Learning Minimum Volume Sets

Clayton Scott* and Robert Nowak†

UW-Madison Technical Report ECE-05-2
cscott@rice.edu, nowak@engr.wisc.edu

June 2005

Abstract

Given a probability measure P and a reference measure μ, one is often interested in the minimum μ-measure set with P-measure at least α. Minimum volume sets of this type summarize the regions of greatest probability mass of P, and are useful for detecting anomalies and constructing confidence regions. This paper addresses the problem of estimating minimum volume sets based on independent samples distributed according to P. Other than these samples, no other information is available regarding P, but the reference measure μ is assumed to be known. We introduce rules for estimating minimum volume sets that parallel the empirical risk minimization and structural risk minimization principles in classification. As in classification, we show that the performances of our estimators are controlled by the rate of uniform convergence of empirical to true probabilities over the class from which the estimator is drawn. Thus we obtain finite sample size performance bounds in terms of VC dimension and related quantities. We also demonstrate strong universal consistency, an oracle inequality, and rates of convergence. The proposed estimators are illustrated with histogram and decision tree set estimation rules.

1 Introduction

Given a probability measure P and a reference measure μ, the minimum volume set (MV-set) with mass at least $0 < \alpha < 1$ is

$$G_\alpha^* = \arg\min\{\mu(G) : P(G) \geq \alpha, G \text{ measurable}\}.$$

MV-sets summarize regions where the mass of P is most concentrated. For example, if P is a multivariate Gaussian distribution and μ is the Lebesgue measure, then the MV-sets are ellipsoids. An MV-set for a two-component Gaussian mixture is illustrated in Figure 1. Applications of minimum volume sets include outlier/anomaly detection, determining highest posterior density or multivariate confidence regions, tests for multimodality, and clustering. See Polonik [1997], Walther [1997], Schölkopf et al. [2001] and references therein for additional applications.

*Department of Statistics, Rice University, 6100 Main St, MS-138, Houston, TX 77005. Supported by an NSF VIGRE postdoctoral training grant.

†Department of Electrical and Computer Engineering, University of Wisconsin-Madison, 1415 Engineering Dr, Madison, WI 53706

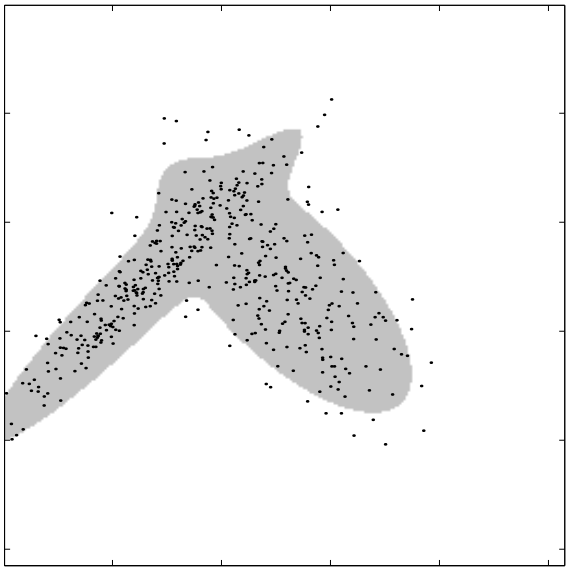

Figure 1: Minimum volume set (gray region) of a two-component Gaussian mixture. Also shown are 500 points drawn independently from this distribution.

This paper considers the problem of MV-set estimation using a training sample drawn from P, which in most practical settings is the only information one has about P. The specifications to the estimation process are the significance level α, the reference measure μ, and a collection of candidate sets $\mathcal{G}$.

A major theme of this work is the strong parallel between MV-set estimation and binary classification. In particular, we find that uniform convergence (of true probability to empirical probability over the class of sets $\mathcal{G}$) plays a central role in controlling the performance of MV-set estimators. Thus, we derive distribution free finite sample performance bounds in terms of familiar quantities such as VC dimension. In fact, as we will see, any uniform convergence bound can be directly converted to a rule for MV-set estimation.

In Section 2 we introduce a rule for MV-set estimation analogous to empirical risk minimization in classification, and shows that this rule obeys similar finite sample size performance guarantees. Section 3 extends the results of the previous section to allow $\mathcal{G}$ to grow in a controlled way with sample size, leading to MV-set estimators that are strongly universally consistent. Section 4 introduces an MV-set estimation rule similar in spirit to structural risk minimization in classification, and develops an oracle-type inequality for this estimator. The oracle inequality guarantees that the estimator automatically adapts its complexity to the problem at hand. Section 5 introduces a tuning parameter to the proposed rules that allows the user to affect the tradeoff between volume error and mass error without sacrificing theoretical properties. Section 6 provides a "case study" of tree-structured set estimators to illustrate the power of the oracle inequality for deriving rates of convergence. Section 7 includes a set of numerical experiments that explores the proposed theory (and algorithmic issues) using histogram and decision tree rules in two dimensions. Section 8 includes concluding remarks and avenues for potential future investigations. Detailed proofs of the main results of the paper are relegated to the appendices. Throughout the paper, the theoretical results are illustrated in detail through several examples, including VC classes, histograms, and

decision trees.

1.1 Previous work

All previous theoretical work on data-based MV-set estimation has been asymptotic in nature, to our knowledge. Thus, ours are the first known finite sample bounds. Polonik [1997] proves consistency and rates of convergence for an estimator based on the so-called generalized quantile function (discussed in more detail in Section 2.4). His consistency results place restrictions on the MV-set G^*_α (e.g, $\mu(G^*_\alpha)$ is continuous in α), whereas our consistency result holds universally, i.e., for all distributions P. Walther [1997] studies an approach based on "granulometric smoothing," which involves applying certain morphological smoothing operations to the α-mass level set of a kernel density estimate. His rates, like those of Polonik, apply under smoothness assumptions on the density. In contrast, our rate of convergence results in Section 6 depend on the smoothness of the boundary of G^*_α.

Algorithms for MV-set estimation have been developed for convex sets [Sager, 1979] and ellipsoidal sets [Hartigan, 1987] in two dimensions. Unfortunately, for more complicated problems (dimension > 2 and non-convex sets), there has been a disparity between practical MV-set estimators and theoretical results. Polonik [1997] makes no comment on the practicality of his estimators. The smoothing estimators of Walther [1997] in practice must approximate the theoretical estimator via iterative level set estimation. On the other hand, computationally efficient procedures like those in Schölkopf et al. [2001] and Huo and Lu [2004] are motivated by the minimum volume set paradigm, but their performance relative to G^*_α is not known. Our proposed algorithms for histograms and decision trees are practical in low dimensional settings, but appear to be constrained by the same computational limitations as empirical risk minimization in binary classification.

More broadly, MV-set estimation theory has similarities (in terms of the nature of results and technical devices) to other set estimation problems, such as classification, discrimination analysis, density support estimation (which corresponds to the case $\alpha = 1$), and density level set estimation, to which we now turn.

1.2 Connection to density level sets

The MV-set estimation problem is closely related to density level set estimation [Tsybakov, 1997, Ben-David and Lindenbaum, 1997, Cuevas and Rodriguez-Casal, 2003, Steinwart et al., 2005, Vert and Vert, 2005] and excess mass estimation problems [Müller and Sawitzki, 1991, Polonik, 1995]. Indeed, it is well known that density level sets are minimum volume sets [Nunez-Garcia et al., 2003].

The main difference between density level sets and MV-sets is that the former require the specification of a density level of interest, rather than the specification of the mass α to be enclosed. Since the density is in general unknown, it seems that specifying α is much more reasonable and intuitive than setting a density level for problems like anomaly detection. Suppose for example that one is interested in a reference measure of the form $c\mu$, where μ is Lebesgue measure and $c > 0$. The choice of c does not change the minimum volume set, but it does affect the γ level set. Since there is no way a priori to choose the best c, the invariance of the minimum volume set seems highly desirable. Further remarks along these lines are given in the concluding section.

The connections between MV-sets and density level sets will be important later in this paper. To make the connection precise the following assumption on the data-generating distribution and

reference measure is needed. We emphasize that this assumption is not necessary for the results in Sections 2 and 3, where we demonstrate distribution free error bounds and universal consistency.

A1 P has a density f with respect to μ.

A key result relating density level and MV-sets is the following, stated without proof (see, e.g., Nunez-Garcia et al. [2003]).

Lemma 1. *Under assumption* **A1** *there exists* γ_α *such that for any MV-set* G^*_α,

$$\{x : f(x) > \gamma_\alpha\} \subset G^*_\alpha \subset \{x : f(x) \geq \gamma_\alpha\}.$$

Note that every density level set is an MV-set, but not conversely. If, however, $\mu(\{x : f(x) = \gamma_\alpha\}) = 0$, then the three sets in the Lemma coincide.

1.3 Notation

Let $(\mathcal{X}, \mathcal{B})$ be a measure space with $\mathcal{X} \subset \mathbb{R}^d$. Let X be a random variable taking values in $\mathcal{X}$ with distribution P. Let $S = (X_1, \ldots, X_n)$ be an independent and identically distributed (IID) sample drawn according to P. Let G denote a subset of $\mathcal{X}$, and let $\mathcal{G}$ be a collection of such subsets. Let $\widehat{P}$ denote the empirical measure based on S

$$\widehat{P}(G) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}\left(X_i \in G\right).$$

Here $\mathbb{I}(\cdot)$ is the indicator function. The notation μ will denote a measure[1] on $\mathcal{X}$. Denote by f the density of P with respect to μ (when it exists), $\gamma > 0$ a level of the density, and $\alpha \in (0, 1)$ a user-specified mass constraint. Define

$$\mu^*_\alpha = \inf_G \{\mu(G) : P(G) \geq \alpha\}, \tag{1}$$

where the inf is over all measurable sets. A minimum volume set, G^*_α, is a minimizer of (1) when it exists. A partial list of notations used in the paper is given in Fig. 2.

2 Minimum Volume Sets and Empirical Risk Minimization

We introduce a procedure inspired by the empirical risk minimization (ERM) principle for classification. In classification, ERM selects a classifier from a fixed set of classifiers by minimizing the empirical error (risk) of a training sample. Vapnik and Chervonenkis established the basic theoretical properties of ERM [see Vapnik, 1998, Devroye et al., 1996], and we find similar properties in the minimum volume setting.

In this and the next section our assumptions are quite general. Thus, let $(\mathcal{X}, \mathcal{B})$ be a measure space, and μ a measure on $\mathcal{X}$. We do not assume P has a density with respect to μ.

Let $\mathcal{G}$ be a class of sets. Given $\alpha \in (0, 1)$, denote

$$\mathcal{G}_\alpha \quad = \quad \{G \in \mathcal{G} : P(G) \geq \alpha\},$$

[1] Although we do not emphasize it, the results of Sections 2 and 3 only require μ to be a real-valued function.

symbol	meaning
P	data generating distribution
S	a training sample
$\widehat{P}$	empirical version of P
μ	volume/reference measure
α	mass constraint $\in (0,1)$
G	a set
G^*_α	a minimum volume set
μ^*_α	volume of G^*_α
$\mathcal{G}$	a class of sets
$G_{\mathcal{G},\alpha}$	best approximation to G^*_α from $\mathcal{G}$
$\mu_{\mathcal{G},\alpha}$	volume of $G_{\mathcal{G},\alpha}$
δ	a confidence parameter
ϕ	a complexity penalty
$\mathcal{G}_\alpha$	$\{G \in \mathcal{G} : P(G) \geq \alpha\}$
$\widehat{\mathcal{G}}_\alpha$	$\{G \in \mathcal{G} : \widehat{P}(G) \geq \alpha - \phi(G,S,\delta)\}$
$\widehat{G}_{\mathcal{G},\alpha}$	a minimum volume set estimate
f	the density of P with respect to μ, if it exists
γ	a density level
γ_α	the density level corresponding to mass α

Figure 2: Notations. A " $\widehat{\ }$ " denotes a data-dependent quantity, while a " $*$ " indicates a quantity that is optimal with respect to all measurable subsets of $\mathcal{X}$.

the collection of all sets in $\mathcal{G}$ with mass at least α. Define

$$\mu_{\mathcal{G},\alpha} = \inf\{\mu(G) : G \in \mathcal{G}_\alpha\} \tag{2}$$

and

$$G_{\mathcal{G},\alpha} = \arg\min\{\mu(G) : G \in \mathcal{G}_\alpha\} \tag{3}$$

when it exists. Thus $G_{\mathcal{G},\alpha}$ is the best approximation to the minimum volume set G^*_α from $\mathcal{G}$.

Empirical versions of $\mathcal{G}_\alpha$ and $G_{\mathcal{G},\alpha}$ are defined as follows. Let $\phi(G, S, \delta)$ be a function of $G \in \mathcal{G}$, the training sample S, and a confidence parameter $\delta \in (0, 1)$. Set

$$\widehat{\mathcal{G}}_\alpha = \{G \in \mathcal{G} : \widehat{P}(G) \geq \alpha - \phi(G, S, \delta)\}$$

and

$$\widehat{G}_{\mathcal{G},\alpha} = \arg\min\{\mu(G) : G \in \widehat{\mathcal{G}}_\alpha\}. \tag{4}$$

We refer to the rule in (4) as MV-ERM because of the analogy with empirical risk minimization in classification. A discussion of the existence and uniqueness of the above quantities is deferred to Section 2.5.

The quantity ϕ acts as a kind of "tolerance" by which the empirical mass may deviate from the targeted value α. Throughout this paper we assume that ϕ satisfies the following.

Definition 1. *We say ϕ is a (distribution free)* complexity penalty *for $\mathcal{G}$ if and only if for all distributions P and all $\delta \in (0, 1)$,*

$$P^n\left(\left\{S : \sup_{G\in\mathcal{G}}\left(\left|P(G) - \widehat{P}(G)\right| - \phi(G, S, \delta)\right) > 0\right\}\right) \leq \delta.$$

Thus, ϕ controls the rate of uniform convergence of $\widehat{P}(G)$ to $P(G)$ for $G \in \mathcal{G}$. It is well known that the performance of ERM (for binary classification) relative to the performance of the best classifier in the given class is controlled by the uniform convergence of true to empirical probabilities. A similar result holds for MV-ERM.

Theorem 1. *If ϕ is a complexity penalty for $\mathcal{G}$, then*

$$P^n\left(\left(P(\widehat{G}_{\mathcal{G},\alpha}) < \alpha - 2\phi(\widehat{G}_{\mathcal{G},\alpha}, S, \delta)\right) \text{ or } \left(\mu(\widehat{G}_{\mathcal{G},\alpha}) > \mu_{\mathcal{G},\alpha}\right)\right) \leq \delta.$$

Proof. Consider the sets

$$\begin{aligned}
\Theta_P &= \{S : P(\widehat{G}_{\mathcal{G},\alpha}) < \alpha - 2\phi(\widehat{G}_{\mathcal{G},\alpha}, S, \delta)\},\\
\Theta_\mu &= \{S : \mu(\widehat{G}_{\mathcal{G},\alpha}) > \mu(G_{\mathcal{G},\alpha})\},\\
\Omega_P &= \left\{S : \sup_{G\in\mathcal{G}}\left(\left|P(G) - \widehat{P}(G)\right| - \tfrac{1}{2}\phi(G, S, \delta)\right) > 0\right\}.
\end{aligned}$$

Lemma 2. *With Θ_P, Θ_μ, and Ω_P as defined above we have*

$$\Theta_P \cup \Theta_\mu \subset \Omega_P.$$

The proof is given in Appendix A, and follows closely the proof of Lemma 1 in Cannon et al. [2002]. The result now follows easily. □

Lemma 2 may be understood by analogy with the result from classification that says $\mathcal{R}(\widehat{f}) - \inf_{f\in\mathcal{F}} \mathcal{R}(f) \leq 2\sup_{f\in\mathcal{F}} |\mathcal{R}(f) - \widehat{\mathcal{R}}(f)|$ (see Devroye et al. [1996], Ch. 8). Here $\mathcal{R}$ and $\widehat{\mathcal{R}}$ are the true and empirical risks, $\widehat{f}$ is the empirical risk minimizer, and $\mathcal{F}$ is a set of classifiers. Just as this result relates uniform convergence to empirical risk minimization in classification, so does Lemma 2 relate uniform convergence to the performance of MV-ERM.

The theorem above allows direct translation of uniform convergence results into performance guarantees on MV-ERM. Fortunately, many penalties (uniform convergence results) are known. In the next two subsections we take a closer look at penalties for VC classes and countable classes, and a Rademacher penalty.

2.1 Example: VC Classes

Let $\mathcal{G}$ be a class of sets with VC dimension V, and define

$$\phi(G, S, \delta) = \sqrt{32\frac{V\log n + \log(8/\delta)}{n}}. \tag{5}$$

By a version of the VC inequality [Devroye et al., 1996], we know that ϕ is a complexity penalty for $\mathcal{G}$, and therefore Theorem 1 applies.

To view this result in perhaps a more recognizable way, let $\epsilon > 0$ and choose δ such that $2\phi(G, S, \delta) = \epsilon$ for all $G \in \mathcal{G}$ and all S. By inverting the relationship between δ and ϵ, we have the following.

Corollary 1. *With the notation defined above,*

$$P^n\left(\left(P(\widehat{G}_{\mathcal{G},\alpha}) < \alpha - \epsilon\right) \text{ or } \left(\mu(\widehat{G}_{\mathcal{G},\alpha}) > \mu_{\mathcal{G},\alpha}\right)\right) \leq 8n^V e^{-n\epsilon^2/128}.$$

Thus, for any fixed $\epsilon > 0$, the probability of being within ϵ of the target mass α and being less than the target volume $\mu_{\mathcal{G},\alpha}$ approaches one exponentially fast as the sample size increases. This result may also be used to calculate a distribution free upper bound on the sample size needed to be within a given tolerance ϵ of α and with a given confidence $1 - \delta$. In particular, the sample size will grow no faster than a polynomial in $1/\epsilon$ and $1/\delta$, paralleling results for classification.

2.2 Example: Countable Classes

Suppose $\mathcal{G}$ is a countable class of sets. Assume that to every $G \in \mathcal{G}$ a number $[\![G]\!]$ is assigned such that

$$\sum_{G\in\mathcal{G}} 2^{-[\![G]\!]} \leq 1. \tag{6}$$

In light of the Kraft inequality for prefix[2] codes [Cover and Thomas, 1991], $[\![G]\!]$ may be defined as the codelength of a codeword for G in a prefix code for $\mathcal{G}$. Let $\delta > 0$ and define

$$\phi(G, S, \delta) = \sqrt{\frac{[\![G]\!]\log 2 + \log(2/\delta)}{2n}}. \tag{7}$$

[2] A prefix code is a collection of codewords (strings of 0s and 1s) such that no codeword is a prefix of another.

By Chernoff's bound together with the union bound, ϕ is a penalty for $\mathcal{G}$. Therefore Theorem 1 applies and we have a result analogous to the Occam's Razor bound for classification [see Langford, 2005].

As a special case, suppose $\mathcal{G}$ is finite and take $[\![G]\!] = \log_2 |\mathcal{G}|$. Setting $2\phi(G, S, \delta) = \epsilon$ and inverting the relationship between δ and ϵ, we have the following.

Corollary 2. *For the MV-ERM estimate $\widehat{G}_{\mathcal{G},\alpha}$ from a finite class $\mathcal{G}$*

$$P^n\left(\left(P(\widehat{G}_{\mathcal{G},\alpha}) < \alpha - \epsilon\right) or \left(\mu(\widehat{G}_{\mathcal{G},\alpha}) > \mu_{\mathcal{G},\alpha}\right)\right) \leq 2|\mathcal{G}|e^{-n\epsilon^2/2}.$$

As with VC classes, these inequalities may be used for sample size calculations.

2.3 The Rademacher Penalty for Sets

The Rademacher penalty was originally studied in the context of classification by Koltchinskii [2001] and Bartlett et al. [2002]. For a succinct exposition of its basic properties, see Bousquet et al. [2004]. An analogous penalty exists for sets. Let $\sigma_1, \ldots, \sigma_n$ be Rademacher random variables, i.e., independent random variables taking on the values 1 and -1 with equal probability. Denote $\widehat{P}_{(\sigma_i)}(G) = \frac{1}{n}\sum_{i=1}^n \sigma_i \mathbb{I}(X_i \in G)$. We define the Rademacher average

$$\rho(\mathcal{G}) = \mathbf{E}\left[\sup_{G\in\mathcal{G}} \widehat{P}_{(\sigma_i)}(G)\right]$$

and the conditional Rademacher average

$$\widehat{\rho}(\mathcal{G}, S) = \mathbf{E}_{(\sigma_i)}\left[\sup_{G\in\mathcal{G}} \widehat{P}_{(\sigma_i)}(G)\right],$$

where the second expectation is with respect the Rademacher random variables only, and conditioned on the sample S.

Proposition 1. *With probability at least $1-\delta$ over the draw of S,*

$$P(G) - \widehat{P}(G) \leq 2\rho(\mathcal{G}) + \sqrt{\frac{\log(1/\delta)}{2n}}$$

for all $G \in \mathcal{G}$. With probability at least $1-\delta$ over the draw of S,

$$P(G) - \widehat{P}(G) \leq 2\widehat{\rho}(\mathcal{G}, S) + \sqrt{\frac{2\log(2/\delta)}{n}}$$

for all $G \in \mathcal{G}$.

The proof of this result follows exactly the same lines as the proof of Theorem 5 in Bousquet et al. [2004], and is omitted.

Assume $\mathcal{G}$ satisfies the property that $G \in \mathcal{G} \Rightarrow \overline{G} \in \mathcal{G}$, where $\overline{G}$ denotes the compliment of G. Then $\widehat{P}(G) - P(G) = P(\overline{G}) - \widehat{P}(\overline{G})$, and so the upper bounds of Proposition 1 also apply to $|P(G) - \widehat{P}(G)|$. Thus we are able to define the conditional Rademacher penalty

$$\phi(G, S, \delta) = 2\widehat{\rho}(\mathcal{G}, S) + \sqrt{\frac{2\log(2/\delta)}{n}}.$$

By the above Proposition, this is a complexity penalty according to Definition 1. The conditional Rademacher penalty is studied further in Section 7 and in Appendix F, where it is shown that $\widehat{\rho}(\mathcal{G}, S)$ can be computed efficiently for sets based on a fixed partition of $\mathcal{X}$ (such as histograms and trees).

2.4 Comparison to Generalized Quantile Processes

Polonik [1997] studies the *empirical quantile function*

$$\widehat{V}_\alpha = \inf\{\mu(G) : \widehat{P}(G) \geq \alpha\},$$

and the MV-set estimate that achieves the minimum (when it exists). The only difference compared with MV-ERM is the absence of the term $\phi(G, S, \delta)$ in the constraint. Thus, MV-ERM will tend to produce estimates with smaller volume and smaller mass. While Polonik proves only asymptotic properties of his estimator, we have demonstrated finite sample bounds for MV-ERM. Moreover, in Section 5, we show that the results of this section extend to a generalization of MV-ERM where ϕ is replaced by $\nu\phi$, where ν is any number $-1 \leq \nu \leq 1$. Thus finite sample bounds also exist for Polonik's estimator ($\nu = 0$).

2.5 Existence and Uniqueness

In this section we discuss the existence and uniqueness of the sets $G_{\mathcal{G},\alpha}$ and $\widehat{G}_{\mathcal{G},\alpha}$. Regarding the former, it is really not necessary that a minimizer exist. All of our results are stated in terms of $\mu_{\mathcal{G},\alpha}$, which certainly exists. When a minimizer exists, its uniqueness is not an issue for the same reason. Our results above involve only $\mu_{\mathcal{G},\alpha}$, which is the same regardless of which minimizer is chosen. Yet one may wonder whether convergence of the volume and mass to their optimal values implies convergence to the MV-set (when it is unique) in any sense. A result in this direction is presented in Theorem 3 below.

For the MV-ERM estimate $\widehat{G}_{\mathcal{G},\alpha}$, uniqueness is again not an issue because all results hold even if the minimizer is chosen arbitrarily. As for existence, we must be more careful. We cannot make the same argument as for $G_{\mathcal{G},\alpha}$ because we are ultimately interested in a concrete set estimate, not just its volume and mass. Clearly, if $\mathcal{G}$ is finite, $\widehat{G}_{\mathcal{G},\alpha}$ exists. For more general sets, existence must be examined on a case-by-case basis. For example, if $\mathcal{X} \subset \mathbb{R}^d$, μ is the Lebesgue measure, and $\mathcal{G}$ is the VC class of spherical or ellipsoidal sets, then $\widehat{G}_{\mathcal{G},\alpha}$ can be seen to exist.

In the event that $\widehat{G}_{\mathcal{G},\alpha}$ does not exist, it suffices to let $\widehat{G}_{\mathcal{G},\alpha}$ be a set whose volume comes within ϵ of the infimum, where ϵ is arbitrarily small. Then our results still hold with $\mu(\widehat{G}_{\mathcal{G},\alpha})$ replaced by $\mu(\widehat{G}_{\mathcal{G},\alpha}) - \epsilon$. The consistency and rate of convergence results below are unchanged, as we may take $\epsilon \to 0$ arbitrarily fast as a function of n.

3 Consistency

A minimum volume set estimator is consistent if its volume and mass tend to the optimal values μ^*_α and α as $n \to \infty$. Formally, define the error quantity

$$\mathcal{M}(G) := (\mu(G) - \mu^*_\alpha)_+ + (\alpha - P(G))_+ ,$$

where $(x)_+ = \max(x, 0)$. We are interested in MV-set estimators such that $\mathcal{M}(\widehat{G}_{\mathcal{G},\alpha})$ tends to zero as $n \to \infty$.

Definition 2. *A learning rule $\widehat{G}_{\mathcal{G},\alpha}$ is* strongly consistent *if*

$$\lim_{n\to\infty} \mathcal{M}(\widehat{G}_{\mathcal{G},\alpha}) = 0 \quad \textit{with probability 1.}$$

If $\widehat{G}_{\mathcal{G},\alpha}$ is strongly consistent for every possible distribution of X, then $\widehat{G}_{\mathcal{G},\alpha}$ is strongly universally *consistent.*

In this section we show that if the approximating power of $\mathcal{G}$ increases in a certain way as a function of n, then MV-ERM leads to a universally consistent learning rule.

To see how consistency might result from MV-ERM, it helps to rewrite Theorem 1 as follows. Let $\mathcal{G}$ be fixed and let $\phi(G, S, \delta)$ be a penalty for $\mathcal{G}$. Then with probability at least $1 - \delta$, both

$$\mu(\widehat{G}_{\mathcal{G},\alpha}) - \mu^*_\alpha \leq \mu(G_{\mathcal{G},\alpha}) - \mu^*_\alpha \tag{8}$$

and

$$\alpha - P(\widehat{G}_{\mathcal{G},\alpha}) \leq 2\phi(\widehat{G}_{\mathcal{G},\alpha}, S, \delta) \tag{9}$$

hold. We refer to the left-hand side of (8) as the *excess volume* of the class $\mathcal{G}$ and the left-hand side of (9) as the *missing mass* of $\widehat{G}_{\mathcal{G},\alpha}$. The upper bounds on the right-hand sides are an approximation error and a stochastic error, respectively.

The idea is to let $\mathcal{G}$ grow with n so that both errors tend to zero as $n \to \infty$. If $\mathcal{G}$ does not change with n, universal consistency is impossible. Either the approximation error will be nonzero for most distributions (when $\mathcal{G}$ is too small) or the bound on the stochastic error will be too large (otherwise). For example, if a class has universal approximation capabilities, its VC dimension is necessarily infinite [Devroye et al., 1996, Ch. 18].

To have both stochastic and approximation errors tend to zero, we apply MV-ERM to a class $\mathcal{G}^k$ from a sequence of classes $\mathcal{G}^1, \mathcal{G}^2, \ldots$, where $k = k(n)$ grows with the sample size. Given such a sequence, define

$$\widehat{G}_{\mathcal{G}^k,\alpha} \quad = \quad \arg\min\{\mu(G) : G \in \widehat{\mathcal{G}}^k_\alpha\}, \tag{10}$$

where

$$\widehat{\mathcal{G}}^k_\alpha = \{G \in \mathcal{G}^k : \widehat{P}(G) \geq \alpha - \phi_k(G, S, \delta)\}$$

and ϕ_k is a penalty for $\mathcal{G}^k$.

Theorem 2. *Choose $k = k(n)$ and $\delta = \delta(n)$ such that*

1. *$k(n) \to \infty$ as $n \to \infty$*
2. *$\sum_{n=1}^{\infty} \delta(n) < \infty$*

Assume the sequence of sets $\mathcal{G}^k$ and penalties ϕ_k satisfy

$$\lim_{k\to\infty} \inf_{G \in \mathcal{G}^k_\alpha} \mu(G) = \mu^*_\alpha \tag{11}$$

and

$$\lim_{n\to\infty} \sup_{G \in \mathcal{G}^k_\alpha} \phi_k(G, S, \delta(n)) = o(1). \tag{12}$$

Then $\widehat{G}_{\mathcal{G}^k,\alpha}$ is strongly universally consistent.

The proof is given in Appendix B. We now give some examples that satisfy these conditions.

3.1 Example: Hierarchy of VC Classes

Assume $\mathcal{G}^1, \mathcal{G}^2, \ldots,$ is a family of VC classes with VC dimensions $V_1 < V_2 < \ldots$. For $G \in \mathcal{G}^k$ define

$$\phi_k(G, S, \delta) = \sqrt{32 \frac{V_k \log n + \log(8/\delta)}{n}}. \tag{13}$$

By taking $\delta(n) \asymp n^{-\beta}$ for some $\beta > 1$ and k such that $V_k = o(n/\log n)$ the assumption in (12) is satisfied. Examples of families of VC classes satisfying (11) include generalized linear discriminant rules with appropriately chosen basis functions and neural networks [Lugosi and Zeger, 1995].

3.2 Example: Histograms

Assume $\mathcal{X} = [0,1]^d$, and let $\mathcal{G}^k$ be the class of all sets formed by taking unions of cells in a regular partition of $\mathcal{X}$ into hypercubes of sidelength $1/k$. Each $\mathcal{G}^k$ has 2^{k^d} members and we may therefore apply the penalty for finite sets discussed in Section 2.2. To satisfy the Kraft inequality (6) it suffices to take $[\![G]\!] = k^d$. The penalty for $G \in \mathcal{G}^k$ is then

$$\phi_k(G, S, \delta) = \sqrt{2 \frac{k^d \log 2 + \log(2/\delta)}{2n}}. \tag{14}$$

By taking $\delta(n) \asymp n^{-\beta}$ for some $\beta > 1$ and k such that $k^d = o(n)$ the assumption in (12) is satisfied. The assumption in (11) is satisfied by the well-known universal approximation capabilities of histograms. Thus the conditions for consistency of histograms for minimum volume set estimation are exactly parallel to the conditions for consistency of histogram rules for classification [Devroye et al., 1996, Ch. 9]. Dyadic decision trees, discussed below in Section 6, are another countable family for which consistency results are possible.

3.3 The Symmetric Difference Performance Metric

An alternative measure of performance for an MV-set estimator is the μ-measure of the symmetric difference, $\mu(\widehat{G}_{\mathcal{G},\alpha} \Delta G^*_\alpha)$, where $A \Delta B = (A \backslash B) \cup (B \backslash A)$. Although this performance metric has been commonly adopted in the study of density level sets, it is less desirable for our purposes. First, unlike with density level sets, there may not be a unique MV-set (imagine the case where the density of P has a plateau). Second, as pointed out by Steinwart et al. [2005], there is no known way to estimate the accuracy of this measure using only samples from P. Nonetheless, the symmetric difference metric coincides asymptotically with our error metric $\mathcal{M}$ in the sense of the following result. The theorem uses the notation γ_α to denote the density level corresponding to the MV-set, as discussed in Section 1.2.

Theorem 3. *Let G_n denote a sequence of sets. If G^*_α is a minimum volume set and $\mu(G_n \Delta G^*_\alpha) \to 0$ with n, then $\mathcal{M}(G_n) \to 0$. Conversely, assume μ is a probability measure, P has a bounded density f with respect to μ, and $\mu(\{x : f(x) = \gamma_\alpha\}) = 0$. If $\mathcal{M}(G_n) \to 0$, then $\mu(G_n \Delta G^*_\alpha) \to 0$.*

The proof is given in Appendix C. The assumptions of the second part of the theorem ensure that G^*_α is unique, otherwise the converse statement need not be true. The proof of the converse reveals yet another connection between MV-set estimation and classification. In particular, we show that $\mathcal{M}(G_n)$ bounds the excess classification risk for a certain classification problem. The converse statement then follows from a result of Steinwart et al. [2005] who show that this excess classification risk and the μ-measure of the symmetric difference tend to zero simultaneously.

4 Structural Risk Minimization and an Oracle Inequality

In the previous section on consistency the rate of convergence of the two errors to zero is determined by the choice of $k = k(n)$, which must be chosen a priori. Hence it is possible that the excess volume decays much more quickly than the missing mass, or vice versa. In this section we introduce a new rule called MV-SRM, inspired by the principle of structural risk minimization (SRM) from the theory of classification [Vapnik, 1982, Lugosi and Zeger, 1996], that automatically balances the two errors.

The results of this and subsequent sections are no longer distribution free. In particular, we assume

A1 P has a density f with respect to μ.

A2 for all $\alpha' \in (0,1)$, $G^*_{\alpha'}$ exists and $P(G^*_{\alpha'}) = \alpha'$.

Note that **A2** holds if f has no plateaus, i.e., $\mu(\{x : f(x) = \gamma\}) = 0$ for all $\gamma > 0$. This is a commonly made assumption in the study of density level sets. However, **A2** is somewhat more general. It still holds, for example, if μ is absolutely continuous with respect to Lebesgue measure, even if f has plateaus.

Recall from Section 1.2 that under assumption **A1**, there exists $\gamma_\alpha > 0$ such for any MV-set G^*_α,

$$\{x : f(x) > \gamma_\alpha\} \subset G^*_\alpha \subset \{x : f(x) \geq \gamma_\alpha\}.$$

Let $\mathcal{G}$ be a class of sets. Intuitively, view $\mathcal{G}$ as a collection of sets of varying capacities, such as a union of VC classes or a union of finite classes (examples are given below). Let $\phi(G, S, \delta)$ be a penalty for $\mathcal{G}$. The MV-SRM principle selects the set

$$\widehat{G}_{\mathcal{G},\alpha} \;=\; \arg\min_{G \in \mathcal{G}} \left\{ \mu(G) + \phi(G, S, \delta) : \widehat{P}(G) \geq \alpha - \phi(G, S, \delta) \right\}. \tag{15}$$

Note that MV-SRM is different from MV-ERM because it minimizes a complexity penalized volume instead of simply the volume. We have the following oracle inequality for MV-SRM.

Theorem 4. *Let $\widehat{G}_{\mathcal{G},\alpha}$ be the MV-set estimator in (15) and assume* **A1** *and* **A2** *hold. With probability at least $1 - \delta$ over the training sample S,*

$$\mathcal{M}(\widehat{G}_{\mathcal{G},\alpha}) \leq \left(1 + \frac{1}{\gamma_\alpha}\right) \inf_{G \in \mathcal{G}_\alpha} \left\{ \; \mu(G) - \mu^*_\alpha + 2\phi(G, S, \delta) \; \right\}. \tag{16}$$

Although the value of $1/\gamma_\alpha$ is in practice unknown, it can be bounded by

$$\frac{1}{\gamma_\alpha} \leq \frac{\mu(\mathcal{X}) - \mu^*_\alpha}{1 - \alpha} \leq \frac{\mu(\mathcal{X})}{1 - \alpha}.$$

This follows from the bound $1 - \alpha \leq \gamma_\alpha \cdot (\mu(\mathcal{X}) - \mu^*_\alpha)$ on the mass outside the minimum volume set. If μ is a probability measure, then $1/\gamma_\alpha \leq 1/(1 - \alpha)$.

The oracle inequality says that MV-SRM performs about as well as the set chosen by an oracle to optimize the tradeoff between excess volume and missing mass.

4.1 Example: Union of VC Classes

Consider $\mathcal{G} = \cup_{k=1}^{K} \mathcal{G}^k$, where $\mathcal{G}^k$ has VC dimension V_k, $V_1 < V_2 < \cdots$, and K is possibly infinite. A penalty for $\mathcal{G}$ can be obtained by defining, for $G \in \mathcal{G}^k$,

$$\phi(G, S, \delta) = \phi_k(G, S, \delta 2^{-k}),$$

where ϕ_k is the penalty from Equation (13). Then ϕ is a penalty for $\mathcal{G}$ because ϕ_k is a penalty for $\mathcal{G}^k$, and by applying the union bound and the fact $\sum_{k \geq 1} 2^{-k} \leq 1$. In this case, MV-SRM adaptively selects an MV-set estimate from a VC class that balances approximation and stochastic errors.

To be more concrete, suppose $\mathcal{G}^k$ is the collection of sets whose boundaries are defined by polynomials of degree k. It may happen that for certain distributions, the MV-set is well-approximated by a quadratic region (such as an ellipse), while for other distributions a higher degree polynomial is required. If the appropriate polynomial degree for the MV-set is not known in advance, as would be the case in practice, then MV-SRM adaptively chooses an estimator of a certain degree that does about as well as if the best degree was known in advance.

4.2 Example: Union of Histograms

Let $\mathcal{G} = \cup_{k=1}^{K} \mathcal{G}^k$, where $\mathcal{G}^k$ is as in Section 3.2. As with VC classes, we obtain a penalty for $\mathcal{G}$ by defining, for $G \in \mathcal{G}^k$,

$$\phi(G, S, \delta) = \phi_k(G, S, \delta 2^{-k}),$$

where ϕ_k is the penalty from Equation (14). Then MV-SRM adaptively chooses a partition resolution k that approximates the MV-set about as well as possible without overfitting the training data. This example is studied experimentally in Section 7.

5 Damping the Penalty

In Theorem 1, the reader may have noticed that MV-ERM does not equitably balance the excess volume ($\mu(\widehat{G}_{\mathcal{G},\alpha})$ relative to its optimal value) with the missing mass ($P(\widehat{G}_{\mathcal{G},\alpha})$ relative to α). Indeed, with high probability, $\mu(\widehat{G}_{\mathcal{G},\alpha})$ is *less than* $\mu(G_{\mathcal{G},\alpha})$, while $P(\widehat{G}_{\mathcal{G},\alpha})$ is only guaranteed to be within $\phi(\widehat{G}_{\mathcal{G},\alpha})$ of α. The net effect is that MV-ERM (and MV-SRM) underestimates the MV-set. Our experiments in Section 7 demonstrate this to be the case.

In this section we introduce variants of MV-ERM and MV-SRM that allow the total error to be shared between the volume and mass, instead of all of the error residing in the mass term. Our approach is to introduce a damping factor $-1 \leq \nu \leq 1$ that scales the penalty. We will see that the resulting MV-set estimators obey performance guarantees like those we have already seen, but with the total error redistributed between the volume and mass. The reason for not introducing this more general framework initially is that the results are slightly less general, more involved to state, and to some extent follow as corollaries to the original ($\nu = 1$) framework.

The extensions of this section encompass the generalized quantile estimate of Polonik [1997], which corresponds to $\nu = 0$. Thus we have finite sample size guarantees for that estimator to match Polonik's asymptotic analysis. The case $\nu = -1$ is also of interest. If it is crucial that the estimate satisfies the mass constraint $P(\widehat{G}_{\mathcal{G},\alpha}) \geq \alpha$ (note that this involves the *true* probability measure P), setting $\nu = -1$ ensures this to be the case with probability at least $1 - \delta$.

First we consider damping the penalty in MV-ERM. Assume that the penalty is independent of $G \in \mathcal{G}$ and of the sample S, although it can depend on n and δ. That is, $\phi(G, S, \delta) = \phi(n, \delta)$. For example, ϕ may be the penalty in (5) for VC classes or (7) for finite classes. Let $\nu \leq 1$ and define

$$\widehat{G}^{\nu}_{\mathcal{G},\alpha} = \arg\min_{G \in \mathcal{G}} \left\{ \mu(G) : \widehat{P}(G) \geq \alpha - \nu\phi(n, \delta) \right\}.$$

Since ϕ is independent of $G \in \mathcal{G}$, $\widehat{G}^{\nu}_{\mathcal{G},\alpha}$ coincides with the MV-ERM estimate (as originally formulated) $\widehat{G}_{\mathcal{G},\alpha'}$ but at the adjusted mass constraint $\alpha' = \alpha + (1-\nu)\phi(n, \delta)$. Therefore, we may apply Theorem 1 to obtain the following.

Corollary 3. *Let $\alpha' = \alpha + (1-\nu)\phi(n, \delta)$. Then*

$$P^n \left(\left(P(\widehat{G}^{\nu}_{\mathcal{G},\alpha}) < \alpha - (1+\nu)\phi(n, \delta) \right) \text{ or } \left(\mu(\widehat{G}_{\mathcal{G},\alpha}) > \mu_{\mathcal{G},\alpha'} \right) \right) \leq \delta.$$

Relative to the original formulation of MV-ERM, the bound on the missing mass is decreased by a factor $(1+\nu)/2$. On the other hand, the volume is now bounded by $\mu_{\mathcal{G},\alpha'} = \mu_{\mathcal{G},\alpha} + (\mu_{\mathcal{G},\alpha'} - \mu_{\mathcal{G},\alpha})$. Thus the bound on the excess volume is increased from 0 to $\mu_{\mathcal{G},\alpha'} - \mu_{\mathcal{G},\alpha}$. This may be interpreted as a stochastic component of the excess volume. Relative to the MV-set, $\mu(\widehat{G}_{\mathcal{G},\alpha})$ has only an approximation error, whereas $\mu(\widehat{G}^{\nu}_{\mathcal{G},\alpha})$ has both approximation and stochastic errors. The advantage is that now the stochastic error of the mass is decreased.

A similar construction applies to MV-SRM. Now assume $\mathcal{G} = \cup_{k=1}^{K} \mathcal{G}^k$. Given a scale parameter ν, define

$$\widehat{G}^{\nu}_{\mathcal{G},\alpha} = \arg\min_{G \in \mathcal{G}} \left\{ \mu(G) + (1+\nu)\phi(G, S, \delta) : \widehat{P}(G) \geq \alpha - \nu\phi(G, S, \delta) \right\}.$$

As above, assume ϕ is independent of the sample and constant on each $\mathcal{G}^k$. Denote $\epsilon_k(n, \delta) = \phi(G, S, \delta)$ for $G \in \mathcal{G}^k$. Observe that computing $\widehat{G}^{\nu}_{\mathcal{G},\alpha}$ is equivalent to computing the MV-ERM estimate on each $\mathcal{G}^k$ at the level $\alpha(k, \nu) = \alpha + (1-\nu)\epsilon_k(n, \delta)$, and then minimizing the penalized volume over these MV-ERM estimates.

Like the original MV-SRM, this modified procedure also obeys an oracle inequality. Recall the notation $\mathcal{G}^k_{\alpha(k,\nu)} = \{G \in \mathcal{G}^k : P(G) \geq \alpha(k, \nu)\} = \{G \in \mathcal{G}^k : P(G) \geq \alpha + (1-\nu)\epsilon_k(n, \delta)\}$.

Theorem 5. *Let $-1 \leq \nu \leq 1$. Set $\alpha(k, \nu) = \alpha + (1-\nu)\epsilon_k(n, \delta)$. Assume* **A1** *and* **A2** *hold. With probability at least $1 - \delta$,*

$$\mathcal{M}(\widehat{G}^{\nu}_{\mathcal{G},\alpha}) \leq \left(1 + \frac{1}{\gamma_\alpha} \right) \min_{1 \leq k \leq K} \left[\inf_{G \in \mathcal{G}^k_{\alpha(k,\nu)}} \left\{ \mu(G) - \mu^*_{\alpha(k,\nu)} \right\} + C_k \epsilon_k(n, \delta) \right], \tag{17}$$

where $C_k = \left((1+\nu) + \frac{1}{\gamma_{\alpha(k,\nu)}}(1-\nu) \right)$.

Here $\gamma_{\alpha(k,\nu)}$ is the density level corresponding to the MV-set with mass $\alpha(k, \nu)$. It may be bounded above in terms of known quantities, as discussed in the previous section. The proof of the theorem is found in Appendix E. Notice that in the case $\nu = 1$ we recover Theorem 4 (under the stated assumptions on $\mathcal{G}$ and ϕ). Also note that $\mathcal{G}^k_{\alpha(k,\nu)}$ will be empty of $\alpha(k, \nu) > 1$, in which case those k should be excluded from the min.

To understand the result, assume that the rate at which $\mathcal{G}^k_\alpha$ approximates G^*_α is independent of α. In other words, the rate at which $\inf_{G \in \mathcal{G}^k_\alpha} \mu(G) - \mu^*_\alpha$ tends to zero as k increases is the

same for all α. Then in the theorem we may replace the expression $\inf_{G \in \mathcal{G}^k_{\alpha(k,\nu)}} \mu(G) - \mu^*_{\alpha(k,\nu)}$ with $\inf_{G \in \mathcal{G}^k_\alpha} \mu(G) - \mu^*_\alpha$. Thus, the ν-damped MV-SRM error decays at the same rate is the original MV-SRM, and adaptively selects the appropriate model class $\mathcal{G}^k$ from which to draw the estimate. Furthermore, damping the penalty by ν has the effect of decreasing the stochastic mass error and adding a stochastic error to the volume. This follows from the above discussion of MV-ERM and the observation that the MV-SRM coincides with an MV-SRM estimate over $\mathcal{G}^k$ for some k. The improved balancing of volume and mass error is confirmed by our experiments in Section 7.

6 Rates of Convergence for Tree-Structured Set Estimators

In this section we illustrate the application of MV-SRM, when combined with an appropriate analysis of the approximation error, to the study of rates of convergence. To preview the main result of this section (Theorem 7), we will consider the class of distributions such that the decision boundary has Lipschitz smoothness (loosely speaking) and d' of the d features are relevant. The best rate of convergence for this class is $n^{-1/d'}$. We will show that MV-SRM can achieve this rate (within a log factor) without knowing d' or which features are relevant. This demonstrates the strength of the oracle inequality, from which the result is derived.

To obtain these rates we apply MV-SRM to sets based on a special family of decision trees called dyadic decision trees (DDTs) [Scott and Nowak, 2004]. Before introducing DDTs, however, we first introduce the class of distributions $\mathcal{D}$ with which our study is concerned. Throughout this section we assume $\mathcal{X} = [0,1]^d$ and μ is the Lebesgue (equivalently, uniform) measure.

6.1 The Box-Counting Class

Before introducing $\mathcal{D}$ we need some additional notation. Let m denote a positive integer, and define $\mathcal{P}_m$ to be the collection of m^d cells formed by the regular partition of $[0,1]^d$ into hypercubes of sidelength $1/m$. Let $c_1, c_2 > 0$ be positive real numbers. Let G^*_α be a minimum volume set, assumed to exist, and let ∂G^*_α be the topological boundary of G^*_α. Finally, let $N_m(\partial G^*_\alpha)$ denote the number of cells in $\mathcal{P}_m$ that intersect ∂G^*_α.

We define the *box-counting* class to be the set $\mathcal{D}_{\text{BOX}} = \mathcal{D}_{\text{BOX}}(c_1, c_2)$ of all distributions satisfying

A1' : X has a density f with respect to μ and f is essentially bounded by c_1.

A3 : $\exists G^*_\alpha$ such that $N_m(\partial G^*_\alpha) \leq c_2 m^{d-1}$ for all m.

Note that since μ is the Lebesgue measure, assumption **A2** from above follows from **A1**, so we do not need to assume it explicitly here. Assumption **A1'** is a slight strengthening of **A1** and implies $P(A) \leq c_1 \mu(A)$ for all measurable sets A. Assumption **A3** essentially requires the boundary of the minimum volume set G^*_α to have Lipschitz smoothness, and thus one would expect the optimal rate of convergence to be $n^{-1/d}$ (the typical rate for set estimation problems characterized by Lipschitz smoothness). See Scott and Nowak [2004] for further discussion of the box-counting assumption.

6.2 Dyadic Decision Trees

Let T denote a tree structured classifier $T : [0,1]^d \to \{0,1\}$. Each such T gives rise to a set $G_T = \{x \in [0,1]^d : T(x) = 1\}$. In this subsection we introduce a certain class of trees, and later consider MV-SRM over the induced class of sets.

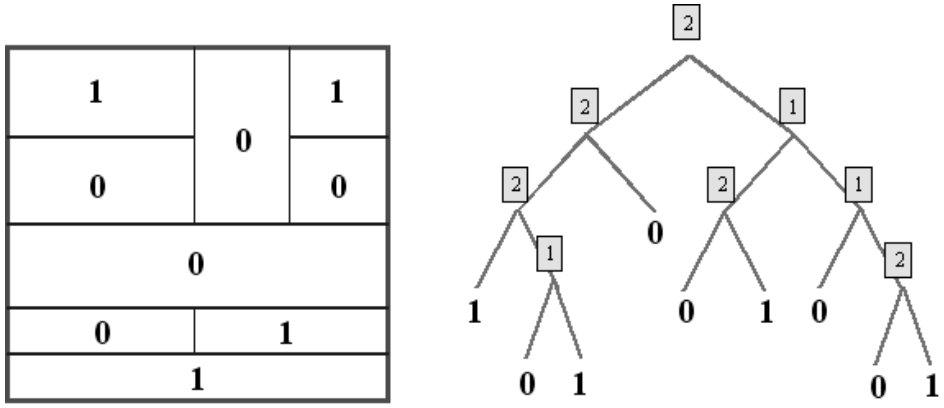

Figure 3: A dyadic decision tree (right) with the associated recursive dyadic partition (left) in $d = 2$ dimensions. Each internal node of the tree is labeled with an integer from 1 to d indicating the coordinate being split at that node. The leaf nodes are decorated with class labels.

Scott and Nowak [2005b, 2004] demonstrate that *dyadic decision trees* (DDTs) offer a computationally feasible classifier that also achieves optimal rates of convergence (for standard classification) under a wide range of conditions. DDTs are especially well suited for rate of convergence studies. Indeed, bounding the approximation error is handled by the restriction to dyadic splits, which allows us to take advantage of recent insights from multiresolution analysis and nonlinear approximations [DeVore, 1998, Cohen et al., 2001, Donoho, 1999]. An analysis similar to that of Scott and Nowak [2004] applies to MV-SRM for DDTs, leading to similar results: optimal rates of convergence for a computationally efficient learning algorithm.

A dyadic decision tree is a decision tree that divides the input space by means of axis-orthogonal dyadic splits. More precisely, a DDT T is a binary tree (with a distinguished root node) specified by assigning (1) an integer $c(v) \in \{1, \ldots, d\}$ to each internal node v of T (corresponding to the coordinate that gets split at that node); (2) a binary label 0 or 1 to each leaf node of T. The nodes of DDTs correspond to hyperrectangles (cells) in $[0,1]^d$. Given a hyperrectangle $A = \prod_{c=1}^{d} [a_c, b_c]$, let $A^{c,1}$ and $A^{c,2}$ denote the hyperrectangles formed by splitting A at its midpoint along coordinate c. Specifically, define $A^{c,1} = \{x \in A \mid x_c \leq (a_c + b_c)/2\}$ and $A^{c,2} = A \backslash A^{c,1}$.

Each node of T is associated with a cell according to the following rules: (1) The root node is associated with $[0,1]^d$; (2) If v is an internal node associated with the cell A, then the children of v are associated with $A^{c(v),1}$ and $A^{c(v),2}$. See Figure 3. Note that every T corresponds to a set $G_T \in [0,1]^d$ (the regions labeled 1), and we think of DDTs as both classifiers and sets interchangeably.

Let $L = L(n)$ be a natural number and define $\mathcal{T}^L$ to be the collection of all DDTs such that (1) no leaf cell has a sidelength smaller than 2^{-L}, and (2) any two leaf nodes that are siblings have different labels. Condition (1) says that when traversing a path from the root to a leaf no coordinate is split more than L times. Condition (2) means that it is impossible to "prune" at any internal node and still have the same set/classifier. Also define $\mathcal{A}^L$ to be the collection of all cells A that correspond to nodes of DDTs in $\mathcal{T}^L$. Define $\pi(T)$ to be the collection of "leaf" cells of T. For a cell $A \in \mathcal{A}^L$, let $j(A)$ denote the depth of A when viewed as a node in some DDT. Observe that when μ is the Lebesgue measure, $\mu(A) = 2^{-j(A)}$.

6.3 MV-SRM with Dyadic Decision Trees

We study MV-SRM over the family $\mathcal{G}^L = \{G_T : T \in \mathcal{T}^L\}$, where L is set by the user. To simplify the notation, at times we will suppress the dependence of ϕ on the training sample S and confidence parameter δ. Thus our MV set estimator has the form

$$\widehat{G}_\alpha = \arg\min_{G \in \mathcal{G}^L} \left\{ \mu(G) + \phi(G) \,|\, \widehat{P}(G) + \phi(G) \geq \alpha \right\}. \tag{18}$$

It remains to specify the penalty ϕ. There are a number of ways to produce ϕ satisfying

$$P^n \left(\left\{ S : \sup_{G \in \mathcal{G}^L} \left(\left| P(G) - \widehat{P}(G) \right| - \phi(G, S, \delta) \right) > 0 \right\} \right) \leq \delta.$$

Since $\mathcal{G}^L$ is countable (in fact, finite), one approach is to devise a prefix code for $\mathcal{G}^L$ and apply the penalty in Section 2.2. Instead, we employ a different penalty which has the advantage that it leads to minimax optimal rates of convergence. Introduce the notation $[\![A]\!] = (3 + \log_2 d) j(A)$, which may be thought of as the codelength of A in a prefix code for $\mathcal{A}^L$, and define the *minimax* penalty

$$\phi(G_T) := \sum_{A \in \pi(T)} \sqrt{8 \max \left(\widehat{P}(A), \frac{[\![A]\!] \log 2 + \log(2/\delta)}{n} \right) \frac{[\![A]\!] \log 2 + \log(2/\delta)}{n}}. \tag{19}$$

For each $A \in \pi(T)$, set $\ell(A) = 1$ if $A \subset G_T$ and 0 otherwise. The bound originates from writing

$$P(G_T) - \widehat{P}(G_T) = \sum_{A \in \pi(T) : \ell(A) = 1} P(A) - \widehat{P}(A)$$

and

$$\begin{aligned} \widehat{P}(G_T) - P(G_T) &= P(\overline{G_T}) - \widehat{P}(\overline{G_T}) \\ &= \sum_{A \in \pi(T) : \ell(A) = 0} P(A) - \widehat{P}(A) \end{aligned}$$

from which it follows that

$$|P(G_T) - \widehat{P}(G_T)| \leq \sum_{A \in \pi(T)} P(A) - \widehat{P}(A). \tag{20}$$

The event $X \in A$ is a Bernoulli trial with probability of success $P(A)$, and so bounding the right hand side of (20) simply involves applying a concentration inequality for binomials to each $A \in \mathcal{A}^L$. There are many ways to do this (additive Chernoff, relative Chernoff, exact tail inversion, etc.), but the one we have chosen is particularly convenient for rate of convergence analysis. For further discussion, see Scott and Nowak [2004]. Proof of the following result is nearly identical to a similar result in Scott and Nowak [2004], and is omitted.

Proposition 2. *Let ϕ be as in (19) and let $\delta \in (0, 1)$. With probability at least $1 - \delta$ over the draw of S,*

$$|P(G) - \widehat{P}(G)| \leq \phi(G)$$

for all $G \in \mathcal{G}^L$. Thus ϕ is a complexity penalty for $\mathcal{G}^L$.

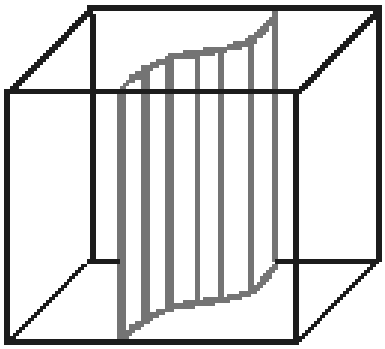

Figure 4: Cartoon illustrating relevant data dimension. If the X^3 axis is irrelevant, then the boundary of the MV-set is a "vertical sheet" over a curve in the (X^1, X^2) plane.

The MV-SRM procedure over $\mathcal{G}^L$ with the above penalty leads to an optimal rate of convergence for the box-counting class.

Theorem 6. *Choose $L = L(n)$ and $\delta = \delta(n)$ such that*

1. $2^{L(n)} \succcurlyeq (n/\log n)^{1/d}$
2. $\delta(n) = O(\sqrt{\log n/n})$ *and* $\log(1/\delta(n)) = O(\log n)$

Define $\widehat{G}_\alpha$ as in (18) with ϕ as in (19). For $d \geq 2$ we have

$$\sup_{\mathcal{D}_{\text{BOX}}} \mathbf{E}^n \mathcal{M}(\widehat{G}_\alpha) \preccurlyeq \left(\frac{\log n}{n}\right)^{\frac{1}{d}}. \tag{21}$$

We omit the proof, since this theorem is a special case of Theorem 7 below. Note that the condition on δ is satisfied if $\delta(n) \asymp n^{-\beta}$ for some $\beta > 1/2$.

6.4 Adapting to relevant features

The previous result could have been obtained without using MV-SRM. Instead, we could have applied MV-ERM to a fixed hierarchy $\mathcal{G}^{L(1)}, \mathcal{G}^{L(2)}, \ldots$ where $L(n) \asymp (n/\log n)^{1/d}$. The strength of MV-SRM and the associated oracle inequality is in its ability to adapt to favorable conditions on the data generating distribution which may not be known in advance. Here we illustrate this idea when the number of relevant features is not known in advance.

We define the *relevant data dimension* to be the number $d' \leq d$ of relevant features. A feature X^i, $i = 1, \ldots, d$, is said to be relevant provided $f(X)$ is not constant when X^i is varied from 0 to 1. For example, if $d = 2$ and $d' = 1$, then ∂G^*_α is a horizontal or vertical line segment (or union of such line segments). If $d = 3$ and $d' = 1$, then ∂G^*_α is a plane (or union of planes) orthogonal to one of the axes. If $d = 3$ and the third coordinate is irrelevant ($d' = 2$), then ∂G^*_α is a "vertical sheet" over a curve in the (X^1, X^2) plane (see Figure 4).

Let $\mathcal{D}'_{\text{BOX}} = \mathcal{D}'_{\text{BOX}}(c_1, c_2, d')$ be the set of all product measures P^n such that **A1'** and **A3** hold for the underlying distribution P, and X has relevant data dimension $d' \geq 2$. An argument of Scott and Nowak [2004] implies that the expected minimax rate for d' relevant features is $n^{-1/d'}$. By the following result, MV-SRM can achieve this rate to within a log factor.

Theorem 7. *Choose $L = L(n)$ and $\delta = \delta(n)$ such that*

1. $2^{L(n)} \succcurlyeq n/\log n$
2. $\delta(n) = O(\sqrt{\log n/n})$ *and* $\log(1/\delta(n)) = O(\log n)$

Define $\widehat{G}_\alpha$ as in (18) with ϕ as in (19). If $d' \geq 2$ then

$$\sup_{\mathcal{D}'_{\text{BOX}}} \mathbf{E}^n \mathcal{M}(\widehat{G}_\alpha) \preccurlyeq \left(\frac{\log n}{n}\right)^{\frac{1}{d'}}. \tag{22}$$

The proof hinges on the oracle inequality. The details of the proof are very similar to the proof of a result in Scott and Nowak [2004] and are therefore omitted. Here we just give a sketch of how the oracle inequality comes into play.

Let $K \leq L$ and let $G_K^* \in \mathcal{G}^K$ be such that (i) $\mu(G_K^*) = \arg\min_{G \in \mathcal{G}^K} \mu(G) - \mu_\alpha^*$; and (ii) G_K^* is based on the smallest possible partition among all sets satisfying (i). Set $m = 2^K$. It can be shown that

$$\mu(G_K^*) - \mu_\alpha^* + 2\phi(G_K^*, S, \delta) \preccurlyeq m^{-1} + m^{d'/2-1}\sqrt{\frac{\log n}{n}}$$

in expectation. This upper bound is minimized when $m \asymp (n/\log n)^{1/d'}$, in which case we obtain the stated rate. Here the oracle inequality is crucial because m depends on d', which is not known in advance. The oracle inequality tells us that MV-SRM performs as if it knew the optimal K.

Note that the set estimation rule does not require knowledge of the constants c_1 and c_2, nor d', nor which features are relevant. Thus the rule is completely automatic and adaptive.

7 Experiments

In this section we conduct some simple numerical experiments to illustrate the rules for MV-set estimation proposed in this work. Our objective is not an extensive comparison with competing methods, but rather to demonstrate that our estimators behave in a way that agrees with the theory, to gain insight into the behavior of various penalties, and to examine basic algorithmic issues. Throughout this section we take $\mathcal{X} = [0,1]^d$ and μ to be the Lebesgue (equivalently, uniform) measure.

7.1 Histograms

We devised a simple numerical experiment to illustrate MV-SRM in the case of histograms (see Sections 3.2 and 4.2). In this case, MV-SRM can be implemented exactly with a simple procedure. First, compute the MV-ERM estimate for each $\mathcal{G}^k$, $k = 1, \ldots, K$, where $1/k$ is the bin-width. To do this, for each k, sort the cells of the partition according to the number of samples in the cell. Then, begin incorporating cells into the estimate one cell at a time, starting with the most populated, until the empirical mass constraint is satisfied. Finally, once all MV-ERM estimates have been computed, choose the one that minimizes the penalized volume.

We consider two penalties. Both penalties are defined via $\phi(G, S, \delta) = \phi_k(G, S, \delta 2^{-k})$ for $G \in \mathcal{G}^k$, where ϕ_k is a penalty for $\mathcal{G}^k$. The first is based on the simple Occam-style bound of Section 3.2. For $G \in \mathcal{G}^k$, set

$$\phi_k^{Occ}(G, S, \delta) = \sqrt{\frac{k^d \log 2 + \log(2/\delta)}{2n}}.$$

The second is the (conditional) Rademacher penalty. For $G \in \mathcal{G}^k$, set

$$\phi_k^{Rad}(G, S, \delta) = \frac{2}{n}\mathbf{E}_{(\sigma_i)}\left[\sup_{G' \in \mathcal{G}^k} \sum_{i=1}^{n} \sigma_i \mathbb{I}\left(X_i \in G'\right)\right] + \sqrt{\frac{2\log(2/\delta)}{n}}.$$

Here $\sigma_1, \ldots, \sigma_n$ are Rademacher random variables, i.e., independent random variables taking on the values 1 and -1 with equal probability. Fortunately, the conditional expectation with respect to these variables can be evaluated exactly in the case of partition-based rules such as the histogram. See Appendix F for details.

As a data set we consider $\mathcal{X} = [0,1]^d$, the unit square, and data generated by a two-dimensional truncated Gaussian distribution, centered at the point $(1/2, 1/2)$ and having spherical variance with parameter $\sigma = 0.15$. Other parameter settings are $\alpha = 0.8$, $K = 40$, and $\delta = 0.05$. All experiments were conducted at nine different sample sizes, logarithmically spaced from 100 to 1000000, and repeated 100 times. Figure 5 shows a representative training sample and MV-ERM estimates with $\nu = 1, 0$, and -1. These examples clearly demonstrate that the larger ν, the smaller the estimate.

Figure 6 depicts the error $\mathcal{M}(\widehat{G})$ of the MV-SRM estimate with $\nu = 1$. The Occam's Razor penalty consistently outperforms the Rademacher penalty. For comparison, a damped version ($\nu = 0$) was also evaluated. It is clear from the graphs that $\nu = 0$ outperforms $\nu = 1$. This happens because the damped version distributes the error more evenly between mass and volume, as discussed in Section 5.

Figure 7 depicts the penalized volume of the MV-ERM estimates ($\nu = 1$) as a function of the resolution k, where $1/k$ is the sidelength of the histogram cell. MV-SRM selects the resolution where this curve is minimized. Clearly the Occam's Razor bound is tighter than the Rademacher bound (look at the right side of the graph), which explains why Occam outperforms Rademacher. Figure 8 depicts the average resolution of the estimate (top) and the average symmetric difference with respect to the true MV-set, for various sample sizes. These graphs are for $\nu = 1$. The graphs for $\nu = 0$ do not change considerably. Thus, while damping seems to have a noticeable effect on the error quantity $\mathcal{M}$, the effect on the symmetric difference is much less pronounced.

7.2 Dyadic decision trees

Implementing MV-SRM for dyadic decision trees is much more challenging than for histograms. In Appendix G we give an exact algorithm, although admittedly this algorithm is quite time consuming and thus of marginal use when seeking to conduct a large number of experiments. Instead, in this section we suggest an approximate algorithm based on a reformulation of the constrained optimization problem defining MV-SRM in terms of its Lagrangian, coupled with a bisection search to find the appropriate Lagrange multiplier. If the penalty is additive, then the unconstrained Lagrangian can be minimized efficiently using existing algorithmic approaches.

A penalty for a DDT is said to be *additive* if it can be written in the form

$$\phi(G_T) = \sum_{A \in \pi(T)} \psi(A)$$

for some ψ. If ϕ is additive the optimization in (18) can be re-written as

$$\min_{T \in \mathcal{T}^L} \sum_{A \in \pi(T)} [\mu(A)\ell(A) + (1+\nu)\psi(A)] \quad \text{subject to} \quad \sum_{A \in \pi(T)} \left[\widehat{P}(A)\ell(A) + \nu\psi(A)\right] \geq \alpha$$

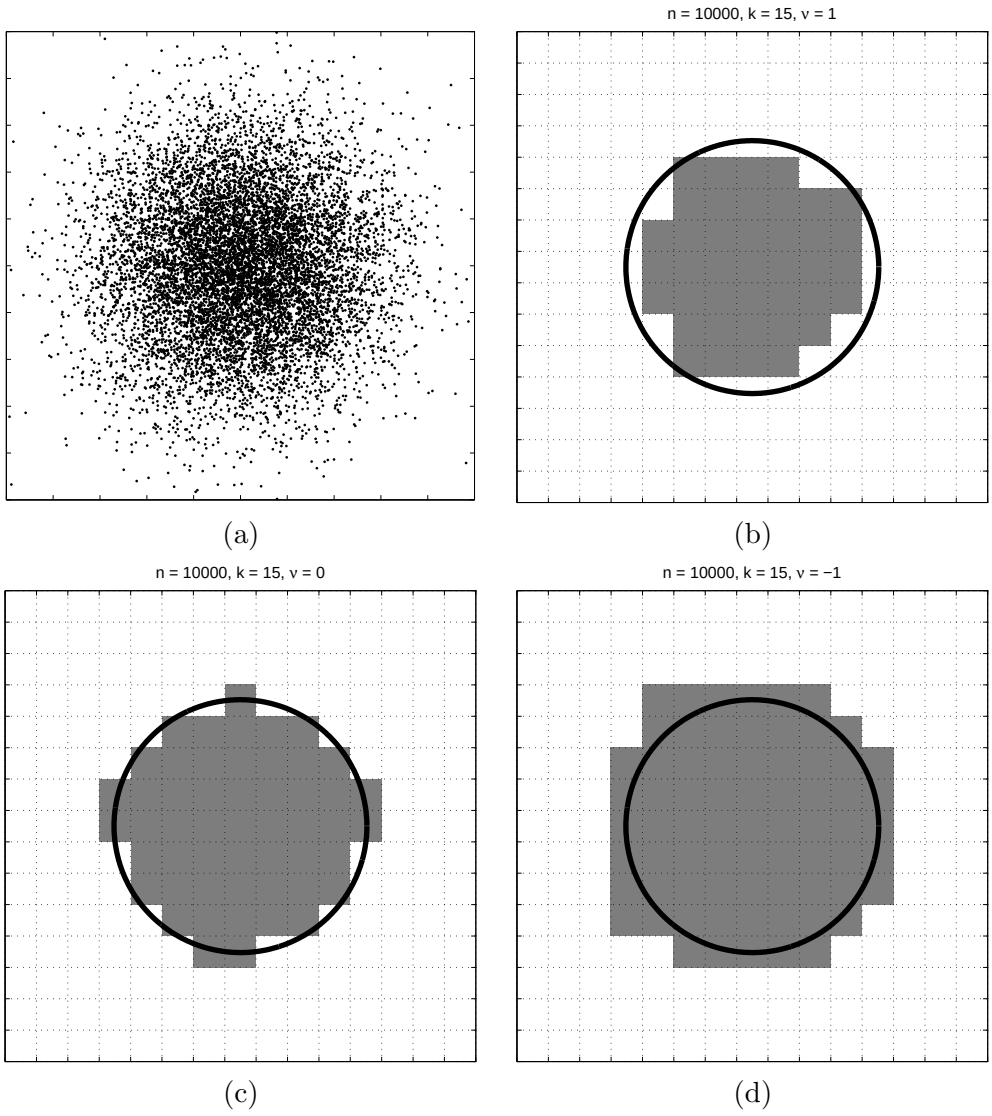

Figure 5: Data and three representative MV-ERM histogram estimates for the data in Section 7.1. The shaded region is the MV-set estimate, and the solid circle indicates the true MV-set. All estimates are based on the Occam bound. (a) 10000 realizations used for training. (b) MV-ERM estimate with a bin-width of 1/15 and $\nu = 1$. (c) $\nu = 0$. (d) $\nu = -1$. Clearly, the larger ν, the smaller the estimate.

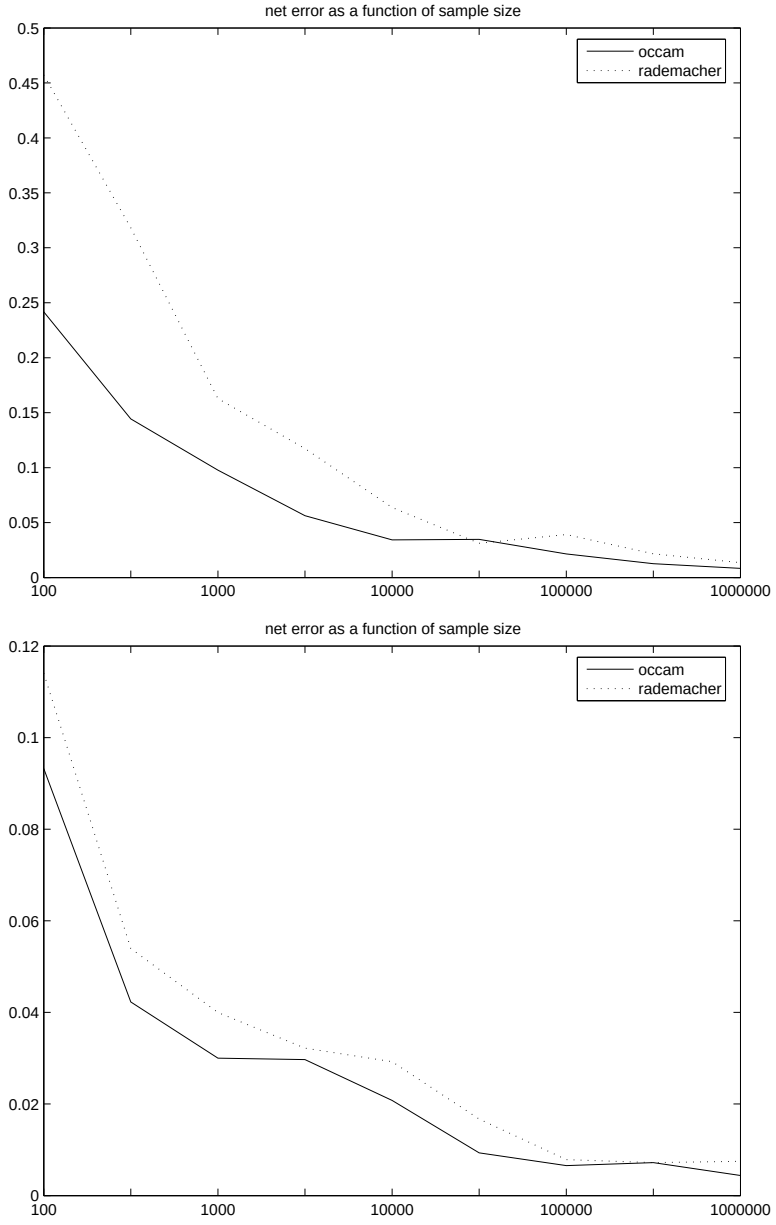

Figure 6: The error $\mathcal{M}(\widehat{G}_{\mathcal{G},\alpha})$ as a function of sample size for the histogram experiments in Section 7.1. All results are averaged over 100 repetitions for each training sample size. (Top) Results for the original MV-SRM algorithm ($\nu = 1$). (Bottom) Results for $\nu = 0$. In this case the error is more evenly distributed between mass and volume, whereas in the former case all the error is in the mass term.

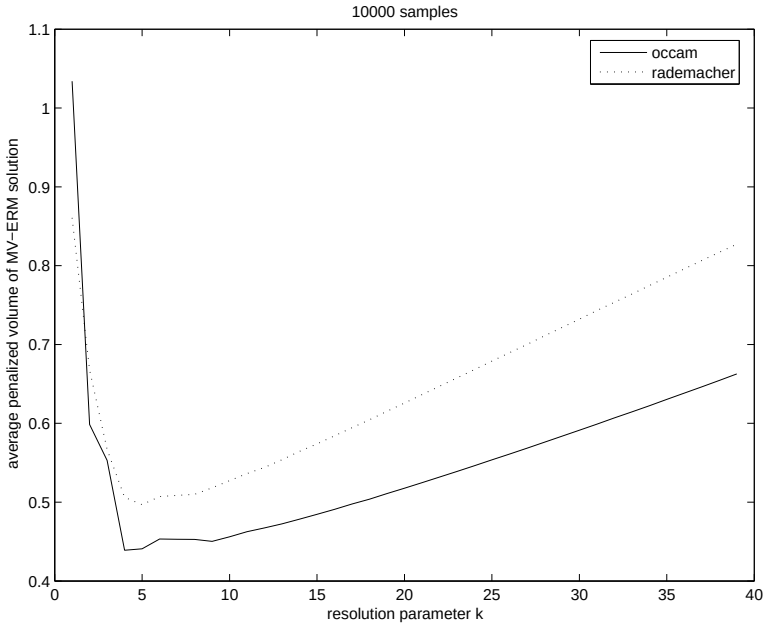

Figure 7: The penalized volume of the MV-ERM estimates $G^k_{\mathcal{G},\alpha}$, as a function of k, where $1/k$ is the sidelength of the histogram cell. The results are for a sample size of 10000. Results represent an average over 100 repetitions. Clearly, the Occam's razor bound is smaller than the Rademacher penalty (look at the right side of the plot), to which we may attribute its improved performance (see Figure 6).

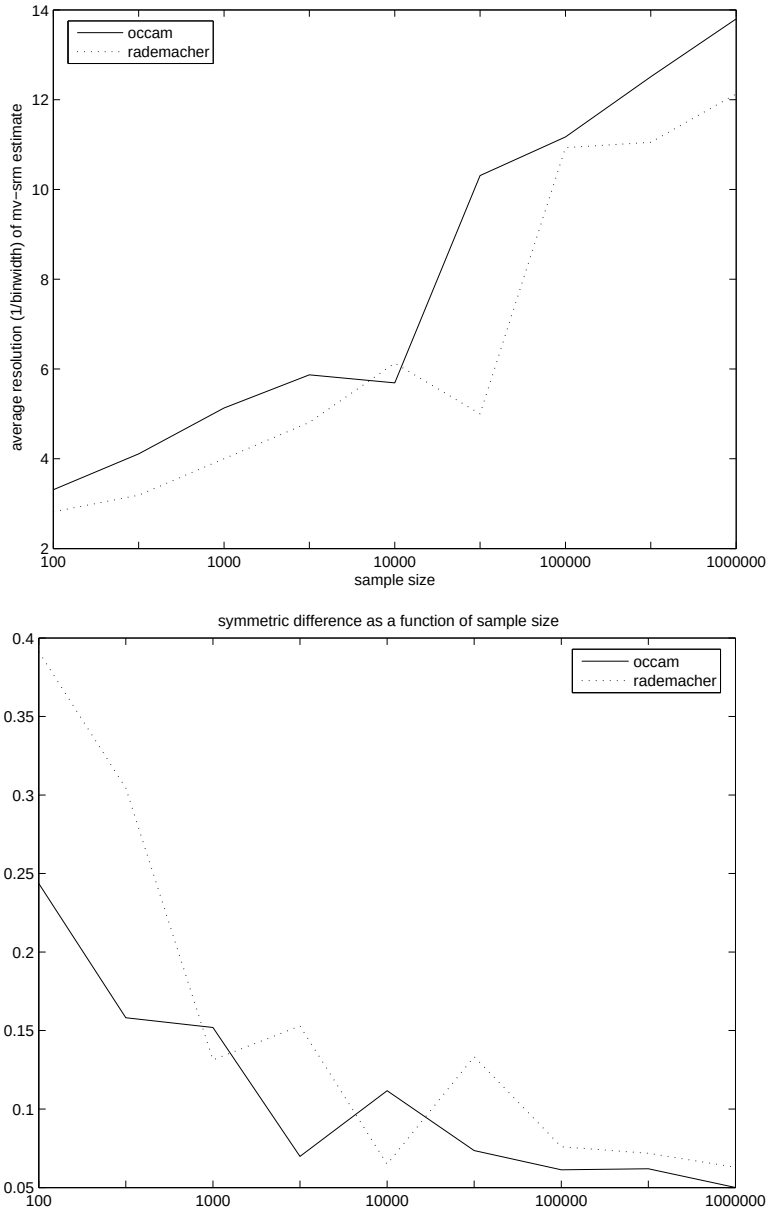

Figure 8: Results from the histogram experiments in Section 7.1. All results are averaged over 100 repetitions for each training sample size, and are for the non-damped version of MV-SRM ($\nu = 1$). (Top) Average value of the resolution parameter k ($1/k$ = sidelength of histogram cells) as a function of sample size. (Bottom) Average value of the symmetric difference between the estimated and true MV-sets. Neither graph changes significantly if ν is varied.

where $\ell(A)$ is the binary label of leaf A ($\ell(A) = 1$ if A is in the candidate set and 0 otherwise). Introducing the Lagrange multiplier $\lambda > 0$, the unconstrained Lagrangian formulation of the problem is

$$\min_T \sum_{A \in T} \left[\mu(A)\ell(A) + (1+\nu)\psi(A) - \lambda \left(\widehat{P}(A)\ell(A) + \nu\psi(A) \right) \right].$$

Inspection of the Lagrangian reveals that the optimal choice of $\ell(A)$ is

$$\ell(A) \;=\; \begin{cases} 1 & \text{if } \lambda \widehat{P}(A) \geq \mu(A), \\ \\ 0 & \text{otherwise} \end{cases}$$

Thus, we have a "per-leaf" cost function

$$\text{cost}(A) \;:= \min(\mu(A) - \lambda \widehat{P}(A), 0) + [1 + \nu(1-\lambda)]\psi(A)$$

For a given value of λ, the optimal tree can be efficiently obtained using the algorithm of Blanchard et al. [2004].

We also note that the above strategy works for tree structures besides the one studied in Section 6. For example, suppose an overfitted tree (with arbitrary, non-dyadic splits) has been constructed by some greedy heuristic (perhaps using an independent dataset). Or, suppose that instead of binary dyadic splits with arbitrary orientation, one only considers "quadsplits" whereby every parent node has 2^d children (in fact, this is the tree structure used for our experiments below). In such cases, optimizing the Lagrangian reduces to a classical pruning problem, and the optimal tree can be found by a simple $O(n)$ dynamic program that has been used since at least the days of CART [Breiman et al., 1984].

Let $\widehat{T}_\lambda$ denote the tree resulting from the Lagrangian optimization above. From standard optimization theory, we know that for each value of λ, $\widehat{T}_\lambda$ will coincide with $\widehat{G}_\alpha$, for a certain value of α. For each value of λ there is a corresponding α, but the converse is not necessarily true. Therefore, the Lagrangian solutions correspond to many, but not all possible solutions of the original MV-SRM optimization with different values of α. Despite this potential limitation, the simplicity of the Lagrangian optimization makes this a very attractive approach to MV-SRM in this case. We can determine the best value of λ for a given target α by repeatedly solving the Lagrangian optimization and finding the setting for λ that meets or comes closest to the original constraint. The search over λ can be conducted efficiently using a bisection search.

In our experiments we do not consider the "free-split" tree structure described in Section 6, in which each parent has two children defined by one of $d = 2$ possible splits. Instead, we assume a quadsplit tree structure, whereby every cell is a square, and every parent has four square children. The total optimization time is $O(mn)$, where m is the number of steps in the bisection search. In our experiments presented below we found that ten steps (i.e., ten Lagrangian tree pruning optimizations) were sufficient to meet the constraint almost exactly (whenever possible).

We consider three complexity penalties. We refer to the first penalty as the *minimax* penalty, since it is inspired by the minimax optimal penalty in (19):

$$\psi^{mm}(A) := (0.01)\sqrt{8 \max\left(\widehat{P}(A), \frac{[\![A]\!] \log 2 + \log(2/\delta)}{n} \right) \frac{[\![A]\!] \log 2 + \log(2/\delta)}{n}}. \tag{23}$$

Note that the penalty is down-weighted by a constant factor of 0.01, since otherwise it is too large to yield meaningful results:[3]

The second penalty is based on the Rademacher penalty (see Section 2.3). Let Π^L denote the set of all partitions π of trees in $\mathcal{T}^L$. Given $\pi_0 \in \Pi^L$, set $\mathcal{G}_{\pi_0} = \{G_T \in \mathcal{G}^L : \pi(T) = \pi_0\}$. Recall $\pi(T)$ denotes the partition associated with the tree T. Combining Proposition 1 with the results of Appendix F, we know that for any fixed π,

$$\sum_{A \in \pi} \sqrt{\frac{\widehat{P}(A)}{n}} + \sqrt{\frac{2\log(2/\delta)}{n}}$$

is a complexity penalty for $\mathcal{G}_\pi$. To obtain a penalty for all $\mathcal{G}^L = \cup_{\pi \in \Pi^L} \mathcal{G}_\pi$, we apply the union bound over all $\pi \in \Pi^L$ and replace δ by $\delta |\Pi^L|^{-1}$. Although distributing the "delta" uniformly across all partitions is perhaps not intuitive (one might expect smaller partitions to be more likely and hence they should receive a larger chunk of the delta), it has the important property that the delta term is the same for all trees, and thus can be dropped for the purposes of minimization. Hence, the effective penalty is additive. In summary, our second penalty, referred to as the Rademacher penalty,[4] is given by

$$\psi^{Rad}(A) \;=\; \sqrt{\frac{\widehat{P}(A)}{n}}. \tag{24}$$

The third penalty is referred to as the modified Rademacher penalty and is given by

$$\psi^{mRad}(A) \;=\; \sqrt{\frac{\widehat{P}(A) + \mu(A)}{n}}. \tag{25}$$

The modified Rademacher penalty is still a valid penalty, since it strictly dominates the basic Rademacher penalty. The basic Rademacher is proportional to the square-root of the empirical P mass and the modified Rademacher is proportional to the square-root of the *total* mass (empirical P mass plus μ mass). In our experiments we have found that the modified Rademacher penalty typically performs better than the basic Rademacher penalty, since it discourages the inclusion of very small isolated leafs containing a single data point (as seen in the experimental results below). Note that, unlike the minimax penalty, the two Rademacher-based penalties are not down-weighted; the true penalties are used.

We illustrate the performance of the dyadic quadtree approach with a two-dimensional Gaussian mixture distribution, taking $\nu = 0$. Figure 1 depicts 500 samples from the Gaussian mixture distribution, along with the true minimum volume set for $\alpha = 0.90$. Figures 9, 10, and 11 depict the minimum volume set estimates based on each of the three penalties, and for sample sizes of 100, 1000, and 10000. Here we use MM, Rad, and mRad to designate the three penalties.

In addition to the minimum volume set estimates based on a single tree, we also show the estimates based on voting over shifted partitions. This amounts to constructing $2^L \times 2^L$ different

[3]Note that here down-weighting is distinct from damping by ν as discussed earlier. With down-weighting, both occurrences of the penalty, in the constraint and in the objective function, are scaled by the same factor. Downweighting is a heuristic, whereas damping has theoretical support.

[4]Technically, this is an upper bound on the Rademacher penalty, but as discussed in Appendix F, this bound is tight to within a factor of $\sqrt{2}$. Using the exact Rademacher yields essentially the same results. Thus, we refer to this upper bound simply as the Rademacher penalty.

trees, each based on a partition offset by an integer multiple of the base sidelength 2^{-L}, and taking a majority vote over all the resulting set estimates to form the final estimate. These estimates are indicated by MM', Rad', and mRad', respectively. Similar methods based on averaging or voting over shifted partitions have been tremendously successful in image processing, and they tend to mitigate the "blockiness" associated with estimates based on a single tree, as is clearly seen in the results depicted. Moreover, because of the significant amount of redundancy in the shifted partitions, the MM', Rad', and mRad' estimates can be computed in just $O(mn \log n)$ operations.

Visual inspection of the resulting minimum volume set estimates (which were "typical" results selected at random) reveals some of the characteristics of the different penalties and their behaviors as a function of the sample size. Notably, the basic Rademacher penalty tends to allow very small and isolated leafs into the final set estimate, which is somewhat unappealing. The modified Rademacher penalty clearly eliminates this problem and provides very reasonable estimates. The (down-weighted) minimax penalty results in set estimates quite similar to those resulting from the modified Rademacher. However, the somewhat arbitrary choice of scaling factor (0.01 in this case) is undesirable. Finally, let us remark on the significant improvement provided by voting over multiple shifted trees. The voting procedure quite dramatically reduces the "blocky" partition associated with estimates based on single trees. Overall, the modified Rademacher penalty coupled with voting over multiple shifted trees appears to perform best in our experiments. In fact, in the case $n = 10000$, this set estimate is almost identical to the true minimum volume set depicted in Figure 1.

8 Conclusions

In this paper we propose two rules, MV-ERM and MV-SRM, for estimation of minimum volume sets. Our theoretical analysis is made possible by relating the performance of these rules to the uniform convergence properties of the class of sets from which the estimate is taken. This in turn lets us apply distribution free uniform convergence results such as the VC inequality to obtain distribution free, finite sample performance guarantees. It also leads to strong universal consistency when the class of candidate sets is allowed to grow in a controlled way. MV-SRM obeys an oracle inequality and thereby automatically selects the appropriate complexity of the set estimator. These theoretical results are illustrated with histograms and dyadic decision trees.

Our estimators, results, and proof techniques for minimum volume sets bear a strong resemblance to existing estimators, results, and proof techniques for supervised classification. This is no coincidence. Minimum volume set estimation is closely linked with hypothesis testing. Assume P has a density with respect to μ, and that μ is a probability measure. Then the minimum volume set with mass α is the acceptance region of the most powerful test of size $1 - \alpha$ for testing $H_0 : X \sim P$ versus $H_1 : X \sim \mu$. But classification and hypothesis testing have the same goals; the difference lies in what knowledge is used to design a classifier/test (training data versus knowledge of the true densities). The problem of learning minimum volume sets stands halfway between these two: For one class the true distribution is known (the reference measure), but for the other only training samples are available.

This observation provides not only intuition for the similarity between MV-set estimation and classification, but it also suggests an alternative approach to MV-set estimation. In particular, suppose it is possible to sample at will from the reference measure. Consider these samples, together with the original training data, to be a labeled training set. Then the MV-set may be estimated by

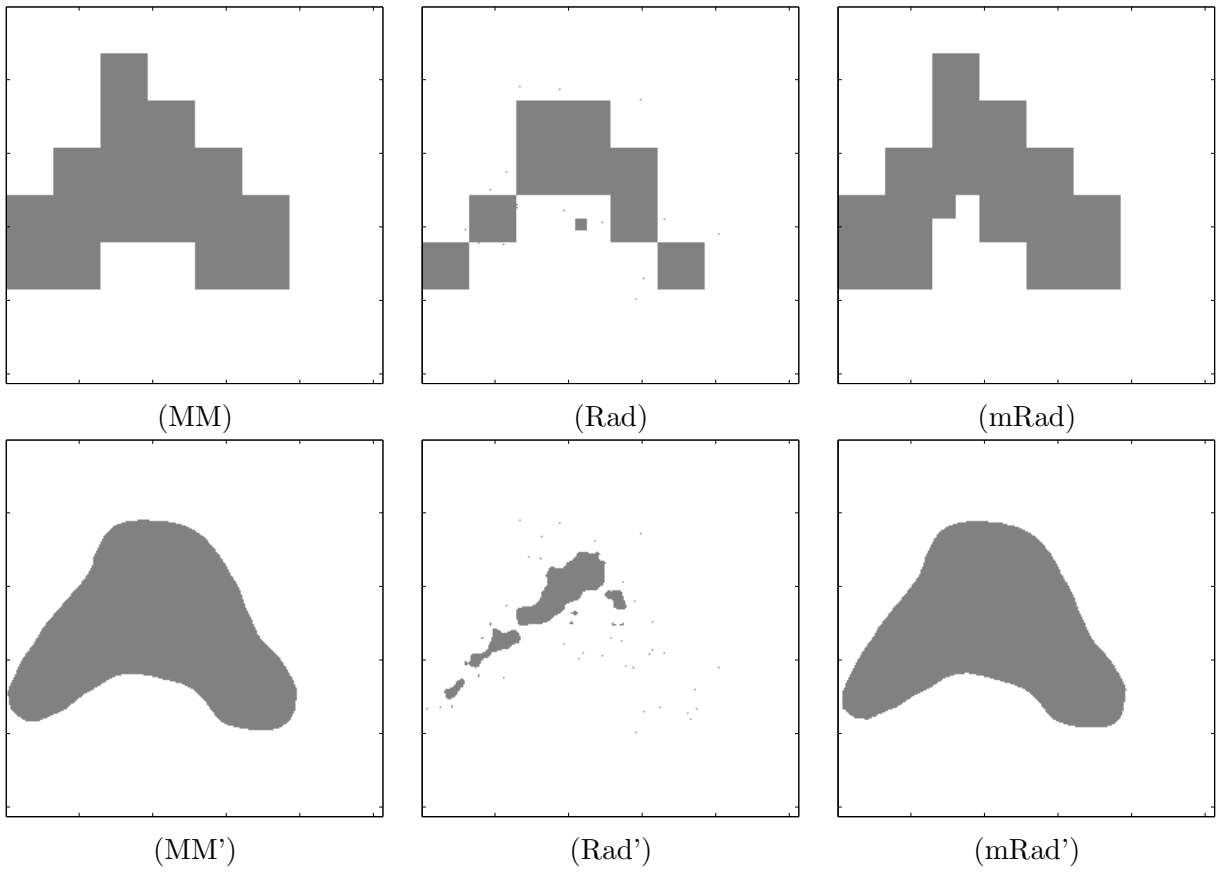

Figure 9: Minimum volume set estimates based on dyadic quadtrees for $\alpha = 0.90$ with $n = 100$ samples. Reconstructions based on MM = minimax penalty (23), Rad = Rademacher penalty (24), and mRad = modified Rademacher penalty (25), and MM', Rad', and mRad' denote the analogous estimates based on voting over multiple trees at different shifts.

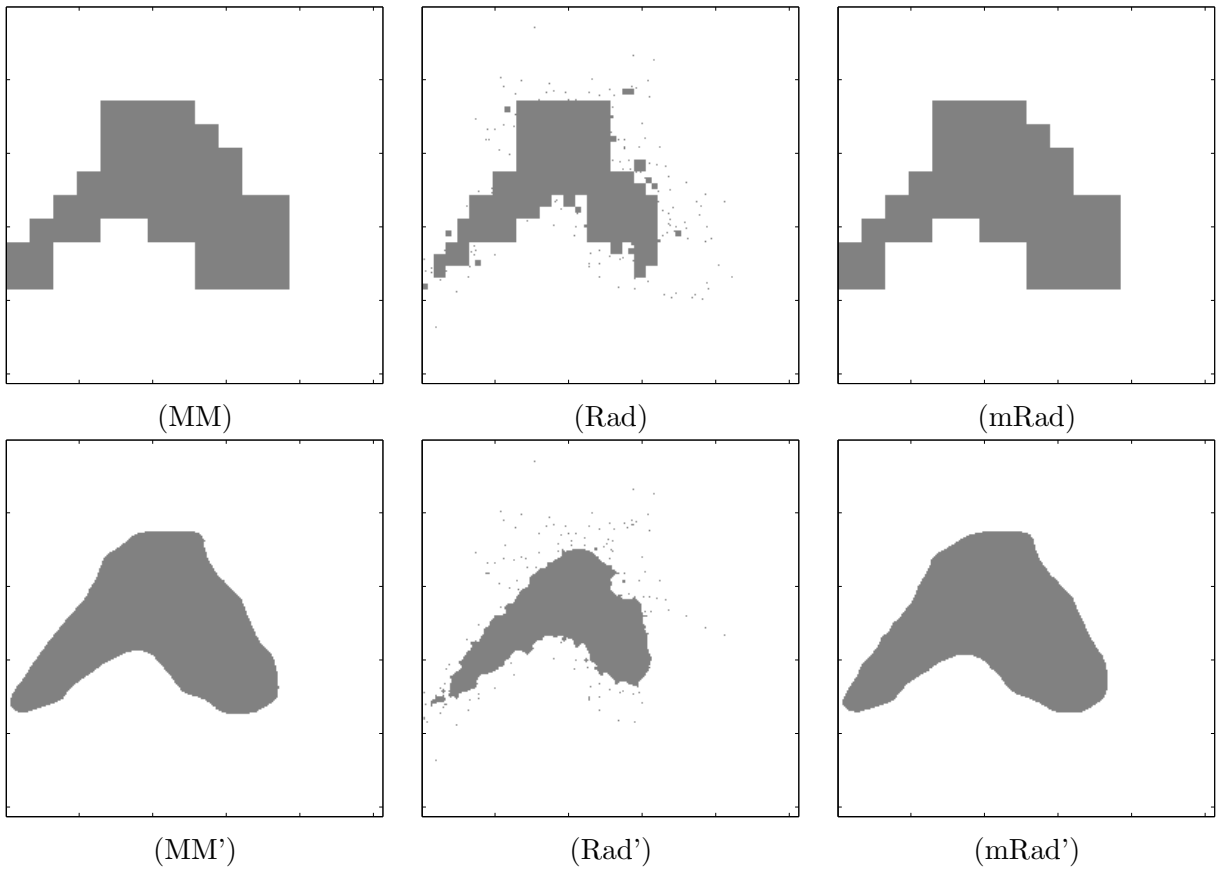

Figure 10: Minimum volume set estimates based on dyadic quadtrees for $\alpha = 0.90$ with $n = 1000$ samples. Reconstructions based on MM = minimax penalty (23), Rad = Rademacher penalty (24), and mRad = modified Rademacher penalty (25), and MM', Rad', and mRad' denote the analogous estimates based on voting over multiple trees at different shifts.

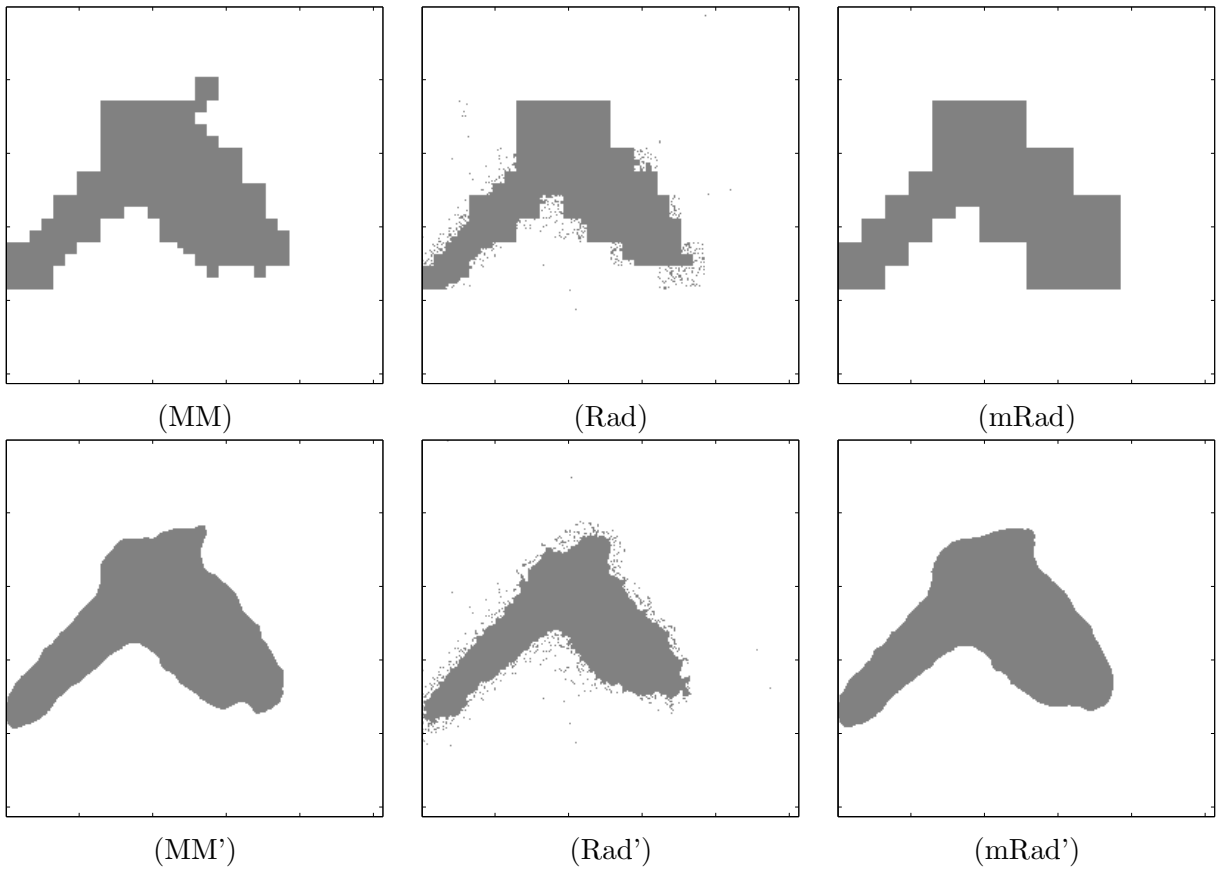

Figure 11: Minimum volume set estimates based on dyadic quadtrees for $\alpha = 0.90$ with $n = 10000$ samples. Reconstructions based on MM = minimax penalty (23), Rad = Rademacher penalty (24), and mRad = modified Rademacher penalty (25), and MM', Rad', and mRad' denote the analogous estimates based on voting over multiple trees at different shifts.

learning a classifier with respect to the Neyman-Pearson criterion [Cannon et al., 2002, Scott and Nowak, 2005a]. Briefly, the Neyman-Pearson classification paradigm involves learning a classifier from training data that minimizes the "miss" generalization error while constraining the "false alarm" generalization error to be less than or equal to a specified size, in our case $1 - \alpha$.

Minimum volume set estimation based on Neyman-Pearson classification offers a distinct advantage over the rules studied in this paper. Indeed, our algorithms for histograms and dyadic decision trees take advantage of the fact that the reference measure μ is easily evaluated for these special types of sets. For more general sets or non-uniform reference measures, direct evaluation of the reference measure may be impractical. Neyman-Pearson classification, in contrast, involves computing the empirical volume based on the training sample, a much easier task. Moreover, in principle one may take an arbitrarily large sample from μ to mitigate finite sample effects. A similar idea has been employed by Steinwart et al. [2005], who sample from μ so as to reduce density level set estimation to cost-sensitive classification. In this setting the advantage of MV-sets over density level sets is further magnified. For example, to sample from a uniform distribution, one must specify its support, which is a priori unknown. Fortunately, MV-sets are invariant to the choice of support, whereas the γ-level set changes with the support of μ.

Acknowledgment

The authors thank Ercan Yildiz and Rebecca Willett for their assistance with the experiments involving dyadic trees, and Gilles Blanchard for his insights into the Rademacher penalty for partition-based estimators.

A Proof of Lemma 2

The proof follows closely the proof of Lemma 1 in Cannon et al. [2002]. Define $\Xi = \{S : \widehat{P}(G_{\mathcal{G},\alpha}) < \alpha - \phi(G_{\mathcal{G},\alpha}, S, \delta)\}$. It is true that $\Theta_\mu \subset \Xi$. To see this, if $S \notin \Xi$ then $G_{\mathcal{G},\alpha} \in \widehat{\mathcal{G}}_\alpha$, and hence $\mu(\widehat{G}_{\mathcal{G},\alpha}) \leq \mu(G_{\mathcal{G},\alpha})$ by definition of $\widehat{G}_{\mathcal{G},\alpha}$. Thus $S \notin \Theta_\mu$. It follows that

$$\Theta_P \cup \Theta_\mu \subset \Theta_P \cup \Xi$$

and hence it suffices to show $\Theta_P \subset \Omega_P$ and $\Xi \subset \Omega_P$.

First, we show that $\Theta_P \subset \Omega_P$. If $S \in \Theta_P$ then

$$P(\widehat{G}_{\mathcal{G},\alpha}) < \alpha - 2\phi(\widehat{G}_{\mathcal{G},\alpha}, S, \delta).$$

This implies

$$\begin{aligned} P(\widehat{G}_{\mathcal{G},\alpha}) - \widehat{P}(\widehat{G}_{\mathcal{G},\alpha}) &< \alpha - 2\phi(\widehat{G}_{\mathcal{G},\alpha}, S, \delta) - \widehat{P}(\widehat{G}_{\mathcal{G},\alpha}) \\ &\leq -\phi(\widehat{G}_{\mathcal{G},\alpha}, S, \delta), \end{aligned}$$

where the last inequality is true because $\widehat{P}(\widehat{G}_{\mathcal{G},\alpha}) \geq \alpha - \phi(\widehat{G}_{\mathcal{G},\alpha}, S, \delta)$. Therefore $S \in \Omega_P$.

Second, we show that $\Xi \subset \Omega_P$. If $S \in \Xi$, then

$$\begin{aligned} \widehat{P}(G_{\mathcal{G},\alpha}) - P(G_{\mathcal{G},\alpha}) &< \alpha - \phi(G_{\mathcal{G},\alpha}, S, \delta) - P(G_{\mathcal{G},\alpha}) \\ &\leq -\phi(G_{\mathcal{G},\alpha}, S, \delta), \end{aligned}$$

where the last inequality holds because $P(G_{\mathcal{G},\alpha}) \geq \alpha$. Thus, $S \in \Omega_P$, and the proof is complete.

B Proof of Theorem 2

By the Borel-Cantelli Lemma [Durrett, 1991], it suffices to show that for any $\epsilon > 0$,

$$\sum_{n=1}^{\infty} P^n(\mathcal{M}(\widehat{G}_{\mathcal{G},\alpha}) > \epsilon) < \infty.$$

We will show this by establishing

$$\sum_{n=1}^{\infty} P^n\left(\left(\mu(\widehat{G}_{\mathcal{G},\alpha}) - \mu^*_\alpha\right)_+ > \frac{\epsilon}{2}\right) < \infty \tag{26}$$

and

$$\sum_{n=1}^{\infty} P^n\left(\left(\alpha - P(\widehat{G}_{\mathcal{G},\alpha})\right)_+ > \frac{\epsilon}{2}.\right) < \infty \tag{27}$$

First consider (26). By assumption (11), there exists K such that $\mu(G^k_{\mathcal{G},\alpha}) - \mu^*_\alpha \leq \epsilon/2$ for all $k \geq K$. Let N be such that $k(n) \geq K$ for $n \geq N$. For any fixed $n \geq N$, consider a sample S of size n. By Theorem 1, it follows that with probability at least $1-\delta(n)$, $\mu(\widehat{G}_{\mathcal{G},\alpha}) - \mu^*_\alpha \leq \mu(G^k_{\mathcal{G},\alpha}) - \mu^*_\alpha \leq \epsilon/2$. Therefore

$$\begin{aligned}
&\sum_{n=1}^{\infty} P^n\left(\left(\mu(\widehat{G}_{\mathcal{G},\alpha}) - \mu^*_\alpha\right)_+ > \frac{\epsilon}{2}\right) \\
&= \sum_{n=1}^{N-1} P^n\left(\left(\mu(\widehat{G}_{\mathcal{G},\alpha}) - \mu^*_\alpha\right)_+ > \frac{\epsilon}{2}\right) + \sum_{n=N}^{\infty} P^n\left(\left(\mu(\widehat{G}_{\mathcal{G},\alpha}) - \mu^*_\alpha\right)_+ > \frac{\epsilon}{2}\right) \\
&\leq \sum_{n=1}^{N-1} P^n\left(\left(\mu(\widehat{G}_{\mathcal{G},\alpha}) - \mu^*_\alpha\right)_+ > \frac{\epsilon}{2}\right) + \sum_{n=N}^{\infty} \delta(n) \\
&< \infty.
\end{aligned}$$

The second inequality follows from the assumed summability of $\delta(n)$.

To establish (27), let N be large enough so that

$$\sup_{G \in \mathcal{G}_\alpha^{k(n)}} 2\phi_k(G, S, \delta(n)) \leq \frac{\epsilon}{2}$$

for all $n \geq N$. For any fixed $n \geq N$, consider a sample S of size n. By Theorem 1, it follows that with probability at least $1 - \delta(n)$, $\alpha - P(\widehat{G}_{\mathcal{G},\alpha}) \leq 2\phi_k(\widehat{G}_{\mathcal{G},\alpha}, S, \delta(n)) \leq \epsilon/2$. Therefore

$$\begin{aligned}
&\sum_{n=1}^{\infty} P^n\left(\left(\alpha - P(\widehat{G}_{\mathcal{G},\alpha})\right)_+ > \frac{\epsilon}{2}\right) \\
&= \sum_{n=1}^{N-1} P^n\left(\left(\alpha - P(\widehat{G}_{\mathcal{G},\alpha})\right)_+ > \frac{\epsilon}{2}\right) + \sum_{n=N}^{\infty} P^n\left(\left(\alpha - P(\widehat{G}_{\mathcal{G},\alpha})\right)_+ > \frac{\epsilon}{2}\right) \\
&\leq \sum_{n=1}^{N-1} P^n\left(\left(\alpha - P(\widehat{G}_{\mathcal{G},\alpha})\right)_+ > \frac{\epsilon}{2}\right) + \sum_{n=N}^{\infty} \delta(n) \\
&< \infty.
\end{aligned}$$

This completes the proof.

C Proof of Theorem 3

The first part of the theorem is straightforward. First, we claim that $(\mu(G_n) - \mu^*_\alpha)_+ \leq \mu(G_n \backslash G^*_\alpha)$. To see this, assume $\mu(G_n) - \mu^*_\alpha \geq 0$, otherwise the statement is trivial. Then

$$\begin{aligned}(\mu(G_n) - \mu^*_\alpha)_+ &= \mu(G_n) - \mu^*_\alpha \\ &= \mu(G_n) - \mu(G^*_\alpha) \\ &\leq \mu(G_n) - \mu(G^*_\alpha \cap G_n) \\ &= \mu(G_n \backslash G^*_\alpha).\end{aligned}$$

Similarly, one can show $(\alpha - P(G_n))_+ \leq P(G^*_\alpha \backslash G_n)$. Suppose f is bounded by B. Then $P(G^*_\alpha \backslash G_n) \leq B\mu(G^*_\alpha \backslash G_n)$. Putting everything together, we deduce $\mathcal{M}(G_n) \leq B(\mu(G_n \backslash G^*_\alpha) + \mu(G^*_\alpha \backslash G_n)) = B\mu(G_n \Delta G^*_\alpha)$ and the result follows.

Now for the second part of the theorem. From Section 1.2, we know $G^*_\alpha = \{x : f(x) = \gamma_\alpha\}$ where γ_α is the unique number such that $\int_{f(x) \geq \gamma_\alpha} f(x) d\mu(x) = \alpha$.

Consider the distribution Q of $(X, Y) \in \mathcal{X} \times \{0, 1\}$ given by the class-conditional distributions $X|Y = 0 \sim P$ and $X|Y = 1 \sim \mu$, and a priori class probabilities $Q(Y = 0) = p = 1 - Q(Y = 1)$, where p will be specified below. Then Q defines a classification problem. Let h^* denote a Bayes classifier with respect to Q (i.e., a classifier with minimum probability of error), and let $h : \mathcal{X} \to \{0, 1\}$ be an arbitrary classifier. The classification risk of h is defined as $\mathcal{R}(h) = Q(h(X) \neq Y)$, and the excess classification risk is $\mathcal{R}(h) - \mathcal{R}(h^*)$. From Bayes decision theory we know that h^* is the rule that compares the likelihood ratio to $p/(1-p)$. But, as discussed in Section 1.2, the likelihood ratio is $1/f$. Therefore, if p is such that $p/(1-p) = 1/\gamma_\alpha$, then $h^*(x) = 1 - \mathbb{I}(x \in G^*_\alpha)$ μ almost everywhere.

Setting $h_n(x) = 1 - \mathbb{I}(x \in G_n)$, we have

$$\begin{aligned}\mathcal{R}(h_n) - \mathcal{R}(h^*) &= Q(h_n(X) \neq Y) - Q(h^*(X) \neq Y) \\ &= (1-p)(\mu(h_n(X) = 0) - \mu(h^*(X) = 0))) + p(P(h_n(X) = 1) - P(h^*(X) = 0)) \\ &= (1-p)(\mu(G_n) - \mu(G^*_\alpha)) + p(1 - P(G_n) - (1 - P(G^*_\alpha))) \\ &= (1-p)(\mu(G_n) - \mu^*_\alpha) + p(\alpha - P(G_n)) \\ &\leq (\mu(G_n) - \mu^*_\alpha) + (\alpha - P(G_n)) \\ &\leq \mathcal{M}(G_n).\end{aligned}$$

Therefore $\mathcal{R}(h_n) \to \mathcal{R}(h^*)$. We now invoke a result of Steinwart et al. [2005] that says, in our notation, that $\mathcal{R}(h_n) \to \mathcal{R}(h^*)$ if and only if $\mu(G_n \Delta G^*_\alpha) \to 0$, and the proof is complete.

D Proof of Theorem 4

Let Ω_P be as in the proof of Theorem 1, and assume $S \in \overline{\Omega_P}$. This holds with probability at least $1 - \delta$. We consider three separate cases: (1) $\mu(\widehat{G}_{\mathcal{G},\alpha}) \geq \mu^*_\alpha$ and $P(\widehat{G}_{\mathcal{G},\alpha}) < \alpha$, (2) $\mu(\widehat{G}_{\mathcal{G},\alpha}) \geq \mu^*_\alpha$ and $P(\widehat{G}_{\mathcal{G},\alpha}) \geq \alpha$, and (3) $\mu(\widehat{G}_{\mathcal{G},\alpha}) < \mu^*_\alpha$ and $P(\widehat{G}_{\mathcal{G},\alpha}) < \alpha$. Note that the case in which both $\alpha \leq P(\widehat{G}_{\mathcal{G},\alpha})$ and $\mu(\widehat{G}_{\mathcal{G},\alpha}) < \mu^*_\alpha$ is impossible by definition of minimum volume sets. We will use the following fact:

Lemma 3. *If $S \in \overline{\Omega_P}$, then $\alpha - P(\widehat{G}_{\mathcal{G},\alpha}) \leq 2\phi(\widehat{G}_{\mathcal{G},\alpha}, S, \delta)$.*

The proof is a repetition of the proof that $\Theta_P \subset \Omega_P$ in Lemma 2.

For the first case we have

$$\begin{aligned}
\mathcal{M}(\widehat{G}_{\mathcal{G},\alpha}) &= \mu(\widehat{G}_{\mathcal{G},\alpha}) - \mu^*_\alpha + \alpha - P(\widehat{G}_{\mathcal{G},\alpha}) \\
&\leq \mu(\widehat{G}_{\mathcal{G},\alpha}) - \mu^*_\alpha + 2\phi(\widehat{G}_{\mathcal{G},\alpha}, S, \delta) \\
&= \inf_{G \in \widehat{\mathcal{G}}_\alpha} \left\{ \mu(G) - \mu^*_\alpha + 2\phi(G, S, \delta) \right\} \\
&\leq \inf_{G \in \mathcal{G}_\alpha} \left\{ \mu(G) - \mu^*_\alpha + 2\phi(G, S, \delta) \right\} \\
&\leq \left(1 + \frac{1}{\gamma_\alpha}\right) \inf_{G \in \mathcal{G}_\alpha} \left\{ \mu(G) - \mu^*_\alpha + 2\phi(G, S, \delta) \right\}.
\end{aligned}$$

The first inequality follows from $S \in \overline{\Theta_P}$. The next line comes from the definition of $\widehat{G}_{\mathcal{G},\alpha}$. The second inequality follows from $S \in \overline{\Omega_P}$, from which it follows that $\mathcal{G}_\alpha \subset \widehat{\mathcal{G}}_\alpha$. The final step is trivial (this constant is needed for case 3).

For the second case, $\mu(\widehat{G}_{\mathcal{G},\alpha}) \geq \mu^*_\alpha$ and $P(\widehat{G}_{\mathcal{G},\alpha}) \geq \alpha$, note

$$\begin{aligned}
\mathcal{M}(\widehat{G}_{\mathcal{G},\alpha}) &= \mu(\widehat{G}_{\mathcal{G},\alpha}) - \mu^*_\alpha \\
&\leq \mu(\widehat{G}_{\mathcal{G},\alpha}) - \mu^*_\alpha + 2\phi(\widehat{G}_{\mathcal{G},\alpha}, S, \delta)
\end{aligned}$$

and proceed as in the first case.

For the third case, $\mu(\widehat{G}_{\mathcal{G},\alpha}) < \mu^*_\alpha$ and $P(\widehat{G}_{\mathcal{G},\alpha}) < \alpha$, we rely on the following lemmas.

Lemma 4. *Let $\epsilon > 0$. Then*

$$\mu^*_\alpha - \mu^*_{\alpha-\epsilon} \leq \frac{\epsilon}{\gamma_\alpha}.$$

Proof. By assumptions **A1** and **A2**, there exist MV-sets $G^*_{\alpha-\epsilon}$ and G^*_α such that

$$\int_{G^*_\alpha} f(x) d\mu(x) = \alpha$$

and

$$\int_{G^*_{\alpha-\epsilon}} f(x) d\mu(x) = \alpha - \epsilon.$$

Furthermore, we may choose $G^*_{\alpha-\epsilon}$ and G^*_α such that $G^*_{\alpha-\epsilon} \subset G^*_\alpha$. Thus

$$\begin{aligned}
\epsilon &= \int_{G^*_\alpha} f(x) d\mu(x) - \int_{G^*_{\alpha-\epsilon}} f(x) d\mu(x) \\
&= \int_{G^*_\alpha \setminus G^*_{\alpha-\epsilon}} f(x) d\mu(x) \\
&\geq \gamma_\alpha \mu(G^*_\alpha \setminus G^*_{\alpha-\epsilon}) \\
&= \gamma_\alpha (\mu^*_\alpha - \mu^*_{\alpha-\epsilon})
\end{aligned}$$

and the result follows. □

Lemma 5. *If $S \in \overline{\Omega_P}$ and $G \in \widehat{\mathcal{G}}_\alpha$, then*

$$\mu^*_\alpha - \mu(G) \leq \frac{1}{\gamma_\alpha} \cdot 2\phi(G, S, \delta).$$

Proof. Denote $\epsilon = 2\phi(G, S, \delta)$. Since $S \in \overline{\Omega_P}$ and $G \in \widehat{\mathcal{G}}_\alpha$, we know

$$P(G) \geq \widehat{P}(G) - \frac{1}{2}\epsilon \geq \alpha - \epsilon.$$

In other words, $G \in \mathcal{G}_{\alpha-\epsilon}$. Therefore, $\mu(G) \geq \mu^*_{\alpha-\epsilon}$ and it suffices to bound $\mu^*_\alpha - \mu^*_{\alpha-\epsilon}$. Now apply the preceding lemma. □

It now follows that

$$\begin{aligned}
\mathcal{M}(\widehat{G}_{\mathcal{G},\alpha}) &= \alpha - P(\widehat{G}_{\mathcal{G},\alpha}) \\
&\leq 2\phi(\widehat{G}_{\mathcal{G},\alpha}, S, \delta) \\
&= \mu(\widehat{G}_{\mathcal{G},\alpha}) - \mu^*_\alpha + \mu^*_\alpha - \mu(\widehat{G}_{\mathcal{G},\alpha}) + 2\phi(\widehat{G}_{\mathcal{G},\alpha}, S, \delta) \\
&\leq \mu(\widehat{G}_{\mathcal{G},\alpha}) - \mu^*_\alpha + \left(1 + \frac{1}{\gamma_\alpha}\right) 2\phi(\widehat{G}_{\mathcal{G},\alpha}, S, \delta) \\
&\leq \left(1 + \frac{1}{\gamma_\alpha}\right) \left(\mu(\widehat{G}_{\mathcal{G},\alpha}) - \mu^*_\alpha + 2\phi(\widehat{G}_{\mathcal{G},\alpha}, S, \delta)\right) \\
&= \left(1 + \frac{1}{\gamma_\alpha}\right) \inf_{G \in \widehat{\mathcal{G}}_\alpha} \left\{ \mu(G) - \mu^*_\alpha + 2\phi(G, S, \delta) \right\} \\
&\leq \left(1 + \frac{1}{\gamma_\alpha}\right) \inf_{G \in \mathcal{G}_\alpha} \left\{ \mu(G) - \mu^*_\alpha + 2\phi(G, S, \delta) \right\}
\end{aligned}$$

The first inequality follows from Lemma 3. The second inequality is by Lemma 5. The next to last line follows from the definition of $\widehat{G}_{\mathcal{G},\alpha}$, and the final step is implied by $S \in \overline{\Omega_P}$ as in case 1. This completes the proof.

E Proof of Theorem 5

The proof follows an argument similar to the previous proof, but a few modifications are necessary. Let Ω_P be as in the proof of Theorem 1 and assume $S \in \overline{\Omega_P}$. This holds with probability at least $1 - \delta$. We consider three separate cases: (1) $\mu(\widehat{G}_{\mathcal{G},\alpha}) \geq \mu^*_\alpha$ and $P(\widehat{G}_{\mathcal{G},\alpha}) < \alpha$, (2) $\mu(\widehat{G}_{\mathcal{G},\alpha}) \geq \mu^*_\alpha$ and $P(\widehat{G}_{\mathcal{G},\alpha}) \geq \alpha$, and (3) $\mu(\widehat{G}_{\mathcal{G},\alpha}) < \mu^*_\alpha$ and $P(\widehat{G}_{\mathcal{G},\alpha}) < \alpha$. We employ the following variant of Lemma 3.

Lemma 6. *If $S \in \overline{\Omega_P}$, then $\alpha - P(\widehat{G}^\nu_{\mathcal{G},\alpha}) \leq (1 + \nu)\phi(\widehat{G}^\nu_{\mathcal{G},\alpha}, S, \delta)$.*

Proof. Suppose the conclusion is not true. Then

$$\begin{aligned}
P(\widehat{G}^\nu_{\mathcal{G},\alpha}) - \widehat{P}(\widehat{G}^\nu_{\mathcal{G},\alpha}) &< \alpha - (1 + \nu)\phi(\widehat{G}_{\mathcal{G},\alpha}, S, \delta) - \widehat{P}(\widehat{G}_{\mathcal{G},\alpha}) \\
&\leq -\phi(\widehat{G}_{\mathcal{G},\alpha}, S, \delta),
\end{aligned}$$

where the last inequality is true because $\widehat{P}(\widehat{G}_{\mathcal{G},\alpha}) \geq \alpha - \nu\phi(\widehat{G}_{\mathcal{G},\alpha}, S, \delta)$. But then $S \in \Omega_P$, contradicting the assumption. □

For the first case we have

$$\begin{aligned}
\mathcal{M}(\widehat{G}^{\nu}_{\mathcal{G},\alpha}) &= \mu(\widehat{G}^{\nu}_{\mathcal{G},\alpha}) - \mu^*_\alpha + \alpha - P(\widehat{G}^{\nu}_{\mathcal{G},\alpha}) \\
&\leq \mu(\widehat{G}^{\nu}_{\mathcal{G},\alpha}) - \mu^*_\alpha + (1+\nu)\phi(\widehat{G}^{\nu}_{\mathcal{G},\alpha}, S, \delta) \\
&= \min_{1\leq k\leq K}\left[\inf_{G\in\widehat{\mathcal{G}}^k_{\alpha(k,\nu)}}\{\mu(G) - \mu^*_\alpha + (1+\nu)\epsilon_k(n,\delta)\}\right] \\
&\leq \min_{1\leq k\leq K}\left[\inf_{G\in\mathcal{G}^k_{\alpha(k,\nu)}}\{\mu(G) - \mu^*_\alpha + (1+\nu)\epsilon_k(n,\delta)\}\right] \\
&= \min_{1\leq k\leq K}\left[\inf_{G\in\mathcal{G}^k_{\alpha(k,\nu)}}\left\{\mu(G) - \mu^*_{\alpha(k,\nu)} + \mu^*_{\alpha(k,\nu)} - \mu^*_\alpha + (1+\nu)\epsilon_k(n,\delta)\right\}\right] \\
&\leq \min_{1\leq k\leq K}\left[\inf_{G\in\mathcal{G}^k_{\alpha(k,\nu)}}\left\{\mu(G) - \mu^*_{\alpha(k,\nu)} + \frac{1}{\gamma_{\alpha(k,\nu)}}(1-\nu)\epsilon_k(n,\delta) + (1+\nu)\epsilon_k(n,\delta)\right\}\right] \\
&= \min_{1\leq k\leq K}\left[\inf_{G\in\mathcal{G}^k_{\alpha(k,\nu)}}\left\{\mu(G) - \mu^*_{\alpha(k,\nu)} + C_k\epsilon_k(n,\delta)\right\}\right] \\
&< \left(1+\frac{1}{\gamma_\alpha}\right)\min_{1\leq k\leq K}\left[\inf_{G\in\mathcal{G}^k_{\alpha(k,\nu)}}\left\{\mu(G) - \mu^*_{\alpha(k,\nu)} + C_k\epsilon_k(n,\delta)\right\}\right]
\end{aligned}$$

The first inequality follows from Lemma 6. The next line comes from the definition of $\widehat{G}^{\nu}_{\mathcal{G},\alpha}$. The second inequality follows from $S \in \overline{\Omega_P}$, from which it follows that $\mathcal{G}^k_{\alpha(k,\nu)} \subset \widehat{\mathcal{G}}^k_{\alpha(k,\nu)}$. The third inequality follows from Lemma 4. The final step is trivial (this constant is needed for case 3).

For the second case, $\mu(\widehat{G}^{\nu}_{\mathcal{G},\alpha}) \geq \mu^*_\alpha$ and $P(\widehat{G}^{\nu}_{\mathcal{G},\alpha}) \geq \alpha$, note

$$\begin{aligned}
\mathcal{M}(\widehat{G}^{\nu}_{\mathcal{G},\alpha}) &= \mu(\widehat{G}^{\nu}_{\mathcal{G},\alpha}) - \mu^*_\alpha \\
&\leq \mu(\widehat{G}^{\nu}_{\mathcal{G},\alpha}) - \mu^*_\alpha + 2\phi(\widehat{G}^{\nu}_{\mathcal{G},\alpha}, S, \delta)
\end{aligned}$$

and proceed as in the first case.

For the third case, $\mu(\widehat{G}_{\mathcal{G},\alpha}) < \mu^*_\alpha$ and $P(\widehat{G}_{\mathcal{G},\alpha}) < \alpha$, we need the following Lemma, which follows like Lemma 5 from Lemma 4.

Lemma 7. *If $S \in \overline{\Omega_P}$ and $G \in \widehat{\mathcal{G}}^k_{\alpha(k,\nu)}$, then*

$$\mu^*_\alpha - \mu(G) \leq \frac{1}{\gamma_\alpha}\cdot(1+\nu)\epsilon_k(n,\delta).$$

We now have

$$\begin{aligned}
\mathcal{M}(\widehat{G}^{\nu}_{\mathcal{G},\alpha}) &= \alpha - P(\widehat{G}^{\nu}_{\mathcal{G},\alpha}) \\
&\leq (1+\nu)\phi(\widehat{G}^{\nu}_{\mathcal{G},\alpha}, S, \delta) \\
&= \mu(\widehat{G}^{\nu}_{\mathcal{G},\alpha}) - \mu^*_\alpha + \mu^*_\alpha - \mu(\widehat{G}^{\nu}_{\mathcal{G},\alpha}) + (1+\nu)\phi(\widehat{G}_{\mathcal{G},\alpha}, S, \delta)
\end{aligned}$$

$$
\begin{aligned}
&\leq \mu(\widehat{G}^{\nu}_{\mathcal{G},\alpha}) - \mu^*_\alpha + \left(1 + \frac{1}{\gamma_\alpha}\right)(1+\nu)\phi(\widehat{G}^{\nu}_{\mathcal{G},\alpha}, S, \delta) \\
&\leq \left(1 + \frac{1}{\gamma_\alpha}\right)\left(\mu(\widehat{G}^{\nu}_{\mathcal{G},\alpha}) - \mu^*_\alpha + (1+\nu)\phi(\widehat{G}^{\nu}_{\mathcal{G},\alpha}, S, \delta)\right) \\
&= \left(1 + \frac{1}{\gamma_\alpha}\right) \min_{1\leq k\leq K}\left[\inf_{G\in\widehat{\mathcal{G}}^k_{\alpha(k,\nu)}} \{\mu(G) - \mu^*_\alpha + (1+\nu)\epsilon_k(n,\delta)\}\right].
\end{aligned}
$$

The first inequality follows from Lemma 6. The second inequality follows from Lemma 7. The last line follows from the definition of $\widehat{G}^{\nu}_{\mathcal{G},\alpha}$. Now bound the expression in square brackets as in case 1 above and the result follows.

F The Rademacher Penalty for Partition-Based Sets

In this appendix we show how the conditional Rademacher penalty introduced in Section 2.3 can be evaluated for a class $\mathcal{G}$ based on a fixed partition. The authors thank Gilles Blanchard for pointing out the properties that follow. Let $\pi = \{A_1, \ldots, A_k\}$ be a fixed, finite partition of $\mathcal{X}$, and let $\mathcal{G}$ be the set of all sets formed by taking the union of cells in π. Thus $|\mathcal{G}| = 2^k$ and every $G \in \mathcal{G}$ is specified by a k-length string of binary digits $\ell(A_1), \ldots, \ell(A_k)$, with $\ell(A) = 1$ if and only if $A \subset G$.

The conditional Rademacher average may be rewritten as follows:

$$
\begin{aligned}
\frac{2}{n}\mathbf{E}_{(\sigma_i)}\left[\sup_{G\in\mathcal{G}}\sum_{i=1}^n \sigma_i \mathbb{I}\left(X_i \in G\right)\right] &= \frac{2}{n}\mathbf{E}_{(\sigma_i)}\left[\sup_{\ell(A)\,:\,A\in\pi}\sum_{i=1}^n \sigma_i \ell(A)\right] \\
&= \frac{2}{n}\sum_{A\in\pi}\mathbf{E}_{(\sigma_i)}\left[\sup_{\ell(A)}\sum_{i:X_i\in A}\sigma_i\ell(A)\right] \\
&=: \sum_{A\in\pi}\psi(A).
\end{aligned}
$$

Thus the penalty is additive (modulo the delta term). Now consider a fixed cell A:

$$
\begin{aligned}
\psi(A) &= \frac{2}{n}\mathbf{E}_{(\sigma_i)}\left[\sup_{\ell(A)}\sum_{i:X_i\in A}\sigma_i\ell(A)\right] \\
&= \frac{1}{n}\mathbf{E}_{(\sigma_i)}\left[\sup_{\ell(A)}\sum_{i:X_i\in A}\sigma_i(2\ell(A)-1)\right] \\
&= \frac{1}{n}\mathbf{E}_{(\sigma_i)}\left[\sup_{\ell(A)}(2\ell(A)-1)\sum_{i:X_i\in A}\sigma_i\right] \\
&= \frac{1}{n}\mathbf{E}_{(\sigma_i)}\left[\left|\sum_{i:X_i\in A}\sigma_i\right|\right].
\end{aligned}
$$

Now let $\mathrm{bin}(M,p,m) = \binom{M}{m}p^m(1-p)^{M-m}$ be the probability of observing m successes in a sequence of M Bernoulli trials having success probability p. Then this last expression can be computed

explicitly as

$$\psi(A) = \frac{1}{n} \sum_{i=0}^{n_A} \mathrm{bin}(n_A, 1/2, i) |n_A - 2i|,$$

where $n_A = |\{i : X_i \in A\}|$. This is the penalty used in the histogram experiments (after the delta term is included).

A more convenient and intuitive penalty may be obtained by bounding

$$\begin{aligned}
\psi(A) &= \frac{1}{n} \mathbf{E}_{(\sigma_i)} \left[\left| \sum_{i:X_i \in A} \sigma_i \right| \right] \\
&\leq \frac{1}{n} \mathbf{E}_{(\sigma_i)} \left[\left(\sum_{i:X_i \in A} \sigma_i \right)^2 \right]^{\frac{1}{2}} \\
&= \frac{1}{n} \mathbf{E}_{(\sigma_i)} \left[\sum_{i:X_i \in A} \sigma_i^2 \right]^{\frac{1}{2}} \\
&= \sqrt{\frac{\widehat{P}(A)}{n}},
\end{aligned}$$

where the inequality is Jensen's. Moreover, by the Khinchin-Kahane inequality [see, e.g., Ledoux and Talagrand, 1991, Lemma 4.1], the converse inequality holds with a factor $\sqrt{2}$, so the bound is tight up to this factor. This is the "Rademacher" penalty employed in the dyadic decision tree experiments.

G Algorithm for MV-SRM with Dyadic Decision Trees

In this section we present an algorithm for MV-SRM over DDTs. We focus on the case where $\nu = 1$, although other cases may be treated similarly. Throughout this section we refer interchangeably to T and $G_T = \{x \in \mathcal{X} : T(x) = 1\}$. The algorithm we present applies to all penalties satisfying

P1 : $\phi(T)$ is additive, meaning it can be written as a sum over the leaves of T.

P2 : $\phi(T)$ does not depend on the labels assigned to the leaves of T.

Combining **P1** and **P2**, we are essentially considering penalties of the form

$$\phi(G_T) = \sum_{A \in \pi(T)} \psi(A),$$

for some function ψ on $\mathcal{A}^L$. The three penalties studied in Section 7 are of this form.

The algorithm for MV-SRM over DDTs is based on the following observations. First, there exists $0 < \alpha_1 < \ldots < \alpha_m = 1$ and DDTs $\widehat{G}_1, \ldots, \widehat{G}_m$ such that $\alpha \in (\alpha_{i-1}, \alpha_i] \Rightarrow \widehat{G}_\alpha = \widehat{G}_i$ (assume $\alpha_0 = 0$ always). To see this, consider Figure 12. This figure shows a hypothetical plot of points of the form $p(T) = (\widehat{P}(T) + \frac{1}{2}\phi(T), \mu(T) + \phi(T))$, $T \in \mathcal{T}^L$. For the value of α represented by the vertical dashed line, the set of feasible DDTs are those with $p(T)$ to the right of the line.

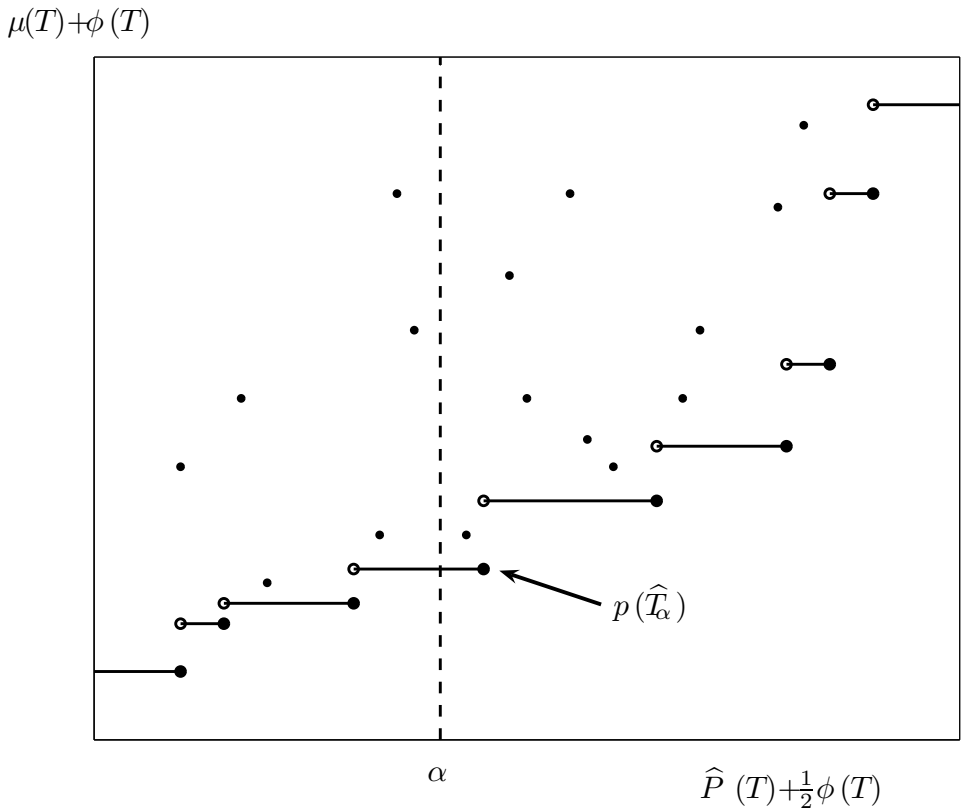

Figure 12: A hypothetical scatter plot of points in $\mathcal{P} = \{p(T) = (\widehat{P}(T) + \frac{1}{2}\phi(T), \mu(T)) \,|\, T \in \mathcal{T}^L\}$. Also shown (solid line) is the graph of $\mu(\widehat{G}_\alpha)$ as a function of $\alpha \in (0, 1)$. For the given value of α shown in the figure (at the vertical dashed line), the trees satisfying the constraint are those such that $p(T)$ is to the right of the line.

Also shown (solid line) is the graph of $\mu(\widehat{G}_\alpha) + \phi(\widehat{G}_\alpha)$ as a function of $\alpha \in (0, 1)$. Since $\mathcal{T}^L$ is finite, this geometric picture clearly shows that (i) $\mu(\widehat{G}_\alpha) + \phi(\widehat{G}_\alpha)$ is a piecewise constant, non-decreasing, and left-continuous function of α; (ii) the α_i are the discontinuities of this function; and (iii) $\alpha_i = \widehat{P}(\widehat{G}_i) + \frac{1}{2}\phi(\widehat{G}_i)$.

Furthermore, the above observation can be made at each cell $A \in \mathcal{A}_L$. Specifically, let MV-DDT(A,α) be the problem of solving MV-SRM over DDTs rooted at A and at the mass constraint α, and let $\widehat{G}_{A,\alpha}$ be the subtree rooted at A that solves this problem. Then there exists $0 < \alpha_{A,1} < \ldots < \alpha_{A,m(A)} = 1$ and DDTs $\widehat{G}_{A,1}, \ldots, \widehat{G}_{A,m(A)}$ such that $\alpha \in (\alpha_{A,i-1}, \alpha_{A,i}) \Rightarrow \widehat{G}_{A,\alpha} = \widehat{G}_{A,i}$.

The second key observation is that $\{\widehat{G}_{A,i}\}_{1 \le i \le m(A)}$ and $\{\alpha_{A,i}\}_{0 \le i \le m(A)}$ can be determined recursively. Suppose that $\widehat{G}_{A,\alpha} = \langle\langle \widehat{G}_{A'}, \widehat{G}_{A''} \rangle\rangle$, where A', A'' are children of A, and $\langle\langle T_1, T_2 \rangle\rangle$ denotes the subtree whose left and right branches are T_1 and T_2, respectively. Now set $\alpha' = \widehat{P}(\widehat{G}_{A'}) + \frac{1}{2}\phi(\widehat{G}_{A'})$ and $\alpha'' = \widehat{P}(\widehat{G}_{A''}) + \frac{1}{2}\phi(\widehat{G}_{A''})$. Then $\widehat{G}_{A'}$ is a solution of MV-DDT(A',α') and $\widehat{G}_{A''}$ is a solution of MV-DDT(A'',α''). Otherwise $\langle\langle \widehat{G}_{A',\alpha'}, \widehat{G}_{A'',\alpha''} \rangle\rangle$ would still satisfy the constraint for MV-DDT(A,α) but have a smaller volume, contradicting the optimality of $\widehat{G}_{A,\alpha}$. Here we have used the fact that μ, P, and ϕ are additive (**P1**). In conclusion, we have argued that for each A and each i, $1 \le i \le m(A)$,

there exists children A' and A'' of A such that $\widehat{G}_{A,i} = \langle\langle \widehat{G}_{A',r}, \widehat{G}_{A'',s} \rangle\rangle$ for some r, s.

This leads to a recursive algorithm for computing $\{\widehat{G}_{A,i}\}_{1 \le i \le m(A)}$ and $\{\alpha_{A,i}\}_{0 \le i \le m(A)}$. Start at the deepest possible cells, where computing these quantities is trivial. At a general cell A, compute all quantities of the form $\langle\langle \widehat{G}_{A',r}, \widehat{G}_{A'',s} \rangle\rangle$ where A', A'' range over all d possible children of A. For each such tree compute the volume, empirical probability, and penalty, which can be easily updated from lower levels because all of these are additive. Using these trees, plot the points $(\widehat{P}(T) + \frac{1}{2}\phi(T), \mu(T) + \phi(T))$ as shown in Figure 12, and determine the largest non-decreasing, piecewise constant, left-continuous function that falls below these points. The discontinuities of this function are the $\alpha_{A,i}$ and the right endpoint of each segment of the function corresponds to $\widehat{G}_{A,i}$.

One issue remains, however, before this approach becomes a practical algorithm. At first glance, the recursive procedure appears to involve visiting all $A \in \mathcal{A}^L$, a potentially huge number of cells. However, given a fixed training sample, most of those cells will be empty. Assume for now that if A is empty, then $\{\widehat{G}_{A,i}\}_{1 \le i \le m(A)}$ and $\{\alpha_{A,i}\}_{0 \le i \le m(A)}$ can be easily determined (we will see why this is true below). Then it is only necessary to perform the recursive update at nonempty cells. It can be shown that each training point intersects precisely $(L+1)^d$ cells in $\mathcal{A}^L$, and hence the total number of cells that need to be visited is $O(nL^d)$ [Blanchard et al., 2004]. Moreover, to take advantage of repeated computations, it pays to determine $\{\widehat{G}_{A,i}\}_{1 \le i \le m(A)}$ and $\{\alpha_{A,i}\}_{0 \le i \le m(A)}$ for all A at the same depth j, before proceeding to update cells at depth $j-1$.

The final piece of the puzzle is an efficient means of determining $\{\widehat{G}_{A,i}\}_{1 \le i \le m(A)}$ and $\{\alpha_{A,i}\}_{0 \le i \le m(A)}$ when A contains no data. Observe that if A contains no data, then for any $\alpha \in (0,1)$, all the leaves of $\widehat{G}_{A,\alpha}$ are given the label 1, and $\mu(\widehat{G}_{A,\alpha}) = 0$. That is, the minimum volume set for the subproblem MV-DDT(A,α) is the empty set. To see this, simply note that changing a label to 0 would not affect the constraint (by **P2**) and would increase the volume. Since all leaves of $\widehat{G}_{A,\alpha}$ are given the same label, and since DDTs in $\mathcal{T}^L$ are not allowed to have sibling leaf nodes with the same label, we conclude that $\widehat{G}_{A,\alpha}$ is the trivial tree consisting of the cell A and labeled 1.

The computational complexity of the algorithm for general ν is difficult to assess. For $\nu = 0$, however, we know that $\widehat{P}(G)$ can take on only the $n+1$ values $0, 1/n, 2/n, \ldots, 1$, and therefore we should see $m \le n+1$. The complexity of the algorithm in Blanchard et al. [2004] is $O(ndL^d)$, ignoring logarithmic factors. Their algorithm produces one update at each nonempty cell A, whereas ours (in the case of $\nu = 0$) requires $m(A) = O(n)$ updates, and hence the overall complexity is $O(n^2 dL^d)$.

References

P. Bartlett, S. Boucheron, and G. Lugosi. Model selection and error estimation. *Machine Learning*, 48:85–113, 2002.

S. Ben-David and M. Lindenbaum. Learning distributions by their density levels: a paradigm for learning without a teacher. *J. Comp. Sys. Sci.*, 55:171–182, 1997.

G. Blanchard, C. Schäfer, and Y. Rozenholc. Oracle bounds and exact algorithm for dyadic classification trees. In J. Shawe-Taylor and Y. Singer, editors, *Learning Theory: 17th Annual Conference on Learning Theory, COLT 2004*, pages 378–392. Springer-Verlag, Heidelberg, 2004.

O. Bousquet, S. Boucheron, and G. Lugosi. Introduction to statistical learning theory. In O. Bousquet, U.v. Luxburg, and G. Rtsch, editors, *Advanced Lectures in Machine Learning*, pages 169–207. Springer, 2004.

L. Breiman, J. Friedman, R. Olshen, and C. Stone. *Classification and Regression Trees*. Wadsworth, Belmont, CA, 1984.

A. Cannon, J. Howse, D. Hush, and C. Scovel. Learning with the Neyman-Pearson and min-max criteria. Technical Report LA-UR 02-2951, Los Alamos National Laboratory, 2002. URL `http://www.c3.lanl.gov/~kelly/ml/pubs/2002_minmax/paper.pdf`.

A. Cohen, W. Dahmen, I. Daubechies, and R. A. DeVore. Tree approximation and optimal encoding. *Applied and Computational Harmonic Analysis*, 11(2):192–226, 2001.

T. Cover and J. Thomas. *Elements of Information Theory*. John Wiley and Sons, New York, 1991.

A. Cuevas and A. Rodriguez-Casal. Set estimation: An overview and some recent developments. *Recent advances and trends in nonparametric statistics*, pages 251–264, 2003.

R. A. DeVore. Nonlinear approximation. *Acta Numerica*, 7:51–150, 1998.

L. Devroye, L. Györfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition*. Springer, New York, 1996.

D. Donoho. Wedgelets: Nearly minimax estimation of edges. *Ann. Stat.*, 27:859–897, 1999.

R. Durrett. *Probability: Theory and Examples*. Wadsworth & Brooks/Cole, Pacific Grove, CA, 1991.

J. Hartigan. Estimation of a convex density contour in two dimensions. *J. Amer. Statist. Assoc.*, 82(397):267–270, 1987.

X. Huo and J. Lu. A network flow approach in finding maximum likelihood estimate of high concentration regions. *Computational Statistics and Data Analysis*, 46(1):33–56, 2004.

V. Koltchinskii. Rademacher penalties and structural risk minimization. *IEEE Trans. Inform. Theory*, 47:1902–1914, 2001.

J. Langford. Tutorial on practical prediction theory for classification. *J. Machine Learning Research*, 6:273–306, 2005.

M. Ledoux and M. Talagrand. *Probability in Banach spaces*. Springer-Verlag, Berlin, 1991.

G. Lugosi and K. Zeger. Nonparametric estimation using empirical risk minimization. *IEEE Trans. Inform. Theory*, 41(3):677–687, 1995.

G. Lugosi and K. Zeger. Concept learning using complexity regularization. *IEEE Trans. Inform. Theory*, 42(1):48–54, 1996.

D. Müller and G Sawitzki. Excess mass estimates and tests for multimodality. *J. Amer. Statist. Assoc.*, 86(415):738–746, 1991.

J. Nunez-Garcia, Z. Kutalik, K.-H.Cho, and O. Wolkenhauer. Level sets and minimum volume sets of probability density functions. *Approximate Reasoning*, 34:25–47, Sept. 2003.

W. Polonik. Measuring mass concentrations and estimating density contour cluster–an excess mass approach. *Ann. Stat.*, 23(3):855–881, 1995.

W. Polonik. Minimum volume sets and generalized quantile processes. *Stochastic Processes and their Applications*, 69:1–24, 1997.

T. W. Sager. An iterative method for estimating a multivariate mode and isopleth. *J. Am. Stat. Asso.*, 74:329–339, 1979.

B. Schölkopf, J. Platt, J. Shawe-Taylor, A. Smola, and R. Williamson. Estimating the support of a high-dimensional distribution. *Neural Computation*, 13(7):1443–1472, 2001.

C. Scott and R. Nowak. Minimax optimal classification with dyadic decision trees. Technical Report TREE0403, Rice University, 2004. URL `http://www.stat.rice.edu/~cscott`.

C. Scott and R. Nowak. A Neyman-Pearson approach to statistical learning. *IEEE Trans. Inform. Theory*, 2005a. (in press).

C. Scott and R. Nowak. On the adaptive properties of decision trees. In Lawrence K. Saul, Yair Weiss, and Léon Bottou, editors, *Advances in Neural Information Processing Systems 17*, Cambridge, MA, 2005b. MIT Press.

I. Steinwart, D. Hush, and C. Scovel. A classification framework for anomaly detection. *J. Machine Learning Research*, 6:211–232, 2005.

A. B. Tsybakov. On nonparametric estimation of density level sets. *Ann. Stat.*, 25:948–969, 1997.

V. Vapnik. *Estimation of Dependencies Based on Empirical Data*. Springer-Verlag, New York, 1982.

V. Vapnik. *Statistical Learning Theory*. Wiley, New York, 1998.

R. Vert and J.-P. Vert. Consistency and convergence rates of one-class SVM and related algorithms. Technical Report 1414, Universit Paris-Sud, 2005.

G. Walther. Granulometric smoothing. *Ann. Stat.*, 25:2273–2299, 1997.