Minimax-Optimal Classification with Dyadic Decision Trees

Clayton Scott*, *Member, IEEE,* and Robert Nowak, *Senior Member, IEEE*

Abstract

Decision trees are among the most popular types of classifiers, with interpretability and ease of implementation being among their chief attributes. Despite the widespread use of decision trees, theoretical analysis of their performance has only begun to emerge in recent years. In this paper it is shown that a new family of decision trees, dyadic decision trees (DDTs), attain nearly optimal (in a minimax sense) rates of convergence for a broad range of classification problems. Furthermore, DDTs are surprisingly adaptive in three important respects: They automatically (1) *adapt* to favorable conditions near the Bayes decision boundary; (2) *focus* on data distributed on lower dimensional manifolds; and (3) *reject* irrelevant features. DDTs are constructed by penalized empirical risk minimization using a new data-dependent penalty and may be computed exactly with computational complexity that is nearly linear in the training sample size. DDTs are the first classifier known to achieve nearly optimal rates for the diverse class of distributions studied here while also being practical and implementable. This is also the first study (of which we are aware) to consider rates for adaptation to intrinsic data dimension and relevant features.

C. Scott is with the Department of Statistics, Rice University, 6100 S. Main, Houston, TX 77005; Vox: 713-348-2695; Fax: 713-348-5476; Email: cscott@rice.edu.

R. Nowak is with the Department of Electrical and Computer Engineering, University of Wisconsin, 1415 Engineering Drive, Madison, WI 53706; Vox: 608-265-3194; Fax: 608-262-1267; Email: nowak@engr.wisc.edu.

This work was supported by the National Science Foundation, Sponsored by the National Science Foundation, grant nos. CCR-0310889 and CCR-0325571, and the Office of Naval Research, grant no. N00014-00-1-0966.

Minimax-Optimal Classification with Dyadic Decision Trees

I. Introduction

Decision trees are among the most popular and widely applied approaches to classification. The hierarchical structure of decision trees makes them easy to interpret and implement. Fast algorithms for growing and pruning decision trees have been the subject of considerable study. Theoretical properties of decision trees including consistency and risk bounds have also been investigated. This paper investigates rates of convergence (to Bayes error) for decision trees, an issue that previously has been largely unexplored.

It is shown that a new class of decision trees called *dyadic decision trees* (DDTs) exhibit near-minimax optimal rates of convergence for a broad range of classification problems. In particular, DDTs are adaptive in several important respects:

Noise Adaptivity: DDTs are capable of automatically adapting to the (unknown) noise level in the neighborhood of the Bayes decision boundary. The noise level is captured by a condition similar to Tsybakov's noise condition [1].

Manifold Focus: When the distribution of features happens to have support on a lower dimensional manifold, DDTs can automatically detect and adapt their structure to the manifold. Thus decision trees learn the "intrinsic" data dimension.

Feature Rejection: If certain features are irrelevant (i.e., independent of the class label), then DDTs can automatically ignore these features. Thus decision trees learn the "relevant" data dimension.

Decision Boundary Adaptivity: If the Bayes decision boundary has γ derivatives, $0 < \gamma \leq 1$, DDTs can adapt to achieve faster rates for smoother boundaries. We consider only trees with axis-orthogonal splits. For more general trees such as perceptron trees, adapting to $\gamma > 1$ should be possible, although retaining implementability may be challenging.

Each of the above properties can be formalized and translated into a class of distributions with known minimax rates of convergence. Adaptivity is a highly desirable quality since in practice the precise characteristics of the distribution are unknown.

Dyadic decision trees are constructed by minimizing a complexity penalized empirical risk over an appropriate family of dyadic partitions. The penalty is data-dependent and comes from a new error

deviance bound for trees. This new bound is tailored specifically to DDTs and therefore involves substantially smaller constants than bounds derived in more general settings. The bound in turn leads to an oracle inequality from which rates of convergence are derived.

A key feature of our penalty is *spatial adaptivity*. Penalties based on standard complexity regularization (as represented by [2], [3], [4]) are proportional to the square root of the size of the tree (number of leaf nodes) and apparently fail to provide optimal rates [5]. In contrast, spatially adaptive penalties depend not only on the size of the tree, but also on the spatial distribution of training samples as well as the "shape" of the tree (e.g., deeper nodes incur a smaller penalty).

Our analysis involves bounding and balancing estimation and approximation errors. To bound the estimation error we apply well known concentration inequalities for sums of Bernoulli trials, most notably the relative Chernoff bound, in a spatially distributed and localized way. Moreover, these bounds hold for all sample sizes and are given in terms of explicit, small constants. 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 approximation [6], [7], [8]. The dyadic structure also leads to computationally tractable classifiers based on algorithms akin to fast wavelet and multiresolution transforms [9]. The computational complexity of DDTs is nearly linear in the training sample size. Optimal rates may be achieved by more general tree classifiers, but these require searches over prohibitively large families of partitions. DDTs are thus preferred because they are simultaneously implementable, analyzable, and sufficiently flexible to achieve optimal rates.

The paper is organized as follows. The remainder of the introduction sets notation and surveys related work. Section II defines dyadic decision trees. Section III presents risk bounds and an oracle inequality for DDTs. Section IV reviews the work of Mammen and Tsybakov [10] and Tsybakov [1] and defines regularity assumptions that help us quantify the four conditions outlined above. Section V presents theorems demonstrating the optimality (and adaptivity) of DDTs under these four conditions. Section VI discusses algorithmic and practical issues related to DDTs. Section VII offers conclusions and discusses directions for future research. The proofs are gathered in Section VIII.

A. Notation

Let Z be a random variable taking values in a set $\mathcal{Z}$, and let $Z^n = \{Z_1, \ldots, Z_n\} \in \mathcal{Z}^n$ be independent and identically distributed (IID) realizations of Z. Let $\mathbb{P}$ be the probability measure for Z, and let $\widehat{\mathbb{P}}_n$ be the empirical estimate of $\mathbb{P}$ based on Z^n: $\widehat{\mathbb{P}}_n(B) = (1/n)\sum_{i=1}^n \mathbb{I}_{\{Z_i \in B\}}$, $B \subseteq \mathcal{Z}$, where $\mathbb{I}$ denotes the indicator function. Let $\mathbb{P}^n$ denote the n-fold product measure on $\mathcal{Z}^n$ induced by $\mathbb{P}$. Let $\mathbb{E}$ and $\mathbb{E}^n$ denote

expectation with respect to $\mathbb{P}$ and $\mathbb{P}^n$, respectively.

In classification we take $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$, where $\mathcal{X}$ is the collection of feature vectors and $\mathcal{Y} = \{0, 1, \ldots, |\mathcal{Y}| - 1\}$ is a finite set of class labels. Let $\mathbb{P}_X$ and $\mathbb{P}_{Y|X}$ denote the marginal with respect to X and the conditional distribution of Y given X, respectively. In this paper we focus on binary classification ($|\mathcal{Y}| = 2$), although the results of Section III can be easily extended to the multi-class case.

A *classifier* is a measurable function $f : \mathcal{X} \to \mathcal{Y}$. Let $\mathcal{F}(\mathcal{X}, \mathcal{Y})$ be the set of all classifiers. Each $f \in \mathcal{F}(\mathcal{X}, \mathcal{Y})$ induces a set $B_f = \{(x, y) \in \mathcal{Z} \,|\, f(x) \neq y\}$. Define the probability of error and empirical error (risk) of f by $R(f) = \mathbb{P}(B_f)$ and $\widehat{R}_n(f) = \widehat{\mathbb{P}}_n(B_f)$, respectively. The *Bayes classifier* is the classifier f^* achieving minimum probability of error and is given by

$$f^*(x) = \arg\max_{y \in \mathcal{Y}} \mathbb{P}_{Y|X}(Y = y \,|\, X = x).$$

When $\mathcal{Y} = \{0, 1\}$ the Bayes classifier may be written

$$f^*(x) = \mathbb{I}_{\{\eta(x) \geq 1/2\}},$$

where $\eta(x) = \mathbb{P}(Y = 1 \,|\, X = x)$ is the a posteriori probability that the correct label is 1. The *Bayes risk* is $R(f^*)$ and denoted R^*. The *excess risk* of f is the difference $R(f) - R^*$. A *discrimination rule* is a measurable function $\widehat{f}_n : \mathcal{Z}^n \to \mathcal{F}(\mathcal{X}, \mathcal{Y})$.

The symbol $[\![\cdot]\!]$ will be used to denote the length of a binary encoding of its argument.

Additional notation is given at the beginning of Section IV.

B. Rates of Convergence in Classification

In this paper we study the rate at which the expected excess risk $\mathbb{E}^n\{R(\widehat{f}_n)\} - R^*$ goes to zero as $n \to \infty$ for $\widehat{f}_n$ based on dyadic decision trees. Marron [11] demonstrates minimax optimal rates under smoothness assumptions on the class-conditional densities. Yang [12] shows that for $\eta(x)$ in certain smoothness classes minimax optimal rates are achieved by appropriate plug-in rules. Both [11] and [12] place global constraints on the distribution, and in both cases optimal classification reduces to optimal density estimation. However, global smoothness assumptions can be overly restrictive for classification since high irregularity away from the Bayes decision boundary may have no effect on the difficulty of the problem.

Tsybakov and collaborators replace global constraints on the distribution by restrictions on η near the Bayes decision boundary. Faster minimax rates are then possible, although existing optimal discrimination rules typically rely on ϵ-nets for their construction and in general are not implementable [10], [1], [13].

Tsybakov and van de Geer [14] offer a more constructive approach using wavelets (essentially an explicit ϵ-net) but their discrimination rule is apparently still intractable and assumes the Bayes decision boundary is a boundary fragment (see Section IV). Other authors derive rates for existing practical discrimination rules, but these works are not comparable to ours, since different distributional assumptions or loss functions are considered. [15], [16], [17], [18], [19].

Our contribution is to demonstrate a discrimination rule that is not only practical and implementable, but also one that adaptively achieves nearly-minimax optimal rates for some of Tsybakov's and related classes. We further investigate issues of adapting to data dimension and rejecting irrelevant features, providing optimal rates in these settings as well. In an earlier paper, we studied rates for dyadic decision trees and demonstrated the (non-adaptive) near-minimax optimality of DDTs in a very special case of the general classes considered herein [20]. We also simplify and improve the bounding techniques used in that work. A more detailed review of rates of convergence is given in Section IV.

C. Decision Trees

In this section we review decision trees, focusing on learning-theoretic developments. For a multi-disciplinary survey of decision trees from a more experimental and heuristic viewpoint, see [21].

A *decision tree*, also known as a classification tree, is a classifier defined by a (usually binary) tree where each internal node is assigned a *predicate* (a "yes" or "no" question that can be asked of the data) and every terminal (or leaf) node is assigned a class label. Decision trees emerged over 20 years ago and flourished thanks in large part to the seminal works of Breiman, Friedman, Olshen and Stone [22] and Quinlan [23]. They have been widely used in practical applications owing to their interpretability and ease of use. Unlike many other techniques, decision trees are also easily constructed to handle discrete and categorical data, multiple classes, and missing values.

Decision tree construction is typically accomplished in two stages: growing and pruning. The growing stage consists of the recursive application of a greedy scheme for selecting predicates (or "splits") at internal nodes. This procedure continues until all training data are perfectly classified. A common approach to greedy split selection is to choose the split maximizing the decrease in "impurity" at the present node, where impurity is measured by a concave function such as entropy or the Gini index. Kearns and Mansour [24] demonstrate that greedy growing using impurity functions implicitly performs boosting. Unfortunately, as noted in [25, chap. 20], split selection using impurity functions cannot lead to consistent discrimination rules. For consistent growing schemes, see [26], [25].

In the pruning stage the output of the growing stage is "pruned back" to avoid overfitting. A variety

of pruning strategies have been proposed (see [27]). At least two groups of pruning methods have been the subject of recent theoretical studies. The first group involves a local, bottom-up algorithm, while the second involves minimizing a global penalization criterion. A representative of the former group is *pessimistic error pruning* employed by C4.5 [23]. In pessimistic pruning a single pass is made through the tree beginning with the leaf nodes and working up. At each internal node an optimal subtree is chosen to be either the node itself or the tree formed by merging the optimal subtrees of each child node. That decision is made by appealing to a heuristic estimate of the error probabilities induced by the two candidates. Both [28] and [29] modify pessimistic pruning to incorporate theoretically motivated local decision criteria, with the latter work proving risk bounds relating the pruned tree's performance to the performance of the best possible pruned tree.

The second kind of pruning criterion to have undergone theoretical scrutiny is especially relevant for the present work. It involves penalized empirical risk minimization (ERM) whereby the pruned tree $\widehat{T}_n$ is the solution of

$$\widehat{T}_n = \arg\min_{T \subset T_{\text{INIT}}} \quad \widehat{R}_n(T) + \Phi_n(T),$$

where T_{INIT} is the initial tree (from stage one) and $\Phi_n(T)$ is a *penalty* that in some sense measures the complexity of T. The most well known example is the *cost-complexity pruning* (CCP) strategy of [22]. In CCP $\Phi_n(T) = \alpha|T|$ where $\alpha > 0$ is a constant and $|T|$ is the number of leaf nodes, or *size*, of T. Such a penalty is advantageous because $\widehat{T}_n$ can be computed rapidly via a simple dynamic program. Despite its widespread use, theoretical justification outside the $R^* = 0$ case has been scarce. Only under a highly specialized "identifiability" assumption (similar to the Tsybakov noise condition in Section IV with $\kappa = 1$) have risk bounds been demonstrated for CCP [30].

Indeed, a penalty that scales linearly with tree size appears to be inappropriate under more general conditions. Mansour and McAllester [31] demonstrate error bounds for a "square root" penalty of the form $\Phi_n(T) = \alpha_n\sqrt{|T|}$. Nobel [32] considers a similar penalty and proves consistency of $\widehat{T}_n$ under certain assumptions on the initial tree produced by the growing stage. Scott and Nowak [5] also derive a square root penalty by applying structural risk minimization to dyadic decision trees.

Recently a few researchers have called into question the validity of basing penalties only on the size of the tree. Berkman and Sandholm [33] argue that any preference for a certain kind of tree implicitly makes prior assumptions on the data. For certain distributions, therefore, larger trees can be better than smaller ones with the same training error. Golea et al. [34] derive bounds in terms of the *effective size*, a quantity that can be substantially smaller than the true size when the training sample is non-uniformly

distributed across the leaves of the tree. Mansour and McAllester [31] introduce a penalty that can be significantly smaller than the square root penalty for unbalanced trees. In papers antecedent to the present work we show that the penalty of Mansour and McAllester can achieve an optimal rate of convergence (in a special case of the class of distributions studied here), while the square root penalty leads to suboptimal rates [5], [20].

The present work examines dyadic decision trees (defined in the next section). Our learning strategy involves penalized ERM, but there are no separate growing and pruning stages; split selection and complexity penalization are performed jointly. This promotes the learning of trees with *ancillary splits*, i.e., splits that don't separate the data but are the necessary ancestor of deeper splits that do. Such splits are missed by greedy growing schemes, which partially explains why DDTs outperform traditional tree classifiers [35] even though DDTs use a restricted set of splits.

We employ a *spatially adaptive, data-dependent* penalty. Like the penalty of [31], our penalty depends on more than just tree size and tends to favor unbalanced trees. In view of [33], our penalty reflects a prior disposition toward Bayes decision boundaries that are well-approximated by unbalanced recursive dyadic partitions.

D. Dyadic Thinking in Statistical Learning

Recursive dyadic partitions (RDPs) play a pivotal role in the present study. Consequently, there are strong connections between DDTs and wavelet and multiresolution methods, which also employ RDPs. For example, Donoho [36] establishes close connections between certain wavelet-based estimators and CART-like analyses. In this section, we briefly comment on the similarities and differences between wavelet methods in statistics and DDTs.

Wavelets have had a tremendous impact on the theory of nonparametric function estimation in recent years. Prior to wavelets, nonparametric methods in statistics were primarily used only by experts because of the complicated nature of their theory and application. Today, however, wavelet thresholding methods for signal denoising are in widespread use because of their ease of implementation, applicability, and broad theoretical foundations. The seminal papers of [37], [38], [39] initiated a flurry of research on wavelets in statistics, and [9] provides a wonderful account of wavelets in signal processing.

Their elegance and popularity aside, wavelet bases can be said to consist of essentially two key elements: a nested hierarchy of recursive, dyadic partitions; and an exact, efficient representation of smoothness. The first element allows one to localize isolated singularities in a concise manner. This is accomplished by repeatedly subdividing intervals or regions to increase resolution in the vicinity

of singularities. The second element then allows remaining smoothness to be concisely represented by polynomial approximations. For example, these two elements are combined in the work of [40], [41] to develop a *multiscale likelihood analysis* that provides a unified framework for wavelet-like modeling, analysis, and regression with data of continuous, count, and categorical types. In the context of classification, the target function to be learned is the Bayes decision rule, a piecewise constant function in which the Bayes decision boundary can be viewed as an "isolated singularity" separating totally smooth (constant) behavior.

Wavelet methods have been most successful in regression problems, especially for denoising signals and images. The orthogonality of wavelets plays a key role in this setting, since it leads to a simple independent sequence model in conjunction with Gaussian white noise removal. In classification problems, however, one usually makes few or no assumptions regarding the underlying data distribution. Consequently, the orthogonality of wavelets does not lead to a simple statistical representation, and therefore wavelets themselves are less natural in classification.

The dyadic partitions underlying wavelets are nonetheless tremendously useful since they can efficiently approximate piecewise constant functions. Johnstone [42] wisely anticipated the potential of dyadic partitions in other learning problems: "We may expect to see more use of 'dyadic thinking' in areas of statistics and data analysis that have little to do directly with wavelets." Our work reported here is a good example of his prediction (see also the recent work of [30], [40], [41]).

II. Dyadic Decision Trees

In this and subsequent sections we assume $\mathcal{X} = [0,1]^d$. We also replace the generic notation f for classifiers with T for decision trees. A *dyadic decision tree* (DDT) is a decision tree that divides the input space by means of axis-orthogonal dyadic splits. More precisely, a dyadic decision tree T is specified by assigning an integer $s(v) \in \{1, \ldots, d\}$ to each internal node v of T (corresponding to the coordinate that is split at that node), and a binary label 0 or 1 to each leaf node.

The nodes of DDTs correspond to hyperrectangles (cells) in $[0,1]^d$ (see Figure 1). Given a hyperrectangle $A = \prod_{r=1}^{d} [a_r, b_r]$, let $A^{s,1}$ and $A^{s,2}$ denote the hyperrectangles formed by splitting A at its midpoint along coordinate s. Specifically, define $A^{s,1} = \{x \in A \mid x^s \leq (a_s + b_s)/2\}$ and $A^{s,2} = A \backslash A^{s,1}$. Each node of a DDT 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 to the cell A, then the children of v are associated to $A^{s(v),1}$ and $A^{s(v),2}$.

Let $\pi(T) = \{A_1, \ldots, A_k\}$ denote the partition induced by T. Let $j(A)$ denote the depth of A and note

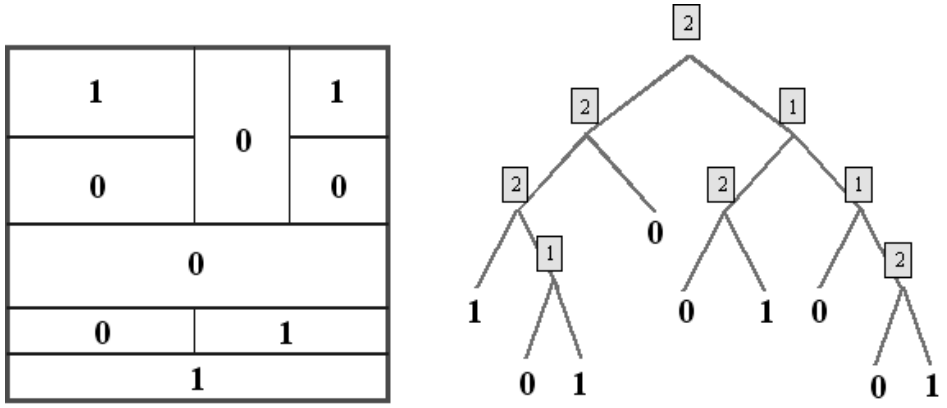


Fig. 1. A dyadic decision tree (right) with the associated recursive dyadic partition (left) when $d = 2$. Each internal node of the tree is labeled with an integer from 1 to d indicating the coordinate being split at that node. The leaf nodes are decorated with class labels.

that $\lambda(A) = 2^{-j(A)}$ where λ is the Lebesgue measure on $\mathbb{R}^d$. Define $\mathcal{T}$ to be the collection of all DDTs and $\mathcal{A}$ to be the collection of all cells corresponding to nodes of trees in $\mathcal{T}$.

Let M be a dyadic integer, that is, $M = 2^L$ for some nonnegative integer L. Define $\mathcal{T}_M$ to be the collection of all DDTs such that no terminal cell has a sidelength smaller than 2^{-L}. In other words, no coordinate is split more than L times when traversing a path from the root to a leaf. We will consider the discrimination rule

$$\widehat{T}_n = \arg\min_{T \in \mathcal{T}_M} \widehat{R}_n(T) + \Phi_n(T) \tag{1}$$

where Φ_n is a "penalty" or regularization term specified below in Equation (9). Computational and experimental aspects of this rule are discussed in Section VI.

A. Cyclic DDTs

In earlier work we considered as special class of DDTs called cyclic DDTs [5], [20]. In a cyclic DDT, $s(v) = 1$ when v is the root node, and $s(u) \equiv s(v) + 1 \pmod{d}$ for every parent-child pair (v, u). In other words, cyclic DDTs may be grown by cycling through the coordinates and splitting cells at the midpoint. Given the forced nature of the splits, cyclic DDTs will not be competitive with more general DDTs, especially when many irrelevant features are present. That said, cyclic DDTs still lead to optimal rates of convergence for the first two conditions outlined in the introduction. Furthermore, penalized ERM with cyclic DDTs is much simpler computationally (see Section VI-A).

III. Risk Bounds for Trees

In this section we introduce error deviance bounds and an oracle inequality for DDTs. The bounding techniques are quite general and can be extended to larger (even uncountable) families of trees (using VC theory, for example) but for the sake of simplicity and smaller constants we confine the discussion to DDTs. These risk bounds can also be easily extended to the case of multiple classes, but here we assume $\mathcal{Y} = \{0, 1\}$.

A. A square root penalty

We begin by recalling the derivation of a square root penalty which, although leading to suboptimal rates, helps motivate our spatially adaptive penalty. The following discussion follows [31] but traces back to [3]. Let $\mathcal{T}$ be a countable collection of trees and assign numbers $[\![T]\!]$ to each $T \in \mathcal{T}$ such that

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

In light of the Kraft inequality for prefix codes[1] [43], $[\![T]\!]$ may be defined as the codelength of a codeword for T in a prefix code for $\mathcal{T}$. Assume $[\![\overline{T}]\!] = [\![T]\!]$, where $\overline{T}$ is the complementary classifier, $\overline{T}(x) = 1 - T(x)$.

Proposition 1 *Let $\delta \in (0, 1]$. With probability at least $1 - \delta$ over the training sample,*

$$|R(T) - \widehat{R}_n(T)| \leq \sqrt{\frac{[\![T]\!] \log 2 + \log(1/\delta)}{2n}} \quad \textit{for all } T \in \mathcal{T}. \tag{2}$$

Proof: By the additive Chernoff bound (see Lemma 4), for any $T \in \mathcal{T}$ and $\delta_T \in (0, 1]$, we have

$$\mathbb{P}^n \left(R(T) - \widehat{R}_n(T) \geq \sqrt{\frac{\log(1/\delta_T)}{2n}} \right) \leq \delta_T.$$

Set $\delta_T = \delta 2^{-[\![T]\!]}$. By the union bound,

$$\mathbb{P}^n \left(\exists T \in \mathcal{T},\ R(T) - \widehat{R}_n(T) \geq \sqrt{\frac{[\![T]\!] \log 2 + \log(1/\delta)}{2n}} \right) \leq \sum_{T \in \mathcal{T}} \delta 2^{-[\![T]\!]} \leq \delta.$$

The reverse inequality follows from $\widehat{R}_n(T) - R(T) = R(\overline{T}) - \widehat{R}_n(\overline{T})$ and from $[\![\overline{T}]\!] = [\![T]\!]$. ■

We call the second term on the right-hand side of (2) the square root penalty. Similar bounds (with larger constants) may be derived using VC theory and structural risk minimization (see for example [32], [5]).

Codelengths for DDTs may be assigned as follows. Let $|T|$ denote the number of leaf nodes in T. Suppose $|T| = k$. Then $2k - 1$ bits are needed to encode the structure of T, and an additional k bits are

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

needed to encode the class labels of the leaves. Finally, we need $\log_2 d$ bits to encode the orientation of the splits at each internal node for a total of $[\![T]\!] = 3k - 1 + (k-1)\log_2 d$ bits. In summary, it is possible to construct a prefix code for $\mathcal{T}$ with $[\![T]\!] \leq (3 + \log_2 d)|T|$. Thus the square root penalty is proportional to the square root of tree size.

B. A spatially adaptive penalty

The square root penalty appears to suffer from slack in the union bound. Every node/cell can be a component of many different trees, but the bounding strategy in Proposition 1 does not take advantage of that redundancy. One possible way around this is to decompose the error deviance as

$$R(T) - \widehat{R}_n(T) = \sum_{A \in \pi(T)} R(T, A) - \widehat{R}_n(T, A), \tag{3}$$

where

$$R(T, A) = \mathbb{P}(T(X) \neq Y, X \in A)$$

and

$$\widehat{R}_n(T, A) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}_{\{T(X_i) \neq Y_i, X_i \in A\}}.$$

Since $n\widehat{R}_n(T, A) \sim \text{Binomial}(n, R(T, A))$, we may still apply standard concentration inequalities for sums of Bernoulli trials. This insight was taken from [31], although we employ a different strategy for bounding the "local deviance" $R(T, A) - \widehat{R}_n(T, A)$.

It turns out that applying the additive Chernoff bound to each term in (3) does not yield optimal rates of convergence. Instead, we employ the *relative* Chernoff bound (see Lemma 4) which implies that for any fixed cell $A \in \mathcal{A}$, with probability at least $1 - \delta$, we have

$$R(T, A) - \widehat{R}_n(T, A) \leq \sqrt{\frac{2p_A \log(1/\delta)}{n}}$$

where $p_A = \mathbb{P}_X(A)$. See the proof of Theorem 1 for details.

To obtain a bound of this form that holds uniformly for all $A \in \mathcal{A}$ we introduce a prefix code for $\mathcal{A}$. Suppose $A \in \mathcal{A}$ corresponds to a node v at depth j. Then $j + 1$ bits can encode the depth of v and j bits are needed to encode the direction (whether to branch "left" or "right") of the splits at each ancestor of v. Finally, an additional $\log_2 d$ bits are needed to encode the orientation of the splits at each ancestor of v, for a total of $[\![A]\!] = 2j + 1 + j \log_2 d$ bits. In summary, it is possible to define a prefix code for $\mathcal{A}$ with $[\![A]\!] \leq (3 + \log_2 d)j(A)$. With these definitions it follows that

$$\sum_{A \in \mathcal{A}} 2^{-[\![A]\!]} \leq 1. \tag{4}$$

Introduce the penalty

$$\Phi'_n(T) = \sum_{A \in \pi(T)} \sqrt{2p_A \frac{[\![A]\!] \log 2 + \log(2/\delta)}{n}}. \tag{5}$$

This penalty is spatially adaptive in the sense that different leaves are penalized differently depending on their depth, since $[\![A]\!] \propto j(A)$ and p_A is smaller for deeper nodes. Thus the penalty depends on the *shape* as well as the size of the tree. We have the following result.

Theorem 1 *With probability at least* $1 - \delta$*,*

$$|R(T) - \widehat{R}_n(T)| \leq \Phi'_n(T) \quad \textit{for all } T \in \mathcal{T}. \tag{6}$$

The proof may be found in Section VIII-A

The relative Chernoff bound allows for the introduction of local probabilities p_A to offset the additional cost of encoding a decision tree incurred by encoding each of its leaves individually. To see the implications of this, suppose the density of X is essentially bounded by c_0. If $A \in \mathcal{A}$ has depth j, then $p_A \leq c_0 \lambda(A) = c_0 2^{-j}$. Thus, while $[\![A]\!]$ increases at a linear rate as a function of j, p_A decays at an exponential rate.

From this discussion it follows that deep nodes contribute less to the spatially adaptive penalty than shallow nodes, and moreover, the penalty *favors unbalanced trees*. Intuitively, if two trees have the same size and empirical risk, minimizing the penalized empirical risk with the spatially adaptive penalty will select the tree that is more unbalanced, whereas a traditional penalty based only on tree size would not distinguish the two trees. This has advantages for classification because we expect unbalanced trees to approximate a $d-1$ (or lower) dimensional decision boundary well (see Figure 2).

Note that achieving spatial adaptivity in the classification setting is somewhat more complicated than in the case of regression. In regression, one normally considers a squared-error loss function. This leads naturally to a penalty that is proportional to the complexity of the model (e.g., squared estimation error grows linearly with degrees of freedom in a linear model). For regression trees, this results in a penalty proportional to tree size [44], [30], [40]. When the models under consideration consist of spatially localized components, as in the case of wavelet methods, then both the squared error and the complexity penalty can often be expressed as a sum of terms, each pertaining to a localized component of the overall model. Such models can be locally adapted to optimize the trade-off between bias and variance.

In classification a $0/1$ loss is used. Traditional estimation error bounds in this case give rise to penalties proportional to the square root of model size, as seen in the square root penalty above. While the (true

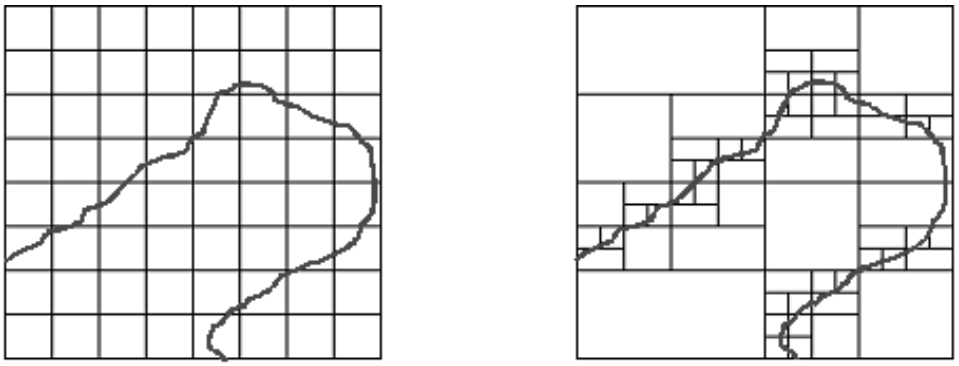


Fig. 2. An unbalanced tree/partition (right) can approximate a decision boundary much better than a balanced tree/partition (left) with the same number of leaf nodes. The suboptimal square root penalty penalizes these two trees equally, while the spatially adaptive penalty favors the unbalanced tree.

and empirical) risk functions in classification may be expressed as a sum over local components, it is no longer possible to easily separate the corresponding penalty terms since the total penalty is the square root of the sum of (what can be interpreted as) local penalties. Thus, the traditional error bounding methods lead to spatially non-separable penalties that inhibit spatial adaptivity. On the other hand, by first spatially decomposing the (true minus empirical) risk and then applying individual bounds to each term, we arrive at a spatially decomposed penalty that engenders spatial adaptivity in the classifier. An alternate approach to designing spatially adaptive classifiers, proposed in [14], is based on approximating the Bayes decision boundary (assumed to be a boundary fragment; see Section IV) with a wavelet series.

Remark 1 The decomposition in (3) is reminiscent of the *chaining* technique in empirical process theory (see [45]). In short, the chaining technique gives tail bounds for empirical processes by bounding the tail event using a "chain" of increasingly refined ϵ-nets. The bound for each ϵ-net in the chain is given in terms of its entropy number, and integrating over the chain gives rise to the so-call entropy integral bound. In our analysis, the cells at a certain depth may be thought of as comprising an ϵ-net, with the prefix code playing a role analogous to entropy. Since our analysis is specific to DDTs, the details differ, but the two approaches operate on similar principles.

Remark 2 In addition to different techniques for bounding the local error deviance, the bound of [31] differs from ours in another respect. Instead of distributing the error deviance over the leaves of T, one distributes the error deviance over some pruned subtree of T called a *root fragment*. The root fragment

is then optimized to yield the smallest bound. Our bound is a special case of this setup where the root fragment is the entire tree. It would be trivial to extend our bound to include root fragments, and this may indeed provide improved performance in practice. The resulting computational task would increase but still remain feasible. We have elected to not introduce root fragments because the penalty and associated algorithm are simpler and to emphasize that general root fragments are not necessary for our analysis.

C. A computable spatially adaptive penalty

The penalty introduced above has one major flaw: it is not computable, since the probabilities p_A depend on the unknown distribution. Fortunately, it is possible to bound p_A (with high probability) in terms of its empirical counterpart, and vice versa.

Recall $p_A = \mathbb{P}_X(A)$ and set $\hat{p}_A = (1/n)\sum_{i=1}^n \mathbb{I}_{\{X_i \in A\}}$. For $\delta \in (0,1]$ define

$$\hat{p}'_A(\delta) = 4\max\left(\hat{p}_A, \frac{[\![A]\!]\log 2 + \log(1/\delta)}{n}\right)$$

and

$$p'_A(\delta) = 4\max\left(p_A, \frac{[\![A]\!]\log 2 + \log(1/\delta)}{2n}\right).$$

Lemma 1 *Let $\delta \in (0,1]$. With probability at least $1-\delta$,*

$$p_A \le \hat{p}'_A(\delta) \quad \textit{for all } A \in \mathcal{A}, \tag{7}$$

and with probability at least $1-\delta$,

$$\hat{p}_A \le p'_A(\delta) \quad \textit{for all } A \in \mathcal{A} \tag{8}$$

We may now define a computable, data-dependent, spatially adaptive penalty by

$$\Phi_n(T) = \sum_{A \in \pi(T)} \sqrt{2\hat{p}'_A(\delta)\frac{[\![A]\!]\log 2 + \log(2/\delta)}{n}}. \tag{9}$$

Combining Theorem 1 and Lemma 1 produces the following.

Theorem 2 *Let $\delta \in (0,1]$. With probability at least $1-2\delta$,*

$$|R(T) - \widehat{R}_n(T)| \le \Phi_n(T) \quad \textit{for all } T \in \mathcal{T}. \tag{10}$$

Henceforth, this is the penalty we use to perform penalized ERM over $\mathcal{T}_M$.

Remark 3 From this point on, for concreteness and simplicity, we take $\delta = 1/n$ and omit the dependence of $\hat{p}'$ and p' on δ. Any choice of δ such that $\delta = O(\sqrt{\log n/n})$ and $\log(1/\delta) = O(\log n)$ would suffice.

D. An Oracle Inequality

Theorem 2 can be converted (using standard techniques) into an oracle inequality that plays a key role in deriving adaptive rates of convergence for DDTs.

Theorem 3 *Let $\widehat{T}_n$ be as in (1) with Φ_n as in (9). Define*

$$\tilde{\Phi}_n(T) = \sum_{A \in \pi(T)} \sqrt{8p'_A \frac{[\![A]\!] \log 2 + \log(2n)}{n}}.$$

With probability at least $1 - 3/n$ over the training sample

$$R(\widehat{T}_n) - R^* \leq \min_{T \in \mathcal{T}_M} \left(R(T) - R^* + 2\tilde{\Phi}_n(T) \right). \tag{11}$$

As a consequence,

$$\mathbb{E}^n\{R(\widehat{T}_n)\} - R^* \leq \min_{T \in \mathcal{T}_M} \left(R(T) - R^* + 2\tilde{\Phi}_n(T) \right) + \frac{3}{n}. \tag{12}$$

The proof is given in Section VIII-C. Note that these inequalities involve the uncomputable penalty $\tilde{\Phi}_n$. A similar theorem with Φ_n replacing $\tilde{\Phi}_n$ is also true, but the above formulation is more convenient for rate of convergence studies.

The expression $R(T) - R^*$ is the approximation error of T, while $\tilde{\Phi}_n(T)$ may be viewed as a bound on the estimation error $R(\widehat{T}_n) - R(T)$. These oracle inequalities say that $\widehat{T}_n$ finds a nearly optimal balance between these two quantities. For further discussion of oracle inequalities, see [46], [47].

IV. Rates of Convergence Under Complexity and Noise Assumptions

We study rates of convergence for classes of distributions inspired by the work of Mammen and Tsybakov [10] and Tsybakov [1]. Those authors examine classes that are indexed by a complexity exponent $\rho > 0$ that reflects the smoothness of the Bayes decision boundary, and a parameter κ that quantifies how "noisy" the distribution is near the Bayes decision boundary. Their choice of "noise assumption" is motivated by their interest in complexity classes with $\rho < 1$. For dyadic decision trees, however, we are concerned with classes having $\rho > 1$, for which a different noise condition is needed. The first two parts of this section review the work of [10] and [1], and the other parts propose new complexity and noise conditions pertinent to DDTs.

For the remainder of the paper assume $\mathcal{Y} = \{0, 1\}$ and $d \geq 2$. Classifiers and measurable subsets of $\mathcal{X}$ are in one-to-one correspondence. Let $\mathcal{G}(\mathcal{X})$ denote the set of all measurable subsets of $\mathcal{X}$ and identify each $f \in \mathcal{F}(\mathcal{X}, \mathcal{Y})$ with $G_f = \{x \in \mathcal{X} : f(x) = 1\} \in \mathcal{G}(\mathcal{X})$. The Bayes decision boundary, denoted ∂G^*, is the topological boundary of the Bayes decision set $G^* = G_{f^*}$. Note that while f^*,

G^*, and ∂G^* depend on the distribution $\mathbb{P}$, this dependence is not reflected in our notation. Given $G_1, G_2 \in \mathcal{G}(\mathcal{X})$, let $\Delta(G_1, G_2) = G_1 \backslash G_2 \cup G_2 \backslash G_1$ denote the symmetric difference. Similarly, define $\Delta(f_1, f_2) = \Delta(G_{f_1}, G_{f_2}) = \{x \in [0,1]^d \ : \ f_1(x) \neq f_2(x)\}$.

To denote rates of decay for integer sequences we write $a_n \preccurlyeq b_n$ if there exists $C > 0$ such that $a_n \leq C b_n$ for n sufficiently large. We write $a_n \asymp b_n$ if both $a_n \preccurlyeq b_n$ and $b_n \preccurlyeq a_n$. If g and h are functions of ϵ, $0 < \epsilon < 1$, we write $g(\epsilon) \preccurlyeq h(\epsilon)$ if there exists $C > 0$ such that $g(\epsilon) \leq C h(\epsilon)$ for ϵ sufficiently small, and $g(\epsilon) \asymp h(\epsilon)$ if $g(\epsilon) \preccurlyeq h(\epsilon)$ and $h(\epsilon) \preccurlyeq g(\epsilon)$.

A. Complexity Assumptions

Complexity assumptions restrict the complexity (regularity) of the Bayes decision boundary ∂G^*. Let $\bar{d}(\cdot, \cdot)$ be a pseudo-metric[2] on $\mathcal{G}(\mathcal{X})$ and let $\mathcal{G} \subseteq \mathcal{G}(\mathcal{X})$. We have in mind the case where $\bar{d}(G, G') = \mathbb{P}_X(\Delta(G, G'))$ and $\mathcal{G}$ is a collection of Bayes decision sets. Since we will be assuming $\mathbb{P}_X$ is essentially bounded with respect to Lebesgue measure λ, it will suffice to consider $\bar{d}(G, G') = \lambda(\Delta(G, G'))$.

Denote by $N(\epsilon, \mathcal{G}, \bar{d})$ the minimum cardinality of a set $\mathcal{G}' \subseteq \mathcal{G}$ such that for any $G \in \mathcal{G}$ there exists $G' \in \mathcal{G}'$ satisfying $\bar{d}(G, G') \leq \epsilon$. Define the *covering entropy* of $\mathcal{G}$ with respect to $\bar{d}$ to be $H(\epsilon, \mathcal{G}, \bar{d}) = \log N(\epsilon, \mathcal{G}, \bar{d})$. We say $\mathcal{G}$ has *covering complexity* $\rho > 0$ with respect to $\bar{d}$ if $H(\epsilon, \mathcal{G}, \bar{d}) \asymp \epsilon^{-\rho}$.

Mammen and Tsybakov [10] cite several examples of $\mathcal{G}$ with known complexities.[3] An important example for the present study is the class of boundary fragments, defined as follows. Let $\gamma > 0$, and take $r = \lceil \gamma \rceil - 1$ to be the largest integer strictly less than γ. Suppose $g : [0,1]^{d-1} \to [0,1]$ is r times differentiable, and let $p_{g,s}$ denote the r-th order Taylor polynomial of g at the point s. For a constant $c_1 > 0$, define $\Sigma(\gamma, c_1)$, the class of functions with Hölder regularity γ, to be the set of all g such that

$$|g(s') - p_{g,s}(s')| \leq c_1 |s - s'|^\gamma \text{ for all } s, s' \in [0,1]^{d-1}.$$

The set G is called a *boundary fragment* of smoothness γ if $G = \text{epi}(g)$ for some $g \in \Sigma(\gamma, c_1)$. Here $\text{epi}(g) = \{(s,t) \in [0,1]^d \ : \ g(s) \leq t\}$ is the epigraph of g. In other words, for a boundary fragment the last coordinate of ∂G^* is a Hölder-γ function of the first $d-1$ coordinates. Let $\mathcal{G}_{\text{BF}}(\gamma, c_1)$ denote the set of all boundary fragments of smoothness γ. Dudley [48] shows that $\mathcal{G}_{\text{BF}}(\gamma, c_1)$ has covering complexity $\rho \geq (d-1)/\gamma$ with respect to Lebesgue measure, with equality if $\gamma \geq 1$.

[2]A pseudo-metric satisfies the usual properties of metrics except $\bar{d}(x, y) = 0$ does not imply $x = y$.

[3]Although they are more interested in *bracketing* entropy rather than covering entropy.

B. Tsybakov's Noise Condition

Tsybakov also introduces what he calls a *margin* assumption (not to be confused with data-dependent notions of margin) that characterizes the level of "noise" near ∂G^* in terms of a noise exponent $\kappa, 1 \leq \kappa \leq \infty$. Fix $c_2 > 0$. A distribution satisfies *Tsybakov's noise condition* with noise exponent κ if

$$\mathbb{P}_X(\Delta(f, f^*)) \leq c_2(R(f) - R^*)^{1/\kappa} \text{ for all } f \in \mathcal{F}(\mathcal{X}, \mathcal{Y}). \tag{13}$$

This condition can be related to the "steepness" of the regression function $\eta(x)$ near the Bayes decision boundary. The case $\kappa = 1$ is the "low noise" case and implies a jump of $\eta(x)$ at the Bayes decision boundary. The case $\kappa = \infty$ is the high noise case and imposes no constraint on the distribution (provided $c_2 \geq 1$). It allows η to be arbitrarily flat at ∂G^*. See [1], [15], [13], [47] for further discussion.

A lower bound for classification under boundary fragment and noise assumptions is given by [10] and [1]. Fix $c_0, c_1, c_2 > 0$ and $1 \leq \kappa \leq \infty$. Define $\mathcal{D}_{\text{BF}}(\gamma, \kappa) = \mathcal{D}_{\text{BF}}(\gamma, \kappa, c_0, c_1, c_2)$ to be the set of all product measures $\mathbb{P}^n$ on $\mathcal{Z}^n$ such that

0A $\mathbb{P}_X(A) \leq c_0 \lambda(A)$ for all measurable $A \subseteq \mathcal{X}$

1A $G^* \in \mathcal{G}_{\text{BF}}(\gamma, c_1)$, where G^* is the Bayes decision set

2A for all $f \in \mathcal{F}(\mathcal{X}, \mathcal{Y})$

$$\mathbb{P}_X(\Delta(f, f^*)) \leq c_2(R(f) - R^*)^{1/\kappa}.$$

Theorem 4 (Mammen and Tsybakov) *Let $d \geq 2$. Then*

$$\inf_{\widehat{f}_n} \sup_{\mathcal{D}_{\text{BF}}(\gamma,\kappa)} \left[\mathbb{E}^n\{R(\widehat{f}_n)\} - R^*\right] \gtrsim n^{-\kappa/(2\kappa+\rho-1)}. \tag{14}$$

where $\rho = (d-1)/\gamma$.

The $\inf$ is over all discrimination rules $\widehat{f}_n : \mathcal{Z}^n \to \mathcal{F}(\mathcal{X}, \mathcal{Y})$ and the $\sup$ is over all $\mathbb{P}^n \in \mathcal{D}_{\text{BF}}(\gamma, \kappa)$.

Mammen and Tsybakov [10] demonstrate that empirical risk minimization (ERM) over $\mathcal{G}_{\text{BF}}(\gamma, c_1)$ yields a classifier achieving this rate when $\rho < 1$. Tsybakov [1] shows that ERM over a suitable "bracketing" net of $\mathcal{G}_{\text{BF}}(\gamma, c_1)$ also achieves the minimax rate for $\rho < 1$. Tsybakov and van de Geer [14] propose a minimum penalized empirical risk classifier (using wavelets, essentially a constructive ϵ-net) that achieves the minimax rate for all ρ (although a strengthened form of **2A** is required; see below). Audibert [13] recovers many of the above results using ϵ-nets and further develops rates under complexity and noise assumptions using PAC-Bayesian techniques. Unfortunately, none of these works provide computationally efficient algorithms for implementing the proposed discrimination rules, and it is unlikely that practical algorithms exist for these rules.

C. The Box-Counting Class

Boundary fragments fragments are theoretically convenient because approximation of boundary fragments reduces to approximation of Hölder regular functions, which can be accomplished constructively by means of wavelets or piecewise polynomials, for example. However, they are not realistic models of decision boundaries. A more realistic class would allow for decision boundaries with arbitrary orientation and multiple connected components. Dudley's classes [48] are a possibility, but constructive approximation of these sets is not well understood. Another option is the class of sets formed by intersecting a finite number of boundary fragments at different "orientations." We prefer instead to a different class that is ideally suited for constructive approximation by DDTs.

We propose a new complexity assumption that generalizes the set of boundary fragments with $\gamma = 1$ (Lipschitz regularity) to sets with arbitrary orientations, piecewise smoothness, and multiple connected components. Thus it is a more realistic assumption than boundary fragments for classification. Let m be a dyadic integer and let $\mathcal{P}_m$ denote the regular partition of $[0,1]^d$ into hypercubes of sidelength $1/m$. Let $N_m(G)$ be the number of cells in $\mathcal{P}_m$ that intersect ∂G. For $c_1 > 0$ define the *box-counting class*[4] $\mathcal{G}_{\text{BOX}}(c_1)$ to be the collection of all sets G such that $N_m(G) \leq c_1 m^{d-1}$ for all m. The following lemma implies $\mathcal{G}_{\text{BOX}}(c_1)$ has covering complexity $\rho \geq d-1$.

Lemma 2 *Boundary fragments with smoothness $\gamma = 1$ satisfy the box-counting assumption. In particular,*

$$\mathcal{G}_{\text{BF}}(1, c_1) \subseteq \mathcal{G}_{\text{BOX}}(c_1')$$

where $c_1' = (c_1\sqrt{d-1} + 2)$.

Proof: Suppose $G = \text{epi}(g)$, where $g \in \Sigma(1, c_1)$. Let S be a hypercube in $[0,1]^{d-1}$ with sidelength $1/m$. The maximum distance between points in S is $\sqrt{d-1}/m$. By the Hölder assumption, g deviates by at most $c_1\sqrt{d-1}/m$ over the cell S. Therefore, g passes through at most $(c_1\sqrt{d-1}+2)m^{d-1}$ cells in $\mathcal{P}_m$. ■

Combining Lemma 2 and Theorem 4 (with $\gamma = 1$ and $\kappa = 1$) gives a lower bound under **0A** and the condition

1B $G^* \in \mathcal{G}_{\text{BOX}}(c_1)$.

[4]The name is taken from the notion of box-counting dimension [49]. Roughly speaking, a set is in a box-counting class when it has box-counting dimension $d-1$. The box-counting dimension is an upper bound on the Hausdorff dimension, and the two dimensions are equal for most "reasonable" sets. For example, if ∂G^* is a smooth k-dimensional submanifold of $\mathbf{R}^d$, then ∂G^* has box-counting dimension k.

In particular we have

Corollary 1 *Let $\mathcal{D}_{\text{BOX}}(c_0, c_1)$ be the set of product measures $\mathbb{P}^n$ on $\mathcal{Z}^n$ such that* **0A** *and* **1B** *hold. Then*

$$\inf_{\widehat{f}_n} \sup_{\mathcal{D}_{\text{BOX}}(c_0,c_1)} \left[\mathbb{E}^n\{R(\widehat{f}_n)\} - R^*\right] \succcurlyeq n^{-1/d}.$$

D. Excluding Low Noise Levels

Recursive dyadic partitions can well-approximate G^* with smoothness $\gamma \leq 1$, and hence covering complexity $\rho \geq (d-1)/\gamma \geq 1$. However, Tsybakov's noise condition can only lead to faster rates of convergence when $\rho < 1$. To see this, note that the lower bound in (14) is tight only when $\rho < 1$. From the definition of $\mathcal{D}_{\text{BF}}(\gamma, \kappa)$ we have $\mathcal{D}_{\text{BF}}(\gamma, 1) \subset \mathcal{D}_{\text{BF}}(\gamma, \kappa)$ for any $\kappa > 1$. If $\rho > 1$, then as $\kappa \to \infty$, the right-hand side of (14) *decreases*. Therefore, the minimax rate for $\mathcal{D}_{\text{BF}}(\gamma, \kappa)$ can be no faster than $n^{-1/(1+\rho)}$, which is the lower bound for $\mathcal{D}_{\text{BF}}(\gamma, 1)$.

In light of the above, to achieve rates faster than $n^{-1/(1+\rho)}$ when $\rho > 1$, clearly an alternate assumption must be made. In fact, a "phase shift" occurs at $\rho = 1$. For $\rho < 1$, faster rates are obtained by excluding distributions with high noise. For $\rho > 1$, however, faster rates require the exclusion of distributions with *low* noise. This may seem at first to be counter-intuitive. Yet recall we are not interested in the rate of the Bayes risk, but of the *excess risk*. It so happens that for $\rho > 1$, the gap between actual and optimal risks is harder to close when the noise level is low.

E. Excluding Low Noise for the Box-Counting Class

We now introduce a new condition that excludes low noise under a concrete complexity assumption, namely, the box-counting assumption. This condition is an alternative to Tsybakov's noise condition, which by the previous discussion cannot yield faster rates for the box-counting class. As discussed below, our condition may be though of as the negation of Tsybakov's noise condition.

Before stating our noise assumption precisely we require additional notation. Fix $c_1 > 0$ and let m be a dyadic integer. Let $K_j(T)$ denote the number of nodes in T at depth j. Define $\mathcal{T}_m(c_1) = \{T \in \mathcal{T}_m : K_j(T) \leq 2c_1 2^{\lceil j/d \rceil (d-1)} \quad \forall j = 1, \ldots, d \log_2 m\}$. Note that when $j = d \log_2 m$, we have $c_1 2^{\lceil j/d \rceil (d-1)} = c_1 m^{d-1}$ which allows $\mathcal{T}_m(c_1)$ to approximate members of the box-counting class. When $j < d \log_2 m$ the condition $K_j(T) \leq 2c_1 2^{\lceil j/d \rceil (d-1)}$ ensures the trees in $\mathcal{T}_m(c_1)$ are *unbalanced*. As is shown in the proof of Theorem 6, this condition is sufficient to ensure that for all $T \in \mathcal{T}_m(c_1)$ (in particular the "oracle tree" T') the bound $\tilde{\Phi}_n(T)$ on estimation error decays at the desired rate. By the

following lemma, $\mathcal{T}_m(c_1)$ is also capable of approximating members of the box-counting class with error on the order of $1/m$.

Lemma 3 *For all $G \in \mathcal{G}_{\text{BOX}}(c_1)$, there exists $T \in \mathcal{T}_m(c_1)$ such that*

$$\lambda(\Delta(T, \mathbb{I}_G)) \leq \frac{c_1}{m}$$

where $\mathbb{I}_G$ is the indicator function on G.

See Section VIII-D for the proof. The lemma says that $\mathcal{T}_m(c_1)$ is an ϵ-net (with respect to Lebesgue measure) for $\mathcal{G}_{\text{BOX}}(c_1)$, with $\epsilon = c_1/m$.

Our condition for excluding low noise levels is defined as follows. Let $\mathcal{D}_{\text{BOX}}(\kappa) = \mathcal{D}_{\text{BOX}}(\kappa, c_0, c_1, c_2)$ be the set of all product measures $\mathbb{P}^n$ on $\mathcal{Z}^n$ such that

0A $\mathbb{P}_X(A) \leq c_0 \lambda(A)$ for all measurable $A \subseteq \mathcal{X}$

1B $G^* \in \mathcal{G}_{\text{BOX}}(c_1)$

2B For every dyadic integer m

$$(R(T_m^*) - R^*)^{1/\kappa} \leq \frac{c_2}{m}$$

where T_m^* minimizes $\lambda(\Delta(T, f^*))$ over $T \in \mathcal{T}_m(c_1)$.

In a sense **2B** is the negation of **2A**. Note that Lemma 3 and **0A** together imply that the approximation error satisfies $\mathbb{P}_X(\Delta(T_m^*, f^*)) \leq c_0 c_1/m$. Under Tsybakov's noise condition, whenever the approximation error is large, the excess risk to the power $1/\kappa$ is at least as large (up to some constant). Under our noise condition, whenever the approximation error is small, the excess risk to the $1/\kappa$ is at least as small. Said another way, Tsybakov's condition in (13) requires the excess risk to the $1/\kappa$ to be greater than the probability of the symmetric difference for all classifiers. Our condition entails the existence of at least one classifier for which that inequality is reversed. In particular, the inequality is reversed for the DDT that best approximates the Bayes classifier.

Remark 4 It would also suffice for our purposes to require **2B** to hold for all dyadic integers greater than some fixed m_0.

To illustrate condition **2B** we give the following example. For the time being suppose $d = 1$. Let $x_0 \in (0, 1)$ be fixed. Assume $\mathbb{P}_X = \lambda$ and $\eta(x) - 1/2 = |x - x_0|^{\kappa - 1}$ for $|x - x_0| \leq 1/m_0$, where m_0 is some fixed dyadic integer and $\kappa > 1$. Also assume $\eta(x) < 1/2$ for $x < x_0$ and $\eta(x) > 1/2$ for $x > x_0$,

so that $G^* = [x_0, 1]$. For $m \geq m_0$ let $[a_m, b_m)$ denote the dyadic interval of length $1/m$ containing x_0. Assume without loss of generality that x_0 is closer to b_m than a_m. Then $\Delta(T_m^*, f^*) = [x_0, b_m]$ and

$$\begin{aligned} R(T_m^*) - R^* &= \int_{x_0}^{b_m} 2|\eta(x) - 1/2|\, dx \\ &= \int_0^{b_m - x_0} 2x^{\kappa-1}\, dx \\ &= \frac{2}{\kappa}(b_m - x_0)^\kappa. \end{aligned}$$

If $c_2 \geq (2/\kappa)^{1/\kappa}$ then

$$(R(T_m^*) - R^*)^{1/\kappa} \leq c_2(b_m - x_0) \leq \frac{c_2}{m}$$

and hence **2B** is satisfied. This example may be extended to higher dimensions, for example, by replacing $|x - x_0|$ with the distance from x to ∂G^*.

We have the following lower bound for learning from $\mathcal{D}_{\text{BOX}}(\kappa)$.

Theorem 5 *Let $d \geq 2$. Then*

$$\inf_{\widehat{f}_n} \sup_{\mathcal{D}_{\text{BOX}}(\kappa)} \left[\mathbb{E}^n\{R(\widehat{f}_n)\} - R^*\right] \succcurlyeq n^{-\kappa/(2\kappa+d-2)}.$$

The proof relies on ideas from [13] and is given in Section VIII-E. In the next section the lower bound is seen to be tight (within a log factor). We conjecture a similar result holds under more general complexity assumptions ($\gamma \neq 1$), but that extends beyond the scope of this paper.

The authors of [14] and [13] introduce a different condition–what we call a *two-sided noise condition* –to exclude low noise levels. Namely, let $\mathcal{F}$ be a collection of candidate classifiers and $c_2 > 0$, and consider all distributions such that

$$\frac{1}{c_2}(R(f) - R^*)^{1/\kappa} \leq \mathbb{P}_X(\Delta(f, f^*)) \leq c_2(R(f) - R^*)^{1/\kappa} \text{ for all } f \in \mathcal{F}. \tag{15}$$

Such a condition does eliminate "low noise" distributions, but it also eliminates "high noise" (forcing the excess risk to behave very uniformly near ∂G^*), and is stronger than we need. In fact, it is not clear that (15) is ever necessary to achieve faster rates. When $\rho < 1$ the right-hand side determines the minimax rate, while for $\rho > 1$ the left-hand side is relevant.

Also note that this condition differs from ours in that we only require the first inequality to hold for classifiers that approximate f^* well. While [13] does prove a lower bound for the two-sided noise condition, that lower bound assumes a different $\mathcal{F}$ than his upper bounds. We believe our formulation is needed to produce lower and upper bounds that apply to the same class of distributions. Unlike Tsybakov's

or the two-sided noise conditions, it appears that the appropriate condition for properly excluding low noise must depend on the set of candidate classifiers.

V. Adaptive Rates for Dyadic Decision Trees

All of our rate of convergence proofs use the oracle inequality in the same basic way. The objective is to find an "oracle tree" $T' \in \mathcal{T}$ such that both $R(T') - R^*$ and $\tilde{\Phi}_n(T')$ decay at the desired rate. This tree is roughly constructed as follows. First form a "regular" dyadic partition (the exact construction will depend on the specific problem) into cells of sidelength $1/m$, for a certain $m \leq M$. Next "prune back" cells that do not intersect ∂G^*. Approximation and estimation error are then bounded using the given assumptions and elementary bounding techniques, and m is calibrated to achieve the desired rate.

This section consists of four subsections, one for each kind of adaptivity we consider. The first three make a box-counting complexity assumption and demonstrate adaptivity to low noise exclusion, intrinsic data dimension, and relevant features. The fourth subsection extends the complexity assumption to Bayes decision boundaries with smoothness $\gamma < 1$. While treating each kind of adaptivity separately allows us to simplify the discussion, all four conditions could be combined into a single result.

We also remark that, while our lower bounds assume dimension $d \geq 2$, the upper bounds (under the first three of the four studied conditions) apply for all $d \geq 1$. If $d = 1$, we get "fast" rates on the order of $\log n/n$, although these rates are not known to be optimal. This also applies when $d' = 1$ or $d'' = 1$, where d' and d'' are the intrinsic and relevant data dimensions, defined below.

A. Adapting to Noise Level

Dyadic decision trees, selected according to the penalized empirical risk criterion discussed earlier, adapt to achieve faster rates when low noise levels are not present. By Theorem 5, this rate is optimal (within a log factor).

Theorem 6 *Choose M such that $M \succcurlyeq (n/\log n)^{1/d}$. Define $\widehat{T}_n$ as in (1) with Φ_n as in (9). Then*

$$\sup_{\mathcal{D}_{\mathrm{BOX}}(\kappa)} \left[\mathbb{E}^n \{ R(\widehat{T}_n) \} - R^* \right] \preccurlyeq \left(\frac{\log n}{n} \right)^{\frac{\kappa}{2\kappa + d - 2}} . \tag{16}$$

The complexity penalized DDT $\widehat{T}_n$ is adaptive in the sense that it is constructed without knowledge of the noise exponent κ or the constants c_0, c_1, c_2. $\widehat{T}_n$ can always be constructed and in favorable circumstances the rate in (16) is achieved. See Section VIII-F for the proof.

B. When the Data Lie on a Manifold

In certain cases it may happen that the feature vectors lie on a manifold in the ambient space $\mathcal{X}$ (see Figure 3 (a)). When this happens, dyadic decision trees automatically adapt to achieve faster rates of convergence. To recast assumptions **0A** and **1B** in terms of a data manifold we again use box-counting ideas. Let $c_0, c_1 > 0$ and $1 \leq d' \leq d$. Recall $\mathcal{P}_m$ denotes the regular partition of $[0,1]^d$ into hypercubes of sidelength $1/m$ and $N_m(G)$ is the number of cells in $\mathcal{P}_m$ that intersect ∂G. The boundedness and complexity assumptions for a d' dimensional manifold are given by

0B For all dyadic integers m and all $A \in \mathcal{P}_m$, $\mathbb{P}_X(A) \leq c_0 m^{-d'}$.

1C For all dyadic integers m, $N_m(G^*) \leq c_1 m^{d'-1}$.

We refer to d' as the *intrinsic data dimension*. In practice, it may be more likely that data "almost" lie on a d'-dimensional manifold. Nonetheless, we believe the adaptivity of DDTs to data dimension depicted in the theorem below reflects a similar capability in less ideal settings.

Let $\mathcal{D}'_{\text{BOX}} = \mathcal{D}'_{\text{BOX}}(c_0, c_1, d')$ be the set of all product measures $\mathbb{P}^n$ on $\mathcal{Z}^n$ such that **0B** and **1C** hold.

Proposition 2 *Let $d' \geq 2$. Then*

$$\inf_{\widehat{f}_n} \sup_{\mathcal{D}'_{\text{BOX}}} \left[\mathbb{E}^n \{ R(\widehat{f}_n) \} - R^* \right] \succcurlyeq n^{-1/d'}.$$

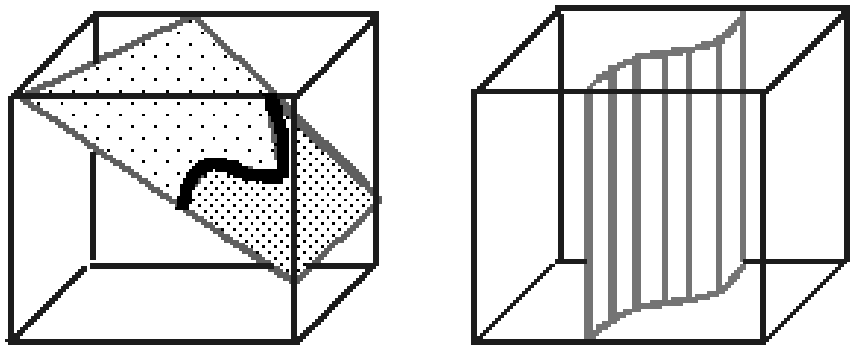
Proof: Assume $Z' = (X', Y')$ satisfies **0A** and **1B** in $[0,1]^{d'}$. Consider the mapping of features $X' = (X^1, \ldots, X^{d'}) \in [0,1]^{d'}$ to $X = (X^1, \ldots, X^{d'}, \zeta, \ldots, \zeta) \in [0,1]^d$, where $\zeta \in [0,1]$ is any non-dyadic rational number. (We disallow dyadic rationals to avoid potential ambiguities in how boxes are counted.) Then $Z = (X, Y')$ satisfies **0B** and **1C** in $[0,1]^d$. Clearly there can be no discrimination rule achieving a rate faster than $n^{-1/d'}$ uniformly over all such Z, as this would lead to a discrimination rule outperforming the minimax rate for Z' given in Corollary 1. ■

Dyadic decision trees can achieve this rate to within a log factor.

Theorem 7 *Choose M such that $M \succcurlyeq n/\log n$. Define $\widehat{T}_n$ as in (1) with Φ_n as in (9). Then*

$$\sup_{\mathcal{D}'_{\text{BOX}}} \left[\mathbb{E}^n \{ R(\widehat{T}_n) \} - R^* \right] \preccurlyeq \left(\frac{\log n}{n} \right)^{\frac{1}{d'}}. \tag{17}$$

Again, $\widehat{T}_n$ is adaptive in that it does not require knowledge of the intrinsic dimension d' or the constants c_0, c_1. The proof may be found in Section VIII-G.



(a) (b)

Fig. 3. Cartoons illustrating intrinsic and relevant dimension. (a) When the data lies on a manifold with dimension $d' < d$, then the Bayes decision boundary has dimension $d' - 1$. Here $d = 3$ and $d' = 2$. (b) If the X^3 axis is irrelevant, then the Bayes decision boundary is a "vertical sheet" over a curve in the (X^1, X^2) plane.

C. Irrelevant Features

We define the *relevant data dimension* to be the number $d'' \leq d$ of features X^i that are not statistically independent of Y. For example, if $d = 2$ and $d'' = 1$, then ∂G^* is a horizontal or vertical line segment (or union of such line segments). If $d = 3$ and $d'' = 1$, then ∂G^* 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^* is a "vertical sheet" over a curve in the (X^1, X^2) plane (see Figure 3 (b)).

Let $\mathcal{D}''_{\text{BOX}} = \mathcal{D}''_{\text{BOX}}(c_0, c_1, d'')$ be the set of all product measures $\mathbb{P}^n$ on $\mathcal{Z}^n$ such that **0A** and **1B** hold and Z has relevant data dimension d''.

Proposition 3 *Let $d'' \geq 2$. Then*

$$\inf_{\widehat{f}_n} \sup_{\mathcal{D}''_{\text{BOX}}} \left[\mathbb{E}^n \{ R(\widehat{f}_n) \} - R^* \right] \succcurlyeq n^{-1/d''}.$$

Proof: Assume $Z'' = (X'', Y'')$ satisfies **0A** and **1B** in $[0,1]^{d''}$. Consider the mapping of features $X'' = (X^1, \ldots, X^{d''}) \in [0,1]^{d''}$ to $X = (X^1, \ldots, X^{d''}, X^{d''+1}, \ldots, X^d) \in [0,1]^d$, where $X^{d''+1}, \ldots, X^d$ are independent of Y. Then $Z = (X, Y'')$ satisfies **0A** and **1B** in $[0,1]^d$ and has relevant data dimension (at most) d''. Clearly there can be no discrimination rule achieving a rate faster than $n^{-1/d''}$ uniformly over all such Z, as this would lead to a discrimination rule outperforming the minimax rate for Z'' given in Corollary 1. ■

Dyadic decision trees can achieve this rate to within a log factor.

Theorem 8 *Choose M such that $M \succcurlyeq n/\log n$. Define $\widehat{T}_n$ as in (1) with Φ_n as in (9). Then*

$$\sup_{\mathcal{D}''_{\text{BOX}}} \left[\mathbb{E}^n \{ R(\widehat{T}_n) \} - R^* \right] \preccurlyeq \left(\frac{\log n}{n} \right)^{\frac{1}{d''}} . \tag{18}$$

As in the previous theorems, our discrimination rule is adaptive in the sense that it does not need to be told c_0, c_1, d'' or which d'' features are relevant. While the theorem does not captures degrees of relevance, we believe it captures the essence of DDTs' feature rejection capability.

Finally, we remark that even if all features are relevant, but the Bayes rule still only depends on $d'' < d$ features, DDTs are still adaptive and decay at the rate given in the previous theorem.

D. Adapting to Bayes Decision Boundary Smoothness

Thus far in this section we have assumed G^* satisfies a box-counting (or related) condition, which essentially includes all ∂G^* with Lipschitz smoothness. When $\gamma < 1$, DDTs can still adaptively attain the minimax rate (within a log factor). Let $\mathcal{D}_{\text{BF}}(\gamma) = \mathcal{D}_{\text{BF}}(\gamma, c_0, c_1)$ denote the set of product measures satisfying **0A** and

1D One coordinate of ∂G^* is a function of the others, where the function has Hölder regularity γ and constant c_1.

Note that **1D** implies G^* is a boundary fragment but with arbitrary "orientation" (which coordinate is a function of the others). It is possible to relax this condition to more general ∂G^* (piecewise Hölder boundaries with multiple connected components) using box-counting ideas (for example), although for we do not pursue this here. Even without this generalization, when compared to [14] DDTs have the advantage (in addition of being implementable) that it is not necessary to know the orientation of ∂G^*, or which side of ∂G^* corresponds to class 1.

Theorem 9 *Choose M such that $M \succcurlyeq (n/\log n)^{1/(d-1)}$. Define $\widehat{T}_n$ as in (1) with Φ_n as in (9). If $d \geq 2$ and $\gamma \leq 1$, then*

$$\sup_{\mathcal{D}_{\text{BF}}(\gamma)} \left[\mathbb{E}^n \{ R(\widehat{T}_n) \} - R^* \right] \preccurlyeq \left(\frac{\log n}{n} \right)^{\frac{\gamma}{\gamma+d-1}} . \tag{19}$$

By Theorem 4 (with $\kappa = 1$) this rate is optimal (within a log factor). The problem of finding practical discrimination rules that adapt to the optimal rate for $\gamma > 1$ is an open problem we are currently pursuing.

VI. Computational Considerations

The data-dependent, spatially adaptive penalty in (9) is additive, meaning it is the sum over its leaves of a certain functional. Additivity of the penalty allows for fast algorithms for constructing $\widehat{T}_n$ when

combined with the fact that most cells contain no data. Indeed, Blanchard et al. [30] show that an algorithm of [36], simplified by a data sparsity argument, may be used to compute $\widehat{T}_n$ in $O(ndL^d \log(nL^d))$ operations, where $L = \log_2 M$ is the maximum number of dyadic refinements along any coordinate. Our theorems on rates of convergence are satisfied by $L \asymp O(\log n)$ in which case the complexity is $O(nd(\log n)^{d+1})$.

For completeness we restate the algorithm, which relies on two key observations. Some notation is needed. Let $\mathcal{A}_M$ be the set of all cells corresponding to nodes of trees in $\mathcal{T}_M$. In other words $\mathcal{A}_M$ is the set of cells obtained by applying no more than $L = \log_2 M$ dyadic splits along each coordinate. For $A \in \mathcal{A}_M$, let T_A denote a subtree rooted at A, and let T_A^* denote the subtree T_A minimizing $\widehat{R}_n(T_A) + \Phi_n(T_A)$, where

$$\widehat{R}_n(T_A) = \frac{1}{n} \sum_{i: X_i \in A} \mathbb{I}_{\{T_A(X_i) \neq Y_i\}}.$$

Recall that $A^{s,1}$ and $A^{s,2}$ denote the children of A when split along coordinate s. If T_1 and T_2 are trees rooted at $A^{s,1}$ and $A^{s,2}$, respectively, denote by $\textsc{merge}(A, T_1, T_2)$ the tree rooted at A having T_1 and T_2 as its left and right branches.

The first key observation is that

$$T_A^* = \arg\min\{\widehat{R}_n(T_A) + \Phi_n(T_A) \mid T_A = \{A\} \text{ or } T_A = \textsc{merge}(A, T_{A^{s,1}}^*, T_{A^{s,2}}^*), s = 1, \ldots, d\}.$$

In other words, the optimal tree rooted at A is either the tree consisting only of A or the tree formed by merging the optimal trees from one of the d possible pairs of children of A. This follows by additivity of the empirical risk and penalty, and leads to a recursive procedure for computing $\widehat{T}_n$. Note that this algorithm is simply a high dimensional analogue of the algorithm of [36] for "dyadic CART" applied to images. The second key observation is that it is not necessary to visit all possible nodes in $\mathcal{A}_M$ because most of them contain no training data (in which case T_A^* is the cell A itself).

Although we are primarily concerned with theoretical properties of DDTs, we note that a recent experimental study [35] demonstrates that DDTs are indeed competitive with state-of-the-art kernel methods while retaining the interpretability of decision trees and outperforming C4.5 on a variety of datasets. The primary drawback of DDTs in practice is the exponential dependence of computational complexity on dimension. When $d > 15$ memory and processor limitations can necessitate heuristic searches or preprocessing in the form of dimensionality reduction [50].

A. Cyclic DDTs

An inspection of their proofs reveals that Theorems 6 and 7 (noise and manifold conditions) hold for cyclic DDTs as well. From a computational point of view, moreover, learning with cyclic DDTs (see Section II-A) is substantially easier. The optimization in (1) reduces to pruning the (unique) cyclic DDT with all leaf nodes at maximum depth. However, many of those leaf nodes will contain no training data, and thus it suffices to prune the tree T_{INIT} constructed as follows: cycle through the coordinates and split (at the midpoint) only those cells that contain data from both classes. T_{INIT} will have at most n non-empty leaves, and every node in T_{INIT} will be an ancestor of such nodes, or one of their children. Each leaf node with data has at most dL ancestors, so T_{INIT} has $O(ndL)$ nodes. Pruning T_{INIT} may be solved via a simple bottom-up tree-pruning algorithm in $O(ndL)$ operations. Our theorems are satisfied by $L \asymp O(\log n)$ in which case the complexity is $O(nd \log n)$.

VII. Conclusions

This paper reports on a new class of decision trees known as dyadic decision trees (DDTs). It establishes four adaptivity properties of DDTs and demonstrates how these properties lead to near minimax optimal rates of convergence for a broad range of pattern classification problems. Specifically, it is shown that DDTs automatically *adapt* to noise and complexity characteristics in the neighborhood of the Bayes decision boundary, *focus* on the manifold containing the training data, which may be lower dimensional than the extrinsic dimension of the feature space, and detect and *reject* irrelevant features.

Although we treat each kind of adaptivity separately for the sake of exposition, there does exist a single classification rule that adapts to all four conditions simultaneously. Specifically, if the resolution parameter M is such that $M \succcurlyeq n/\log n$, and $\widehat{T}_n$ is obtained by penalized empirical risk minimization (using the penalty in (9)) over all DDTs up to resolution M, then

$$\mathbb{E}^n\{R(\widehat{T}_n)\} - R^* \preccurlyeq \left(\frac{\log n}{n}\right)^{\frac{\kappa}{2\kappa+\rho^*-1}}$$

where κ is the noise exponent, $\rho^* = (d^* - 1)/\gamma$, $\gamma \leq 1$ is the Bayes decision boundary smoothness, and d^* is the dimension of the manifold supporting the relevant features.

Two key ingredients in our analysis are a family of classifiers based on recursive dyadic partitions (RDPs) and a novel data-dependent penalty which work together to produce the near optimal rates. By considering RDPs we are able to leverage recent insights from nonlinear approximation theory and multiresolution analysis. RDPs are optimal, in the sense of nonlinear m-term approximation theory, for approximating certain classes of decision boundaries. They are also well suited for approximating low

dimensional manifolds and ignoring irrelevant features. Note that the optimality of DDTs for these two conditions should translate to similar results in density estimation and regression.

The data-dependent penalty favors the unbalanced tree structures that correspond to the optimal approximations to decision boundaries. Furthermore, the penalty is additive, leading to a computationally efficient algorithm. Thus DDTs are the first known practical classifier to attain optimal rates for the broad class of distributions studied here.

An interesting aspect of the new penalty and risk bounds is that they demonstrate the importance of *spatial adaptivity* in classification, a property that has recently revolutionized the theory of nonparametric regression with the advent of wavelets. In the context of classification, the spatial decomposition of the error leads to the new penalty that permits trees of arbitrary depth and size, provided that the bulk of the leaves correspond to "tiny" volumes of the feature space. Our risk bounds demonstrate that it is possible to control the error of arbitrarily large decision trees when most of the leaves are concentrated in a small volume. This suggests a potentially new perspective on generalization error bounds that takes into account the interrelationship between classifier complexity and volume in the concentration of the error. The fact that classifiers may be arbitrarily complex in infinitesimally small volumes is crucial for optimal asymptotic rates and may have important practical consequences as well.

Finally, we comment on one significant issue that still remains. The DDTs investigated in this paper cannot provide more efficient approximations to smoother decision boundaries (cases in which $\gamma > 1$), a limitation that leads to suboptimal rates in such cases. The restriction of DDTs (like most other practical decision trees) to axis-orthogonal splits is one limiting factor in their approximation capabilities. Decision trees with more general splits such as "perceptron trees" [51] offer potential advantages, but the analysis and implementation of more general tree structures becomes quite complicated.

Alternatively, we note that a similar boundary approximation issue has been addressed in the image processing literature in the context of representing edges [52]. Multiresolution methods known as "wedgelets" or "curvelets" [8], [53] can better approximate image edges than their wavelet counterparts, but these methods only provide optimal approximations up to $\gamma = 2$, and they do not appear to scale well to dimensions higher than $d = 3$. However, motivated by these methods, we proposed "polynomial-decorated" DDTs, that is, DDTs with empirical risk minimizing polynomial decision boundaries at the leaf nodes [20]. Such trees yield faster rates but they are computationally prohibitive. Recent risk bounds for polynomial-kernel support vector machines may offer a computationally tractable alternative to this approach [19]. One way or another, we feel that dyadic decision trees, or possibly new variants thereof, hold promise to address these issues.

VIII. PROOFS

Our error deviance bounds for trees are stated with explicit, small constants and hold for all sample sizes. Our rate of convergence upper bounds could also be stated with explicit constants (depending on $d, \kappa, \gamma, c_0, c_1, c_2$, etc.) that hold for all n. To do so would require us to explicitly state how the resolution parameter M grows with n. We have opted not to follow this route, however, for two reasons: the proofs are less cluttered, and the statements of our results are somewhat more general. That said, explicit constants are given (in the proofs) where it does not obfuscate the presentation, and it would be a simple exercise for the interested reader to derive explicit constants throughout.

Our analysis of estimation error employs the following concentration inequalities. The first is known as a *relative* Chernoff bound (see [54]), the second is a standard (additive) Chernoff bound [55], [56], and the last two were proved by [56].

Lemma 4 *Let U be a Bernoulli random variable with $\mathbb{P}(U = 1) = p$, and let $U^n = \{U_1, \ldots, U_n\}$ be IID realizations. Set $\hat{p} = \frac{1}{n}\sum_{i=1}^n U_i$. For all $\epsilon > 0$*

$$\mathbb{P}^n \left\{ \hat{p} \leq (1-\epsilon)p \right\} \leq e^{-np\epsilon^2/2}, \tag{20}$$

$$\mathbb{P}^n \left\{ \hat{p} \geq p + \epsilon \right\} \leq e^{-2n\epsilon^2}, \tag{21}$$

$$\mathbb{P}^n \left\{ \sqrt{\hat{p}} \geq \sqrt{p} + \epsilon \right\} \leq e^{-2n\epsilon^2}, \tag{22}$$

$$\mathbb{P}^n \left\{ \sqrt{p} \geq \sqrt{\hat{p}} + \epsilon \right\} \leq e^{-n\epsilon^2}. \tag{23}$$

Corollary 2 *Under the assumptions of the previous lemma*

$$\mathbb{P}^n \left\{ p - \hat{p} \geq \sqrt{\frac{2p\log(1/\delta)}{n}} \right\} \leq \delta.$$

This is proved by applying (20) with $\epsilon = \sqrt{\frac{2\log(1/\delta)}{pn}}$.

A. Proof of Theorem 1

Let $T \in \mathcal{T}$.

$$\begin{aligned} R(T) - \widehat{R}_n(T) &= \sum_{A \in \pi(T)} R(T, A) - \widehat{R}_n(T, A) \\ &= \sum_{A \in \pi(T)} \mathbb{P}(B_{A,T}) - \widehat{\mathbb{P}}_n(B_{A,T}) \end{aligned}$$

where $B_{A,T} = \{(x,y) \in \mathbb{R}^d \times \{0,1\} \,|\, x \in A,\, T(x) \neq y\}$. For fixed A, T, consider the Bernoulli trial U which equals 1 if $(X,Y) \in B_{A,T}$ and 0 otherwise. By Corollary 2

$$\mathbb{P}(B_{A,T}) - \widehat{\mathbb{P}}_n(B_{A,T}) \leq \sqrt{2\mathbb{P}(B_{A,T}) \frac{([\![A]\!]+1)\log 2 + \log(1/\delta)}{n}},$$

except on a set of probability not exceeding $\delta 2^{-([\![A]\!]+1)}$. We want this to hold for all $B_{A,T}$. Note that the sets $B_{A,T}$ are in 2-to-1 correspondence with cells $A \in \mathcal{A}$, because each cell could have one of two class labels. Using $\mathbb{P}(B_{A,T}) \leq p_A$, the union bound, and applying the same argument for each possible $B_{A,T}$, we have that $R(T) - \widehat{R}_n(T) \leq \Phi'_n(T)$ holds uniformly except on a set of probability not exceeding

$$\sum_{B_{A,T}} \delta 2^{-([\![A]\!]+1)} = \sum_{\substack{A \in \mathcal{A} \\ \text{label} = 0 \text{ or } 1}} \delta 2^{-([\![A]\!]+1)} = \sum_{A \in \mathcal{A}} \delta 2^{-[\![A]\!]} \leq \delta,$$

where the last step follows from the Kraft inequality (4). To prove the reverse inequality, note that $\mathcal{T}$ is closed under complimentation. Therefore $\widehat{R}_n(T) - R(T) = R(\overline{T}) - \widehat{R}_n(\overline{T})$. Moreover, $\Phi'_n(\overline{T}) = \Phi'_n(T)$. The result now follows.

B. Proof of Lemma 1

We prove the second statement. The first follows in a similar fashion. For fixed A

$$\begin{aligned}
\mathbb{P}^n\left\{\hat{p}_A \geq p'_A(\delta)\right\} &= \mathbb{P}^n\left\{\hat{p}_A \geq 4\max(p_A, ([\![A]\!]\log 2 + \log(1/\delta))/(2n))\right\} \\
&= \mathbb{P}^n\left\{\sqrt{\hat{p}_A} \geq 2\max(\sqrt{p_A}, \sqrt{([\![A]\!]\log 2 + \log(1/\delta))/(2n)})\right\} \\
&\leq \mathbb{P}^n\left\{\sqrt{\hat{p}_A} \geq \sqrt{p_A} + \sqrt{([\![A]\!]\log 2 + \log(1/\delta))/(2n)}\right\} \\
&\leq \delta 2^{-[\![A]\!]},
\end{aligned}$$

where the last inequality follows from (22) with $\epsilon = \sqrt{([\![A]\!]\log 2 + \log(1/\delta))/(2n))}$. The result follows by repeating this argument for each A and applying the union bound and Kraft inequality (4).

C. Proof of Theorem 3

Recall that in this and subsequent proofs we take $\delta = 1/n$ in the definition of Φ_n, $\tilde{\Phi}_n$, p', and $\hat{p}'$.

Let $T' \in \mathcal{T}_M$ be the tree minimizing the expression on the right-hand side of (12). Take Ω to be the set of all Z^n such that the events in (8) and (10) hold. Then $\mathbb{P}(\Omega) \geq 1 - 3/n$. Given $Z^n \in \Omega$, we know

$$\begin{aligned}
R(\widehat{T}_n) &\leq \widehat{R}_n(\widehat{T}_n) + \Phi_n(\widehat{T}_n) \\
&\leq \widehat{R}_n(T') + \Phi_n(T') \\
&\leq \widehat{R}_n(T') + \tilde{\Phi}_n(T') \\
&\leq R(T') + 2\tilde{\Phi}_n(T')
\end{aligned}$$

where the first inequality follows from (10), the second from (1), the third from (8), and the fourth again from (10). To see the third step, observe that for $Z^n \in \Omega$

$$\begin{aligned}
\hat{p}'_A &= 4\max\left(\hat{p}_A, \frac{[\![A]\!]\log 2 + \log n}{n}\right) \\
&\leq 4\max\left(p'_A, \frac{[\![A]\!]\log 2 + \log n}{n}\right) \\
&= 4\max\left(4\max\left(p_A, \frac{[\![A]\!]\log 2 + \log n}{2n}\right), \frac{[\![A]\!]\log 2 + \log n}{n}\right) \\
&= 4p'_A.
\end{aligned}$$

The first part of the theorem now follows by subtracting R^* from both sides.

To prove the second part, simply observe

$$\begin{aligned}
\mathbb{E}^n\{R(\widehat{T}_n)\} &= \mathbb{P}^n(\Omega)\mathbb{E}^n\{R(\widehat{T}_n)\,|\,\Omega\} + \mathbb{P}^n(\overline{\Omega})\mathbb{E}^n\{R(\widehat{T}_n)\,|\,\overline{\Omega}\} \\
&\leq \mathbb{E}^n\{R(\widehat{T}_n)\,|\,\Omega\} + \frac{3}{n}
\end{aligned}$$

and apply the result of the first part of the proof.

D. Proof of Lemma 3

Recall $\mathcal{P}_m$ denotes the partition of $[0,1]^d$ into hypercubes of sidelength $1/m$. Let $\mathcal{B}_m$ be the collection of cells in $\mathcal{P}_m$ that intersect ∂G^*. Take T' to be the smallest *cyclic* DDT such that $\mathcal{B}_m \subseteq \pi(T')$. In other words, T' is formed by cycling through the coordinates and dyadicly splitting nodes containing both classes of data. Then T' consists of the cells in $\mathcal{B}_m$, together with their ancestors (according to the forced splitting scheme of cyclic DDTs), together with their children. Choose class labels for the leaves of T' such that $R(T')$ is minimized. Note that T' has depth $J = d\log_2 m$.

To verify $K_j(T') \leq 2c_1 2^{\lceil j/d\rceil(d-1)}$, fix j and set $j' = \lceil j/d\rceil d$. Since $j \leq j'$, $K_j(T') \leq N_{j'}(T')$. By construction, the nodes at depth j in T' are those that intersect ∂G^* together with their siblings. Since nodes at depth j' are hypercubes with sidelength $1/\lceil j/d\rceil$, we have $N_{j'}(T') \leq 2c_1(2^{\lceil j/d\rceil})^{d-1}$ by the box-counting assumption.

Finally, observe

$$\begin{aligned}
\lambda(\Delta(T', f^*)) &\leq \lambda(\cup\{A : A \in \mathcal{B}_m\}) \\
&= \sum_{A\in\mathcal{B}_m} \lambda(A) \\
&= |\mathcal{B}_m| m^{-d} \\
&\leq c_1 m^{d-1} m^{-d} = c_1 m^{-1}.
\end{aligned}$$

E. Proof of Theorem 5

Audibert [13] presents two general approaches for proving minimax lower bounds for classification, one based on Assouad's lemma and the other on Fano's lemma. The basic idea behind Assouad's lemma is prove a lower bound for a finite subset (of the class of interest) indexed by the vertices of a discrete hypercube. A minimax lower bound for a subset then implies a lower bound for the full class of distributions. Fano's lemma follows a similar approach but considers a finite set of distributions indexed by a proper subset of an Assouad hypercube (sometimes called a pyramid). The Fano pyramid has cardinality proportional to the full hypercube but its elements are better separated which eases analysis in some cases. For an overview of minimax lower bounding techniques in nonparametric statistics see [57].

As noted by [13, chap. 3, sec. 6.2], Assouad's lemma is inadequate for excluding low noise levels (at least when using the present proof technique) because the members of the hypercube do not satisfy the low noise exclusion condition. To prove lower bounds for a two-sided noise condition, Audibert applies Birgé's version of Fano's lemma. We follow in particular the techniques laid out in Section 6.2 and Appendix E of [13, chap. 3], with some variations, including a different version of Fano's lemma.

Our strategy is to construct a finite set of probability measures $\mathcal{D}_m \subset \mathcal{D}_{\text{BOX}}(\kappa)$ for which the lower bound holds. We proceed as follows. Let m be a dyadic integer such that $m \asymp n^{1/(2\kappa+d-2)}$. In particular, it will suffice to take $\log_2 m = \lceil \frac{1}{2\kappa+d-2} \log_2 n \rceil$ so that

$$n^{1/(2\kappa+d-2)} \leq m \leq 2n^{1/(2\kappa+d-2)}.$$

Let $\Omega = \{1, \ldots, m\}^{d-1}$. Associate $\xi = (\xi_1, \ldots, \xi_{d-1}) \in \Omega$ with the hypercube

$$A_\xi = \left(\prod_{j=1}^{d-1} \left[\frac{\xi_j - 1}{m}, \frac{\xi_j}{m} \right] \right) \times \left[0, \frac{1}{m} \right] \subseteq [0,1]^d$$

where $\prod$ denotes Cartesian cross-product. To each $\omega \subseteq \Omega$ assign the set

$$G_\omega = \bigcup_{\xi \in \omega} A_\xi.$$

Observe that $\lambda(\Delta(G_{\omega_1}, G_{\omega_2})) \leq \frac{1}{m}$ for all $\omega_1, \omega_2 \subseteq \Omega$.

Lemma 5 *There exists a collection $\mathcal{G}'$ of subsets of $[0,1]^d$ such that*

1) *each $G' \in \mathcal{G}'$ has the form $G' = G_\omega$ for some $\omega \subseteq \Omega$*
2) *for any $G'_1 \neq G'_2$ in $\mathcal{G}'$, $\lambda(\Delta(G'_1, G'_2)) \geq \frac{1}{4m}$*
3) $\log |\mathcal{G}'| \geq \frac{1}{8} m^{d-1}$.

Proof: Subsets of Ω are in one-to-one correspondence with points in the discrete hypercube $\{0,1\}^{m^{d-1}}$. We invoke the following result [58, lemma 7].

Lemma 6 (Huber) *Let $\delta(\sigma,\sigma')$ denote the Hamming distance between σ and σ' in $\{0,1\}^p$. There exists a subset Σ of $\{0,1\}^p$ such that*

- *for any $\sigma \neq \sigma'$ in Σ, $\delta(\sigma,\sigma') \geq \frac{p}{4}$.*
- $\log|\Sigma| \geq \frac{p}{8}$.

Lemma 5 now follows from Lemma 6 with $p = m^{d-1}$ and using $\lambda(A_\xi) = m^{-d}$ for each ξ. ■

Let a be a positive constant to be specified later and set $b = am^{-(\kappa-1)}$. Let $\mathcal{G}'$ be as in Lemma 5 and define $\mathcal{D}'_m$ to be the set of all probability measures $\mathbb{P}$ on $\mathcal{Z}$ such that

(i) $\mathbb{P}_X = \lambda$

(ii) For some $G' \in \mathcal{G}'$

$$\eta(x) = \begin{cases} \frac{1+b}{2} & x \in G' \\ \frac{1-b}{2} & x \notin G'. \end{cases}$$

Now set $\mathcal{D}_m = \{\mathbb{P}^n : \mathbb{P} \in \mathcal{D}'_m\}$. By construction, $\log|\mathcal{D}_m| \geq \frac{1}{8}m^{d-1}$.

Clearly **0A** holds for $\mathcal{D}_m$ provided $c_0 \geq 1$. Condition **1B** requires $N_k(G') \leq c_1 k^{d-1}$ for all k. This holds trivially for $k \leq m$ provided $c_1 \geq 1$. For $k > m$ it also holds provided $c_1 \geq 4d$. To see this, note that every face of a hypercube A_ξ intersects $2(k/m)^{d-1}$ hypercubes of sidelength $1/k$. Since each G' is composed of at most m^{d-1} hypercubes A_ξ, and each A_ξ has $2d$ faces, we have

$$N_k(G') \leq 2d \cdot m^{d-1} \cdot 2(k/m)^{d-1} = 4dk^{d-1}.$$

To verify **2B**, consider $\mathbb{P}^n \in \mathcal{D}_m$ and let f^* be the corresponding Bayes classifier. We need to show

$$(R(T_k^*) - R^*)^{1/\kappa} \leq \frac{c_2}{k}$$

for every dyadic k, where T_k^* minimizes $\lambda(\Delta(T, f^*))$ over all $T \in \mathcal{T}_k(c_1)$. For $k \geq m$ this holds trivially

because $T_k^* = f^*$. Consider $k < m$. By Lemma 3 we know $\lambda(\Delta(T_k^*, f^*)) \leq \frac{c_1}{k}$. Now

$$\begin{aligned} R(T_k^*) - R^* &= \int_{\Delta(T_k^*, f^*)} 2|\eta(x) - 1/2| d\mathbb{P}_X \\ &= b\lambda(\Delta(T_k^*, f^*)) \\ &\leq \frac{c_1 b}{k} \\ &= \frac{a c_1 m^{-(\kappa-1)}}{k} \\ &\leq \frac{a c_1 k^{-(\kappa-1)}}{k} \\ &= a c_1 k^{-\kappa}. \end{aligned}$$

Thus $(R(T_k^*) - R^*)^{1/\kappa} \leq \frac{c_2}{k}$ provided $a \leq c_2^\kappa / c_1$.

It remains to derive a lower bound for the expected excess risk. We employ the following generalization of Fano's lemma due to [57]. Introducing notation, let $\bar{d}$ be a pseudo-metric on a parameter space Θ, and let $\hat{\theta}$ be an estimator of $\theta = \theta(\mathbb{P})$ based on a realization drawn from $\mathbb{P}$.

Lemma 7 (Yu) *Let $r \geq 1$ be an integer and let $\mathcal{M}_r$ contain r probability measures indexed by $j = 1, \ldots, r$ such that for any $j \neq j'$*

$$\bar{d}(\theta(\mathbb{P}_j), \theta(\mathbb{P}_{j'})) \geq \alpha_r$$

and

$$K(\mathbb{P}_j, \mathbb{P}_{j'}) = \int \log(\mathbb{P}_j / \mathbb{P}_{j'}) d\mathbb{P}_j \leq \beta_r.$$

Then

$$\max_j \mathbb{E}_j \bar{d}(\hat{\theta}, \theta(\mathbb{P}_j)) \geq \frac{\alpha_r}{2} \left(1 - \frac{\beta_r + \log 2}{\log r}\right).$$

In the present setting we have $\Theta = \mathcal{F}(\mathcal{X}, \mathcal{Y})$, $\theta(\mathbb{P}) = f^*$, and $\bar{d}(f, f') = \lambda(\Delta(f, f'))$. We apply the lemma with $r = |\mathcal{D}_m|$, $\mathcal{M}_r = \mathcal{D}_m$, $\alpha_r = \frac{1}{4m}$, and $R(\widehat{f}_n) - R^* = b\lambda(\Delta(\widehat{f}_n, f^*))$.

Corollary 3 *Assume that for $\mathbb{P}_j^n, \mathbb{P}_{j'}^n \in \mathcal{D}_m$, $j \neq j'$,*

$$K(\mathbb{P}_j^n, \mathbb{P}_{j'}^n) = \int \log(\mathbb{P}_j^n / \mathbb{P}_{j'}^n) d\mathbb{P}_j \leq \beta_m.$$

Then

$$\max_{\mathcal{D}_m} \left[\mathbb{E}^n\{R(\widehat{f}_n)\} - R^*\right] \geq \frac{b}{8m} \left(1 - \frac{\beta_m + \log 2}{\frac{1}{8} m^{d-1}}\right).$$

Since $b/m \asymp n^{-\kappa/(2\kappa+d-2)}$, it suffices to show $(\beta_m + \log 2)/(\frac{1}{8} m^{d-1})$ is bounded by a constant < 1 for m sufficiently large.

Toward this end, let $\mathbb{P}^n_j, \mathbb{P}^n_{j'} \in \mathcal{D}_m$. We have

$$\begin{aligned} K(\mathbb{P}^n_j, \mathbb{P}^n_{j'}) &= \int_{\mathcal{Z}^n} \log(\mathbb{P}^n_j/\mathbb{P}^n_{j'}) d\mathbb{P}_j \\ &= \int_{\mathcal{X}^n} \left(\int_{\mathcal{Y}^n} \log(\mathbb{P}^n_j/\mathbb{P}^n_{j'}) d(\mathbb{P}^n_j)_{Y|X} \right) d(\mathbb{P}^n_j)_X. \end{aligned}$$

The inner integral is 0 unless $x \in \Delta(f^*_j, f^*_{j'})$. Since all $\mathbb{P}^n \in \mathcal{D}_m$ have a uniform first marginal we have

$$\begin{aligned} K(\mathbb{P}^n_j, \mathbb{P}^n_{j'}) &= n\lambda(\Delta(f^*_j, f^*_{j'})) \left(\frac{1+b}{2} \log\left(\frac{1+b}{1-b}\right) + \frac{1-b}{2} \log\left(\frac{1-b}{1+b}\right) \right) \\ &\leq \frac{nb}{m} \log\left(\frac{1+b}{1-b}\right) \\ &\leq 2.5\frac{nb^2}{m} \end{aligned}$$

where we use the elementary inequality $\log((1+b)/(1-b)) \leq 2.5b$ for $b \leq 0.7$. Thus take $\beta_m = 2.5nb^2/m$. We have

$$\begin{aligned} \frac{\beta_m + \log 2}{\frac{1}{8}m^{d-1}} &\leq 20\frac{nb^2}{m^d} + 8(\log 2)\frac{1}{m^{d-1}} \\ &\leq 20a^2 n^1 n^{-\frac{2\kappa-2}{2\kappa+d-2}} n^{-\frac{d}{2\kappa+d-2}} + 8(\log 2)\frac{1}{m^{d-1}} \\ &= 20a^2 + 8(\log 2)\frac{1}{m^{d-1}} \\ &\leq .5 \end{aligned}$$

provided a is sufficiently small and m sufficiently large. This proves the theorem.

F. Proof of Theorem 6

Let $\mathbb{P}^n \in \mathcal{D}_{\text{BOX}}(\kappa, c_0, c_1, c_2)$. For now let m be an arbitrary dyadic integer. Later we will specify it to balance approximation and estimation errors. By **2B**, there exists $T' \in \mathcal{T}_m(c_1)$ such that

$$R(T') - R^* \leq c_2^\kappa m^{-\kappa}.$$

This bounds the approximation error. Note that T' has depth $J \leq d\ell$ where $m = 2^\ell$.

The estimation error is bounded as follows.

Lemma 8

$$\tilde{\Phi}_n(T') \preccurlyeq m^{d/2-1}\sqrt{\log n/n}$$

Proof: We begin with three observations. First,

$$\begin{aligned} \sqrt{p'_A} &\leq 2\sqrt{p_A + (\llbracket A \rrbracket \log 2 + \log n)/(2n)} \\ &\leq 2(\sqrt{p_A} + \sqrt{(\llbracket A \rrbracket \log 2 + \log n)/(2n)}). \end{aligned}$$

Second, if A corresponds to a node of depth $j = j(A)$, then by **0A**, $p_A \le c_0\lambda(A) = c_0 2^{-j(A)}$. Third, $[\![A]\!] \le (2+\log_2 d)j(A) \le (2+\log_2 d)d\ell \preccurlyeq \log n$. Combining these, we have $\tilde{\Phi}_n(T') \preccurlyeq \tilde{\Phi}_n^1(T') + \tilde{\Phi}_n^2(T')$ where

$$\tilde{\Phi}_n^1(T) = \sum_{A\in\pi(T)} \sqrt{2^{-j(A)}\frac{\log n}{n}}$$

and

$$\tilde{\Phi}_n^2(T) = \sum_{A\in\pi(T)} \sqrt{\frac{\log n}{n}\cdot\frac{\log n}{n}}. \tag{24}$$

We note that $\tilde{\Phi}_n^2(T') \preccurlyeq \tilde{\Phi}_n^1(T')$. This follows from $m \asymp (n/\log n)^{1/(2\kappa+d-2)}$, for then $\log n/n \preccurlyeq m^{-d} = 2^{-d\ell} \le 2^{-j(A)}$ for all A.

It remains to bound $\tilde{\Phi}_n^1(T')$. Let K_j be the number of nodes in T' at depth j. Since $T' \in \mathcal{T}_m(c_1)$ we know $K_j \le 2c_1 2^{\lceil j/d\rceil(d-1)}$ for all $j \le J$. Writing $j = (p-1)d+q$ where $1 \le p \le \ell$ and $1 \le q \le d$ we have

$$\begin{aligned}
\tilde{\Phi}_n^1(T') &\preccurlyeq \sum_{j=1}^{J} 2^{\lceil j/d\rceil(d-1)}\sqrt{2^{-j}\frac{\log n}{n}} \\
&\preccurlyeq \sqrt{\frac{\log n}{n}}\sum_{p=1}^{\ell}\sum_{q=1}^{d} 2^{p(d-1)}\sqrt{2^{-[(p-1)d+q]}} \\
&= \sqrt{\frac{\log n}{n}}\sum_{p=1}^{\ell} 2^{p(d-1)}2^{-(p-1)d/2}\sum_{q=1}^{d} 2^{-q/2} \\
&\preccurlyeq \sqrt{\frac{\log n}{n}}\sum_{p=1}^{\ell} 2^{p(d/2-1)} \\
&\preccurlyeq 2^{\ell(d/2-1)}\sqrt{\frac{\log n}{n}} \\
&= m^{d/2-1}\sqrt{\frac{\log n}{n}}.
\end{aligned}$$

Note that although we use $\preccurlyeq$ instead of $\le$ at some steps, this is only to streamline the presentation and not because we need n sufficiently large. ■

The theorem now follows by the oracle inequality and choosing $m \asymp (n/\log n)^{1/(2\kappa+d-2)}$ and plugging into the above bounds on approximation and estimation error.

G. Proof of Theorem 7

Let $m = 2^\ell$ be a dyadic integer, $1 \le \ell \le L = \log_2 M$, with $m \asymp (n/\log n)^{1/d'}$. Let $\mathcal{B}_m$ be the collection of cells in $\mathcal{P}_m$ that intersect ∂G^*. Take T' to be the smallest cyclic DDT such that

$\mathcal{B}_m \subseteq \pi(T')$. In other words, T' consists of the cells in $\mathcal{B}_m$, together with their ancestors (according to the forced splitting structure of cyclic DDTs) and their ancestors' children. Choose class labels for the leaves of T' such that $R(T')$ is minimized. Note that T' has depth $J = d\ell$. The construction of T' is identical to the proof of Lemma 3; the difference now is that $|\mathcal{B}_m|$ is substantially smaller.

Lemma 9 *For all* m,

$$R(T') - R^* \leq c_0 c_1 m^{-1}.$$

Proof: We have

$$\begin{aligned} R(T') - R^* &\leq \mathbb{P}(\Delta(T', f^*)) \\ &\leq \mathbb{P}(\cup\{A : A \in \mathcal{B}_m\}) \\ &= \sum_{A \in \mathcal{B}_m} \mathbb{P}(A) \\ &\leq c_0 |\mathcal{B}_m| m^{-d'} \\ &\leq c_0 c_1 m^{d'-1} m^{-d'} \\ &= c_0 c_1 m^{-1} \end{aligned}$$

where the third inequality follows from **0B** and the last inequality from **1C**. ■

Next we bound the estimation error.

Lemma 10

$$\tilde{\Phi}_n(T') \preccurlyeq m^{d'/2-1}\sqrt{\log n/n}$$

Proof: If A is a cell at depth $j = j(A)$ in T, then $p_A \leq c_0 2^{-\lfloor j/d \rfloor d'}$ by assumption **0B**. Arguing as in the proof of Theorem 6, we have $\tilde{\Phi}_n(T') \preccurlyeq \tilde{\Phi}_n^1(T') + \tilde{\Phi}_n^2(T')$ where

$$\tilde{\Phi}_n^1(T) = \sum_{A \in \pi(T)} \sqrt{2^{-\lfloor j(A)/d \rfloor d'} \frac{\log n}{n}}$$

and $\tilde{\Phi}_n^2(T)$ is as in (24). Note that $\tilde{\Phi}_n^2(T') \preccurlyeq \tilde{\Phi}_n^1(T')$. This follows from $m \asymp (n/\log n)^{1/d'}$, for then $\log n/n \preccurlyeq m^{-d'} = 2^{-\ell d'} \leq 2^{-\lfloor j(A)/d \rfloor d'}$ for all A.

It remains to bound $\tilde{\Phi}_n^1(T')$. Let K_j be the number of nodes in T' at depth j. Arguing as in the proof of Lemma 3 we have $K_j \leq 2c_1 2^{\lceil j/d \rceil (d'-1)}$. Then

$$\begin{aligned}
\tilde{\Phi}_n^1(T') &\asymp \sum_{j=1}^{J} 2^{\lceil j/d \rceil (d'-1)} \sqrt{2^{-\lfloor j/d \rfloor d'} \frac{\log n}{n}} \\
&= \sqrt{\frac{\log n}{n}} \sum_{p=1}^{\ell} 2^{p(d'-1)} \sum_{q=1}^{d} \sqrt{2^{-\lfloor \frac{(p-1)d+q}{d} \rfloor d'}} \\
&\leq \sqrt{\frac{\log n}{n}} \sum_{p=1}^{\ell} 2^{p(d'-1)} \cdot d \sqrt{2^{-(p-1)d'}} \\
&\asymp \sqrt{\frac{\log n}{n}} \sum_{p=1}^{\ell} 2^{p(d'/2-1)} \\
&\asymp 2^{\ell(d'/2-1)} \sqrt{\frac{\log n}{n}} \\
&= m^{d'/2-1} \sqrt{\frac{\log n}{n}}.
\end{aligned}$$

■

The theorem now follows by the oracle inequality and plugging $m \asymp (n/\log n)^{1/d'}$ into the above bounds on approximation and estimation error.

H. Proof of Theorem 8

Assume without loss of generality that the first d'' coordinates are relevant and the remaining $d-d''$ are statistically independent of Y. Then ∂G^* is the Cartesian product of a "box-counting" curve in $[0,1]^{d''}$ with $[0,1]^{d-d''}$. Formally, we have the following.

Lemma 11 *Let m be a dyadic integer, and consider the partition of $[0,1]^{d''}$ into hypercubes of sidelength $1/m$. Then the projection of ∂G^* onto $[0,1]^{d''}$ intersects at most $c_1 m^{d''-1}$ of those hypercubes.*

Proof: If not, then ∂G^* intersects more than $c_1 m^{d-1}$ members of $\mathcal{P}_m$ in $[0,1]^d$, in violation of the box-counting assumption. ■

Now construct the tree T' as follows. Let $m = 2^\ell$ be a dyadic integer, $1 \leq \ell \leq L$, with $m \asymp (n/\log n)^{1/d''}$. Let $\mathcal{P}''_m$ be the partition of $[0,1]^d$ obtained by splitting the first d'' features uniformly into cells of sidelength $1/m$. Let $\mathcal{B}''_m$ be the collection of cells in $\mathcal{P}''_m$ that intersect ∂G^*. Let T'_{INIT} be the DDT formed by splitting cyclicly through the first d'' features until all leaf nodes have a depth of $J = d''\ell$. Take T' to be the smallest pruned subtree of T'_{INIT} such that $\mathcal{B}''_m \subset \pi(T')$. Choose class labels for the leaves of T' such that $R(T')$ is minimized. Note that T' has depth $J = d''\ell$.

Lemma 12 *For all m,*

$$R(T') - R^* \leq c_0 c_1 m^{-1}.$$

Proof: We have

$$\begin{aligned} R(T') - R^* &\leq \mathbb{P}(\Delta(T', f^*)) \\ &\leq c_0 \lambda(\Delta(T', f^*)) \\ &\leq c_0 \lambda(\cup\{A : A \in \mathcal{B}''_m\}) \\ &= c_0 \sum_{A \in \mathcal{B}''_m} \lambda(A) \\ &= c_0 |\mathcal{B}''_m| m^{-d''} \\ &\leq c_0 c_1 m^{d''-1} m^{-d''} \\ &= c_0 c_1 m^{-1} \end{aligned}$$

where the second inequality follows from **0A** and the last inequality from Lemma 11. ■

The remainder of the proof proceeds in a manner entirely analogous to the proofs of the previous two theorems, where now $K_j \leq 2c_1 2^{\lceil j/d'' \rceil (d''-1)}$.

I. Proof of Theorem 9

Assume without loss of generality that the last coordinate of ∂G^* is a function of the others. Let $m = 2^\ell$ be a dyadic integer, $1 \leq \ell \leq L$, with $m \asymp (n/\log n)^{1/(\gamma+d-1)}$. Let $\tilde{m} = 2^{\tilde{\ell}}$ be the largest dyadic integer not exceeding m^γ. Note that $\tilde{\ell} = \lfloor \gamma \ell \rfloor$. Construct the tree T' as follows. First, cycle through the first $d-1$ coordinates $\ell - \tilde{\ell}$ times, subdividing dyadicly along the way. Then, cycle through all d coordinates $\tilde{\ell}$ times, again subdividing dyadicly at each step. Call this tree T'_{INIT}. The leaves of T'_{INIT} are hyperrectangles with sidelength $2^{-\ell}$ along the first $d-1$ coordinates and sidelength $2^{-\tilde{\ell}}$ along the last coordinate. Finally, form T' by pruning back all cells in T'_{INIT} whose parents do not intersect ∂G^*. Note that T' has depth $J = (\ell - \tilde{\ell})(d-1) + \tilde{\ell} d$.

Lemma 13 *Let K_j denote the number of nodes in T' at depth j. Then*

$$K_j \begin{cases} = 0 & j \leq (\ell - \tilde{\ell})(d-1) \\ \leq C 2^{(\ell - \tilde{\ell} + p)(d-1)} & j = (\ell - \tilde{\ell})(d-1) + (p-1)d + q \end{cases}$$

where $C = 2c_1(d-1)^{\gamma/2} + 4$ and $p = 1, \ldots, \tilde{\ell}$ and $q = 1, \ldots, d$.

Proof: In the first case the result is obvious by construction of T' and the assumption that one coordinate of ∂G^* is a function of the others. For the second case, let $j = (\ell - \tilde{\ell})(d-1) + (p-1)d + q$ for some p and q. Define $\mathcal{P}_m^\gamma(j)$ to be the partition of $[0,1]^d$ formed by the set of cells in T'_{INIT} having depth j. Define $\mathcal{B}_m^\gamma(j)$ to be the set of cells in $\mathcal{P}_m^\gamma(j)$ that intersect ∂G^*. By construction of T' we have $K_j \leq 2|\mathcal{B}_m^\gamma(j)|$. From the fact $|\mathcal{B}_m^\gamma(j)| \leq |\mathcal{B}_m^\gamma(j+1)|$, we conclude

$$K_j \leq 2|\mathcal{B}_m^\gamma(j)| \leq 2|\mathcal{B}_m^\gamma((\ell - \tilde{\ell})(d-1) + pd)|.$$

Thus it remains to show $2|\mathcal{B}_m^\gamma((\ell - \tilde{\ell})(d-1) + pd)| \leq C2^{(\ell-\tilde{\ell}+p)(d-1)}$ for each p. Each cell in $\mathcal{B}_m^\gamma((\ell-\tilde{\ell})(d-1)+pd)$ is a rectangle of the form $U_\sigma \times V_\tau$, where $U_\sigma \subseteq [0,1]^{d-1}, \sigma = 1, \ldots, 2^{(\ell-\tilde{\ell}+p)(d-1)}$ is a hypercube of sidelength $2^{-(\ell-\tilde{\ell}+p)}$, and $V_\tau, \tau = 1, \ldots, 2^p$ is an interval of length 2^{-p}. For each $\sigma = 1, \ldots, 2^{(\ell-\tilde{\ell}+p)(d-1)}$, set $\mathcal{B}_m^\gamma(p,\sigma) = \{U_\sigma \times V_\tau \in \mathcal{B}_m^\gamma((\ell - \tilde{\ell})(d-1) + pd) \mid \tau = 1, \ldots, 2^p\}$.

The lemma will be proved if we can show $|\mathcal{B}_m^\gamma(p,\sigma)| \leq c_1(d-1)^{\gamma/2} + 2$, for then

$$\begin{aligned} K_j &\leq 2|\mathcal{B}_m^\gamma((\ell - \tilde{\ell})(d-1) + pd)| \\ &= \sum_{\sigma=1}^{2^{(\ell-\tilde{\ell}+p)(d-1)}} 2|\mathcal{B}_m^\gamma(p,\sigma)| \\ &\leq C2^{(\ell-\tilde{\ell}+p)(d-1)} \end{aligned}$$

as desired. To prove this fact, recall $\partial G^* = \{(s,t) \in [0,1]^d \mid t = g(s)\}$ for some function $g : [0,1]^{d-1} \to [0,1]$ satisfying $|g(s) - g(s')| \leq c_1|s - s'|^\gamma$ for all $s, s' \in [0,1]^{d-1}$. Therefore, the value of g on a single hypercube U_σ can vary by no more than $c_1(\sqrt{d-1} \cdot 2^{-(\ell-\tilde{\ell}+p)})^\gamma$. Here we use the fact that the maximum distance between points in U_σ is $\sqrt{d-1} \cdot 2^{-(\ell-\tilde{\ell}+p)}$. Since each interval V_τ has length 2^{-p} ,

$$\begin{aligned} |\mathcal{B}_m^\gamma(p,\sigma)| &\leq \frac{c_1(d-1)^{\gamma/2}(2^{-(\ell-\tilde{\ell}+p)\gamma})}{2^{-p}} + 2 \\ &= c_1(d-1)^{\gamma/2}2^{-(\ell\gamma-\tilde{\ell}\gamma+p\gamma-p)} + 2 \\ &\leq c_1(d-1)^{\gamma/2}2^{-(\tilde{\ell}-\tilde{\ell}\gamma+p\gamma-p)} + 2 \\ &= c_1(d-1)^{\gamma/2}2^{-(\tilde{\ell}-p)(1-\gamma)} + 2 \\ &\leq c_1(d-1)^{\gamma/2} + 2. \end{aligned}$$

This proves the lemma. ■

The following lemma bounds the approximation error.

Lemma 14 *For all m,*

$$R(T') - R^* \leq Cm^{-\gamma},$$

where $C = 2c_0(c_1(d-1)^{\gamma/2} + 4)$.

Proof: Recall T' has depth $J = (\ell - \tilde{\ell})(d-1) + \tilde{\ell}d$, and define $\mathcal{B}_m^\gamma(j)$ as in the proof of the Lemma 13.

$$\begin{aligned} R(T') - R^* &\leq \mathbb{P}(\Delta(T', f^*)) \\ &\leq c_0 \lambda(\Delta(T', f^*)) \\ &\leq c_0 \lambda(\cup\{A : A \in \mathcal{B}_m^\gamma(J)\}) \\ &= c_0 \sum_{A \in \mathcal{B}_m^\gamma(J)} \lambda(A). \end{aligned}$$

By construction, $\lambda(A) = 2^{-\ell(d-1)-\tilde{\ell}}$. Noting that $2^{-\tilde{\ell}} = 2^{-\lfloor \gamma \ell \rfloor} \leq 2^{-\gamma\ell+1}$, we have $\lambda(A) \leq 2 \cdot 2^{-\ell(d+\gamma-1))} = 2m^{-(d+\gamma-1)}$. Thus

$$\begin{aligned} R(T') - R^* &\leq 2c_0 |\mathcal{B}_m^\gamma(J)| m^{-(d+\gamma-1)} \\ &\leq C 2^{\ell(d-1)} m^{-(d+\gamma-1)} \\ &= C m^{d-1} m^{-(d+\gamma-1)} \\ &= C m^{-\gamma}. \end{aligned}$$

■

The bound on estimation error decays as follows.

Lemma 15

$$\tilde{\Phi}_n(T') \preccurlyeq m^{(d-\gamma-1)/2} \sqrt{\log n / n}$$

This lemma follows from Lemma 13 and techniques used in the proofs of Theorems 6 and 7. The theorem now follows by the oracle inequality and plugging $m \asymp (n/\log n)^{1/(\gamma+d-1)}$ into the above bounds on approximation and estimation error.

ACKNOWLEDGMENT

The authors thank Rui Castro and Rebecca Willett for their helpful feedback, and Rui Castro for his insights regarding the two-sided noise condition.

REFERENCES

[1] A. B. Tsybakov, "Optimal aggregation of classifiers in statistical learning," *Ann. Stat.*, vol. 32, no. 1, pp. 135–166, 2004.

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

[3] A. Barron, "Complexity regularization with application to artificial neural networks," in *Nonparametric functional estimation and related topics*, G. Roussas, Ed. Dordrecht: NATO ASI series, Kluwer Academic Publishers, 1991, pp. 561–576.

[4] G. Lugosi and K. Zeger, "Concept learning using complexity regularization," *IEEE Trans. Inform. Theory*, vol. 42, no. 1, pp. 48–54, 1996.

[5] C. Scott and R. Nowak, "Dyadic classification trees via structural risk minimization," in *Advances in Neural Information Processing Systems 15*, S. Becker, S. Thrun, and K. Obermayer, Eds. Cambridge, MA: MIT Press, 2003.

[6] R. A. DeVore, "Nonlinear approximation," *Acta Numerica*, vol. 7, pp. 51–150, 1998.

[7] A. Cohen, W. Dahmen, I. Daubechies, and R. A. DeVore, "Tree approximation and optimal encoding," *Applied and Computational Harmonic Analysis*, vol. 11, no. 2, pp. 192–226, 2001.

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

[9] S. Mallat, *A Wavelet Tour of Signal Processing*. San Diego, CA: Academic Press, 1998.

[10] E. Mammen and A. B. Tsybakov, "Smooth discrimination analysis," *Ann. Stat.*, vol. 27, pp. 1808–1829, 1999.

[11] J. S. Marron, "Optimal rates of convergence to Bayes risk in nonparametric discrimination," *Ann. Stat.*, vol. 11, no. 4, pp. 1142–1155, 1983.

[12] Y. Yang, "Minimax nonparametric classification–Part I: Rates of convergence," *IEEE Trans. Inform. Theory*, vol. 45, no. 7, pp. 2271–2284, 1999.

[13] J.-Y. Audibert, "PAC-Bayesian statistical learning theory," Ph.D. dissertation, Université Paris 6, June 2004.

[14] A. B. Tsybakov and S. A. van de Geer, "Square root penalty: adaptation to the margin in classification and in edge estimation," 2004, preprint. [Online]. Available: http://www.proba.jussieu.fr/pageperso/tsybakov/tsybakov.html

[15] P. Bartlett, M. Jordan, and J. McAuliffe, "Convexity, classification, and risk bounds," Department of Statistics, U.C. Berkeley, Tech. Rep. 638, 2003, to appear in *Journal of the American Statistical Association*.

[16] G. Blanchard, G. Lugosi, and N. Vayatis, "On the rate of convergence of regularized boosting classifiers," *J. Machine Learning Research*, vol. 4, pp. 861–894, 2003.

[17] J. C. Scovel and I. Steinwart, "Fast rates for support vector machines," Los Alamos National Laboratory, Tech. Rep. LA-UR 03-9117, 2003.

[18] Q. Wu, Y. Ying, and D. X. Zhou, "Multi-kernel regularized classifiers," *Submitted to J. Complexity*, 2004.

[19] D. X. Zhou and K. Jetter, "Approximation with polynomial kernels and SVM classifiers," *Submitted to Advances in Computational Mathematics*, 2004.

[20] C. Scott and R. Nowak, "Near-minimax optimal classification with dyadic classification trees," in *Advances in Neural Information Processing Systems 16*, S. Thrun, L. Saul, and B. Schölkopf, Eds. Cambridge, MA: MIT Press, 2004.

[21] S. Murthy, "Automatic construction of decision trees from data: A multi-disciplinary survey," *Data Mining and Knowledge Discovery*, vol. 2, no. 4, pp. 345–389, 1998.

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

[23] R. Quinlan, *C4.5: Programs for Machine Learning*. San Mateo: Morgan Kaufmann, 1993.

[24] M. Kearns and Y. Mansour, "On the boosting ability of top-down decision tree learning algorithms," *Journal of Computer and Systems Sciences*, vol. 58, no. 1, pp. 109–128, 1999.

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

[26] L. Gordon and R. Olshen, "Asymptotically efficient solutions to the classification problem," *Ann. Stat.*, vol. 6, no. 3, pp. 515–533, 1978.

[27] F. Esposito, D. Malerba, and G. Semeraro, "A comparitive analysis of methods for pruning decision trees," *IEEE Trans. Patt. Anal. Mach. Intell.*, vol. 19, no. 5, pp. 476–491, 1997.

[28] Y. Mansour, "Pessimistic decision tree pruning," in *Proc. 14th Int. Conf. Machine Learning*, D. H. Fisher, Ed. Nashville, TN: Morgan Kaufmann, 1997, pp. 195–201.

[29] M. Kearns and Y. Mansour, "A fast, bottom-up decision tree pruning algorithm with near-optimal generalization," in *Proc. 15th Int. Conf. Machine Learning*, J. W. Shavlik, Ed. Madison, WI: Morgan Kaufmann, 1998, pp. 269–277.

[30] G. Blanchard, C. Schäfer, and Y. Rozenholc, "Oracle bounds and exact algorithm for dyadic classification trees," in *Learning Theory: 17th Annual Conference on Learning Theory, COLT 2004*, J. Shawe-Taylor and Y. Singer, Eds. Heidelberg: Springer-Verlag, 2004, pp. 378–392.

[31] Y. Mansour and D. McAllester, "Generalization bounds for decision trees," in *Proceedings of the Thirteenth Annual Conference on Computational Learning Theory*, N. Cesa-Bianchi and S. Goldman, Eds., Palo Alto, CA, 2000, pp. 69–74.

[32] A. Nobel, "Analysis of a complexity based pruning scheme for classification trees," *IEEE Trans. Inform. Theory*, vol. 48, no. 8, pp. 2362–2368, 2002.

[33] N. Berkman and T. Sandholm, "What should be minimized in a decision tree: A re-examination," University of Massachusetts at Amherst, Tech. Rep. TR 95-20, 1995.

[34] M. Golea, P. Bartlett, W. S. Lee, and L. Mason, "Generalization in decision trees and DNF: Does size matter?" in *Advances in Neural Information Processing Systems 10*. Cambridge, MA: MIT Press, 1998.

[35] G. Blanchard, C. Schfer, Y. Rozenholc, and K.-R. Mller, "Optimal dyadic decision trees," Tech. Rep., 2005, preprint. [Online]. Available: http://www.math.u-psud.fr/$\sim$blanchard/

[36] D. Donoho, "CART and best-ortho-basis selection: A connection," *Ann. Stat.*, vol. 25, pp. 1870–1911, 1997.

[37] D. Donoho and I. Johnstone, "Ideal adaptation via wavelet shrinkage," *Biometrika*, vol. 81, pp. 425–455, 1994.

[38] ——, "Adapting to unknown smoothness via wavelet shrinkage," *J. Amer. Statist. Assoc.*, vol. 90, pp. 1200–1224, 1995.

[39] D. L. Donoho, I. M. Johnstone, G. Kerkyacharian, and D. Picard, "Wavelet shrinkage: Asymptopia?" *J. Roy. Statist. Soc. B*, vol. 57, no. 432, pp. 301–369, 1995.

[40] E. Kolaczyk and R. Nowak, "Multiscale likelihood analysis and complexity penalized estimation," *Ann. Stat.*, vol. 32, no. 2, pp. 500–527, 2004.

[41] ——, "Multiscale generalized linear models for nonparametric function estimation," To appear in *Biometrika*, vol. 91, no. 4, December 2004.

[42] I. Johnstone, "Wavelets and the theory of nonparametric function estimation," *Phil. Trans. Roy. Soc. Lond. A.*, vol. 357, pp. 2475–2494, 1999.

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

[44] S. Gey and E. Nedelec, "Risk bounds for CART regression trees," in *MSRI Proc. Nonlinear Estimation and Classification*. Springer-Verlag, December 2002.

[45] A. W. van er Vaart and J. A. Wellner, *Weak Convergence and Empirical Processes*. New York: Springer, 1996.

[46] P. Bartlett, S. Boucheron, and G. Lugosi, "Model selection and error estimation," *Machine Learning*, vol. 48, pp. 85–113, 2002.

[47] S. Boucheron, O. Bousquet, and G. Lugosi, "Theory of classification: a survey of recent advances," 2004, preprint. [Online]. Available: http://www.econ.upf.es/~lugosi/

[48] R. Dudley, "Metric entropy of some classes of sets with differentiable boundaries," *J. Approx. Theory*, vol. 10, pp. 227–236, 1974.

[49] K. Falconer, *Fractal Geometry: Mathematical Foundations and Applications*. West Sussex, England: Wiley, 1990.

[50] G. Blanchard, Personal communication, August 2004.

[51] K. Bennett, N. Cristianini, J. Shawe-Taylor, and D. Wu, "Enlarging the margins in perceptron decision trees," *Machine Learning*, vol. 41, pp. 295–313, 2000.

[52] A. P. Korostelev and A. B. Tsybakov, *Minimax Theory of Image Reconstruction*. New York: Springer-Verlag, 1993.

[53] E. Candes and D. Donoho, "Curvelets and curvilinear integrals," *J. Approx. Theory*, vol. 113, pp. 59–90, 2000.

[54] T. Hagerup and C. Rüb, "A guided tour of Chernoff bounds," *Inform. Process. Lett.*, vol. 33, no. 6, pp. 305–308, 1990.

[55] H. Chernoff, "A measure of asymptotic efficiency of tests of a hypothesis based on the sum of observations," *Annals of Mathematical Statistics*, vol. 23, pp. 493–507, 1952.

[56] M. Okamoto, "Some inequalities relating to the partial sum of binomial probabilites," *Annals of the Institute of Statistical Mathematics*, vol. 10, pp. 29–35, 1958.

[57] B. Yu, "Assouad, Fano, and Le Cam," in *Festschrift for Lucien Le Cam*, D. Pollard, E. Torgersen, and G. Yang, Eds. Springer-Verlag, 1997, pp. 423–435.

[58] C. Huber, "Lower bounds for function estimation," in *Festschrift for Lucien Le Cam*, D. Pollard, E. Torgersen, and G. Yang, Eds. Springer-Verlag, 1997, pp. 245–258.