. Osher, and E. Fatemi, *Nonlinear total variation based noise removal algorithms*, Phys. D., 60 (1992), pp. 259–268.
[41] C. Scott and M. Davenport, *Regression level set estimation via cost-sensitive classification*, IEEE Trans. Signal Process., 55 (2007), pp. 2752–2757.
[42] C. Scott and R. Nowak, *Learning minimum volume sets*, J. Machine Learning Res., 7 (2006), pp. 665–704.
[43] C. Scott and R.D. Nowak, *Minimax-optimal classification with dyadic decision trees*, IEEE Trans. Inform. Theory, 52 (2006), pp. 1335–1353.
[44] A. Singh, C. Scott, and R. Nowak, *Adaptive Hausdorff estimation of density level sets*, Ann. Statist., 37 (2009), pp. 2760–2782.
[45] D. Takhar, J. Laska, M. Wakin, M. Duarte, D. Baron, S. Sarvotham, K. Kelly, and R. Baraniuk, *A new compressive imaging camera architecture using optical-domain compression*, in Electronic Imaging 2006, Int. Soc. Opt. Phot., 2006, 606509.
[46] R. Tibshirani, *Regression shrinkage and selection via the lasso*, J. Roy. Statist. Soc. Ser. B, 58 (1996), pp. 267–288.
[47] J. A. Tropp and A. C. Gilbert, *Signal recovery from random measurements via orthogonal matching pursuit*, IEEE Trans. Inform. Theory, 53 (2007), pp. 4655–4666.
[48] A. Tsybakov, *On nonparametric estimation of density level sets*, Ann. Statist., 25 (1997), pp. 948–969.
[49] S. Vaiter, G. Peyré, C. Dossal, and J. Fadili, *Robust sparse analysis regularization*, IEEE Trans. Inform. Theory, 59 (2013), pp. 2001–2016.

[50] V. Vapnik, *The Nature of Statistical Learning Theory*, Springer-Verlag, Berlin, 2000.
[51] R. Vershynin, *Norm of a Random Matrix*, lecture notes on nonasymptotic theory of random matrices, Department of Mathematics, University of Michigan, Ann Arbor, MI, 2006–2007; available online at http://www-personal.umich.edu/∼romanv/teaching/2006-07/280/lec6.pdf.
[52] Y. Wang, J. Yang, W. Yin, and Y. Zhang, *A new alternating minimization algorithm for total variation image reconstruction*, SIAM J. Imaging Sci., 1 (2008), pp. 248–272.
[53] R. Willett and R. Nowak, *Minimax optimal level set estimation*, IEEE Trans. Image Process., 16 (2007), pp. 2965–2979.
[54] Y. Yang, *Minimax nonparametric classification—Part* I: *Rates of convergence*, IEEE Trans. Inform. Theory, 45 (1979), pp. 2271–2284.