RICE UNIVERSITY

Dyadic Decision Trees

by

Clayton D. Scott

A Thesis Submitted
in Partial Fulfillment of the
Requirements for the Degree

Doctor of Philosophy

Approved, Thesis Committee:

Robert D. Nowak, Chair
Adjunct Professor of Electrical and Computer Engineering
Associate Professor of Electrical and Computer Engineering, University of Wisconsin at Madison

---

Don H. Johnson
J.S. Abercrombie Professor of Electrical and Computer Engineering and of Statistics

---

David W. Scott
Noah Harding Professor of Statistics

Houston, Texas

August, 2004

Abstract

Dyadic Decision Trees

by

Clayton D. Scott

This thesis introduces a new family of classifiers called dyadic decision trees (DDTs) and develops their theoretical properties within the framework of statistical learning theory.

First, we show that DDTs achieve optimal rates of convergence for a broad range of classification problems and are 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 selected by penalized empirical risk minimization using a new data-dependent penalty and may be computed exactly and efficiently. DDTs are the first practical classifier known to achieve optimal rates for the diverse class of distributions studied here. This is also the first study (of which we are aware) to consider rates for adaptation to data dimension and relevant features.

Second, we develop the theory of statistical learning using the Neyman-Pearson (NP) criterion. It is shown that concepts from learning with a Bayes error criterion have counterparts in the NP context. Thus, we consider constrained versions of empirical risk minimization and structural risk minimization (NP-SRM), proving performance guarantees for both. We also provide a general condition under which NP-SRM leads to strong universal consistency. Finally, we apply NP-SRM to dyadic decision trees, deriving rates of convergence and providing an explicit algorithm to implement NP-SRM in this setting.

Third, we study the problem of pruning a binary tree by minimizing an objective function that sums an additive cost with a non-additive penalty depending only on tree size. We focus on sub-additive penalties which are motivated by theoretical

results for dyadic and other decision trees. Consider the family of optimal prunings generated by varying the scalar multiplier of a sub-additive penalty. We show this family is a subset of the analogous family produced by an additive penalty. This implies (by known results for additive penalties) that the trees generated by a sub-additive penalty (1) are nested; (2) are unique; and (3) can be computed efficiently. It also implies that an additive penalty is preferable when using cross-validation to select from the family of possible prunings.

Acknowledgments

I would like to thank all who have advised, inspired, and challenged me in my years as a graduate student, including Don Johnson, David Scott, Rich Baraniuk, Mike Orchard, and of course my advisor Rob Nowak. Thank you Rob for your enthusiasm, devotion, and most importantly, your time. You are a great friend and role model.

Many thanks to Rebecca Willett and Rui Castro for their careful reading of Chapters 2 and 3. This thesis is far clearer and has far fewer errors as a result.

Finally, I owe a heartfelt thanks to all of my family for their support, especially my parents, and most important of all, my wife Alena.

Contents

Illustrations

Chapter 1

Introduction

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, very few practical decision tree algorithms have been shown to possess desirable theoretical properties such as consistency and rates of convergence. This thesis introduces a new family of practical classifiers called dyadic decision trees (DDTs) and establishes their theoretical properties in the context of statistical learning theory.

The three main chapters of the thesis each comprise a self-contained paper related to DDTs. The material should be accessible to anyone familiar with the probabilistic theory of classification as put forth, for example, in the masterful textbook of Devroye, Györfi, and Lugosi (1996). Each of the three chapters begins with a detailed summary of the material presented. Here we simply outline the content with broad strokes and try to provide some context.

Chapter 2 is the main component of the thesis. There dyadic decision trees are introduced and their performance is characterized through generalization error bounds and rates of convergence. DDTs are seen to be the first computationally feasible classifier that achieves optimal rates of convergence of a broad range of classification problems.

In Chapter 3 we develop a theoretical framework for statistical learning using the Neyman-Pearson (NP) criterion, which permits the design of classifiers in situations where errors for different classes carry different weights or a priori class probabilities are unknown. This chapter includes a section on DDTs, including rates of convergence and a practical algorithm, but the scope extends beyond decision trees. Indeed it serves as a foundational work for NP learning by reformulating many well known results and techniques from the Bayes error setting.

Chapter 4 is somewhat more modest in scope. In the course of investigating theoretically motivated pruning rules for decision trees, new algorithms needed to be developed. In many cases these algorithms turned out to be rediscovered versions or

simple extensions of existing algorithms. In one case, however, a novel algorithm was found. The existence of this algorithm only appeared as a consequence of a curious theorem about pruning. The theorem in turn appeared after an unexpected pattern emerged while performing some simulations. The proof of the theorem is entirely elementary, requiring only high-school algebra.

The final chapter offers some broad conclusions about common themes that have emerged in this study and speculates on future research directions.

Chapter 2

Minimax Optimal Classification with Dyadic Decision Trees

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 we show that a new family of decision trees, called 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 the behavior of the a posteriori probability in the vicinity of 1/2; (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 data dimension and relevant features.

2.1 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 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 (Tsybakov, 2004).

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 "effective" 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 specificly 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 Barron, 1991; Lugosi and Zeger, 1996; Vapnik, 1982) are proportional to the square root of the size of the tree (number of leaf nodes) and apparently fail to provide optimal rates (Scott and Nowak, 2002). 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 (Cohen, Dahmen, Daubechies, and DeVore, 2001; DeVore, 1998; Donoho, 1999). The dyadic structure also leads to computationally tractable classifiers based on algorithms akin to fast wavelet and multiresolution transforms (Mallat, 1998). 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 prefered 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 2.2 defines dyadic decision trees. Section 2.3 presents risk bounds and an oracle inequality for DDTs. Section 2.4 reviews the work of Mammen and Tsybakov (1999) and Tsybakov (2004) and defines regularity assumptions that help us quantify the four conditions outlined above. Section 2.5 presents theorems demonstrating the optimality (and adaptivity) of DDTs under these four conditions. Section 2.6 discusses algorithmic and practical issues related to DDTs. Section 2.7 offers conclusions and discusses directions for future research. The proofs are gathered in Section 2.8.

2.1.1 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 $\mathcal{Y} = \{0, 1, \ldots, B-1\}$ 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 condition distribution of Y given X, respectively.

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 2.4.

2.1.2 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 (1983) demonstrates minimax optimal rates under smoothness assumptions on the class-conditional densities. Yang (1999) shows that for $\eta(x)$ in certain smoothness classes minimax optimal rates are achieved by appropriate plug-in rules. Both Marron and Yang 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 affect on the difficulty of the problem.

Tsybakov and collaborators replace global constraints on the distribution by re-

strictions 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 (Audibert, 2004; Mammen and Tsybakov, 1999; Tsybakov, 2004). Tsybakov and van de Geer (2004) 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 2.4). Other authors derive rates for existing practical discrimination rules, but these rates are not known to be optimal in the minimax sense considered here (Bartlett, Jordan, and McAuliffe, 2003; Blanchard, Lugosi, and Vayatis, 2003; Scovel and Steinwart, 2004; Wu, Ying, and Zhou, 2004; Zhou and Jetter, 2004).

Our contribution is to demonstrate practical and implementable discrimination rules that adaptively achieve 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 (Scott and Nowak, 2002, 2003) and demonstrated the (non-adaptive) near-minimax optimality of DDTs in a very special case of the general classes considered herein (Scott and Nowak, 2003). We also simplify and improve the bounding techniques used in that work. A more detailed review of rates of convergence is given in Section 2.4.

2.1.3 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 Murthy (1998).

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 (1984) and Quinlan (1993). 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 (1999) demonstrate that greedy growing using impurity functions implicitly performs boosting. Unfortunately, as noted in Devroye et al. (1996, chap. 20), split selection using impurity functions cannot lead to consistent discrimination rules. For consistent growing schemes, see Devroye et al. (1996); Gordon and Olshen (1978).

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 Esposito, Malerba, and Semeraro, 1997). 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 (Quinlan, 1993). 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. Mansour (1997) and Kearns and Mansour (1998) both 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 involves penalized empirical risk minimization (ERM) whereby the pruned tree $\widehat{T}_n$ is the solution of

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

where T_{INIT} is the initial tree (from stage one) and $\Phi(T)$ 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 Breiman et al. (1984). In CCP $\Phi(T) = \alpha|T|$ 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 2.4 with $\kappa = 1$) have risk bounds been demonstrated for CCP (Blanchard, Schäfer, and Rozenholc, 2004).

Indeed, a penalty proportional to tree size appears to be inappropriate under more general conditions. Mansour and McAllester (2000) demonstrate error bounds for a "square root" penalty of the form $\Phi(T) = \alpha_n\sqrt{|T|}$. Nobel (2002) considers a similar penalty and showed consistency of $\widehat{T}_n$ under certain assumptions on the initial tree produced by the growing stage. Scott and Nowak (2002) 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 (1995) 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, Bartlett, Lee, and Mason (2003) 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 (2000) 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 appears to lead to suboptimal to suboptimal rates (Scott and Nowak, 2002, 2003).

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. In both cases we employ a *spatially adaptive*, *data-dependent* penalty. Like the penalty of Mansour and McAllester (2000), our penalty depends on more than just tree size and tends to favor unbalanced trees. In view of Berkman and Sandholm (1995), our penalty reflects a prior disposition toward Bayes decision boundaries that are well

approximated by unbalanced recursive dyadic partitions.

2.1.4 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 (1997) 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 Donoho and Johnstone (1994, 1995); Donoho, Johnstone, Kerkyacharian, and Picard (1995) initiated a flurry of research on wavelets in statistics, and the textbook by Mallat (1998) 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 Kolaczyk and Nowak (2004a,b) 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 piece-wise 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 (1999) 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 Blanchard et al., 2004; Kolaczyk and Nowak, 2004a,b).

2.2 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 2.1). 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 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 nonegative integer L. Define $\mathcal{T}_M$ to be the collection of all DDTs such that no terminal cell has a sidelength smaller

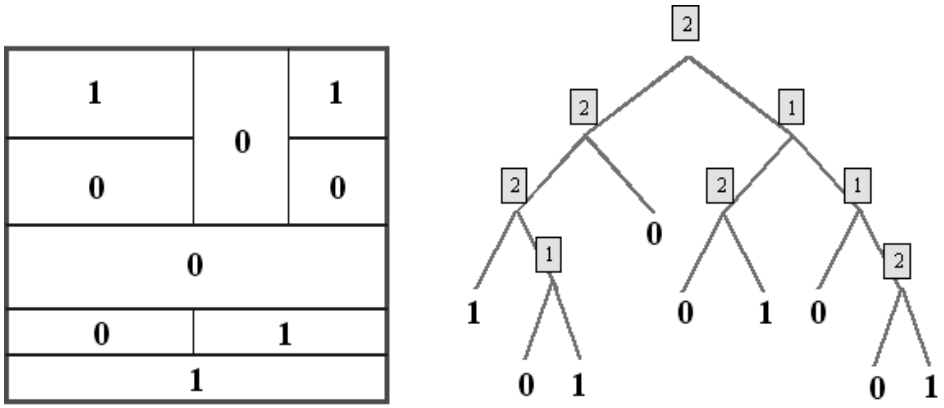


Figure 2.1 : A dyadic decision tree (right) with the associated recursive dyadic parition (left) assuming $d = 2$. Each internal node of the tree is labeled with an integer from 1 to d indicating the coordinate being split at that node. The leaf nodes are decorated with class labels.

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

$$\widehat{T}_n = \arg\min_{T \in \mathcal{T}_M} \widehat{R}_n(T) + \Phi(T) \tag{2.1}$$

where Φ is a "penalty" or regularization term specified below in Equation (2.9). Computational and experimental aspects of this rule are discussed in Section 2.6.

2.2.1 Cyclic DDTs

In earlier work we considered as special class of DDTs called cyclic DDTs (Scott and Nowak, 2002, 2003). 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 2.6.1).

2.3 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, but for the sake of simplicity and smaller constants we confine the discussion to DDTs. We begin by reviewing a penalty based only on tree size.

2.3.1 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 is taken from Mansour and McAllester (2000) but traces back to the work of Barron (1991). 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* (Cover and Thomas, 1991), $\|T\|$ may be defined as the codelength of a codeword for T in a prefix code for $\mathcal{T}$.

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

$$R(T) \leq \widehat{R}_n(T) + \sqrt{\frac{\|T\| \log 2 + \log(1/\delta)}{2n}} \quad \textit{for all } T \in \mathcal{T}. \tag{2.2}$$

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

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

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

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

□

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

Similar bounds (with larger constants) may be derived using VC theory and structural risk minimization (see for example Nobel, 2002; Scott and Nowak, 2002).

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 $k \log_2 B$ bits are 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 $\|T\| = 2k - 1 + k \log_2 B + (k-1) \log_2 d$ bits. In summary, it is possible to construct a prefix code for $\mathcal{T}$ with $\|T\| \leq (2 + \log_2 B + \log_2 d)|T|$. Thus the square root penalty is proportional to the square root of tree size.

2.3.2 A spatially adaptive penalty

The square root penalty appears to suffer from slack in the union bound. Many trees share leaf nodes, 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

$$R(T) - \widehat{R}_n(T) = \sum_{A \in \pi(T)} R(T, A) - \widehat{R}_n(T, A), \tag{2.3}$$

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 Mansour and McAllester (2000), 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 (2.3) does not yield optimal rates of convergence. Instead, we employ the *relative* Chernoff bound (see Lemma 4) which implies that with probability at least $1 - \delta$, for any fixed cell $A \in \mathcal{A}$ we have

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

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 $j-1$ 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 $\|A\| = 2j + (j-1)\log_2 d$ bits. In summary, it is possible to define a prefix code for $\mathcal{A}$ with $\|A\| \leq (2 + \log_2 d)j(A)$. With these definition it follows that

$$\sum_{A \in \mathcal{A}} 2^{-\|A\|} \leq 1. \tag{2.4}$$

Introduce the penalty

$$\Phi'_n(T) = \sum_{A \in \pi(T)} \sqrt{2p_A \frac{\|A\| \log 2 + \log(2/\delta)}{n}}. \tag{2.5}$$

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* $1 - \delta$*,*

$$R(T) \leq \widehat{R}_n(T) + \Phi'_n(T) \quad \textit{for all } T \in \mathcal{T}. \tag{2.6}$$

The proof may be found in Section 2.8.1

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

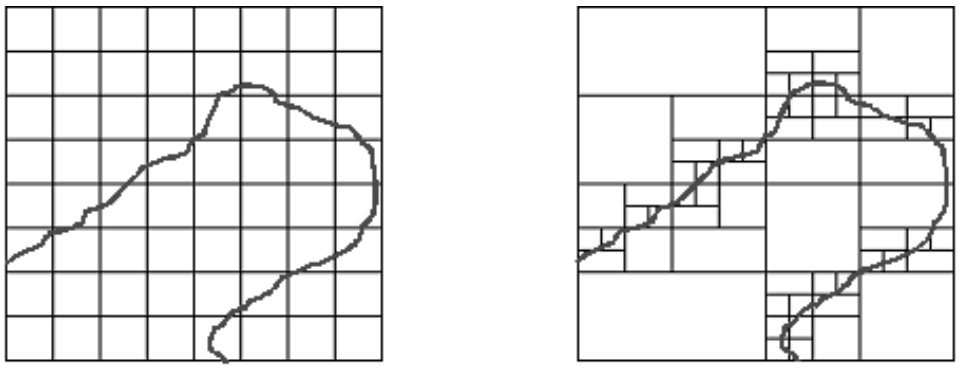


Figure 2.2 : An unbalanced tree/partition (right) can approximate a decision boundary much better than a balanced tree/partition (left) with the same number of leaf nodes. The suboptimal square root penalty penalizes these two trees equally, while the spatially adaptive penalty favors the unbalanced tree.

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 2.2).

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 (Blanchard et al., 2004; Gey and Nedelec, 2002; Kolaczyk and Nowak, 2004b). 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 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 inhibits 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 Tsybakov and van de Geer (2004), is based on approximating the Bayes decision boundary (assumed to be a boundary fragment; see Section 2.4) with a wavelet series.

Remark 1. In addition to different techniques for bounding the local error deviance, the bound of Mansour and McAllester 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 framents, 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.

2.3.3 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 $1-\delta$,*

$$p_A \leq \hat{p}'_A(\delta) \quad \textit{for all } A \in \mathcal{A}, \tag{2.7}$$

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

$$\hat{p}_A \leq p'_A(\delta) \quad \textit{for all } A \in \mathcal{A} \tag{2.8}$$

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

$$\Phi(T) = \sum_{A \in \pi(T)} \sqrt{2\hat{p}'_A(\delta) \frac{\|A\| \log 2 + \log(2/\delta)}{n}}. \tag{2.9}$$

Combining Theorem 1 and Lemma 1 produces the following.

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

$$R(T) \leq \widehat{R}_n(T) + \Phi(T) \quad \textit{for all } T \in \mathcal{T}. \tag{2.10}$$

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

Remark 2. 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.

2.3.4 An Oracle Inequality

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

Theorem 3. *Let $\widehat{T}_n$ be as in (2.1) with Φ as in (2.9). Define*

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

With probability at least $1 - 4/n$ over the training sample

$$R(\widehat{T}_n) - R^* \leq \min_{T \in \mathcal{T}_M} \Big(R(T) - R^* + \tilde{\Phi}_n(T) \Big) + \sqrt{\frac{\log n}{2n}}. \tag{2.11}$$

As a consequence,

$$\mathbb{E}^n\{R(\widehat{T}_n)\} - R^* \leq \min_{T \in \mathcal{T}_M} \left(R(T) - R^* + \tilde{\Phi}_n(T) \right) + \sqrt{\frac{\log n}{2n}} + \frac{4}{n}. \tag{2.12}$$

The proof is given in Section 2.8.3 Note that these inequalities involve the uncomputable penalty $\tilde{\Phi}_n$. A similar theorem with Φ 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 Bartlett, Boucheron, and Lugosi (2002); Bousquet, Boucheron, and Lugosi (2004).

2.4 Rates of Convergence Under Complexity and Noise Assumptions

We study rates of convergence for classes of distributions based on the work Mammen and Tsybakov (1999) and Tsybakov (2004). In this section we review their classes and propose modifications pertinent to DDTs. In general, these classes 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. 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 $G^* = G_{f^*}$. 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 all n. 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

$h(\epsilon) \preccurlyeq g(\epsilon)$.

2.4.1 Complexity Assumptions

Complexity assumptions restrict the complexity (regularity) of the Bayes decision boundary ∂G^*. Let $\bar{d}(\cdot,\cdot)$ be a pseudo-metric[†] 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 $\bar{d} = \lambda(\Delta(G, G'))$.

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 $H(\epsilon, \mathcal{G}, \bar{d}) \asymp \epsilon^{-\rho}$ as $\epsilon \to 0$.

Denote by $N_B(\epsilon, \mathcal{G}, \bar{d})$ the minimum number of pairs (G_j^-, G_j^+) such that (i) for each j, $\bar{d}(G_j^-, G_j^+) \leq \epsilon$; and (ii) for all $G \in \mathcal{G}$, there exists j such that $G_j^- \subseteq G \subseteq G_j^+$. Define the *bracketing entropy* of $\mathcal{G}$ with respect to $\bar{d}$ to be $H_B(\delta, \mathcal{G}, \bar{d}) = \log N_B(\delta, \mathcal{G}, \bar{d})$. We say $\mathcal{G}$ has *bracketing complexity* $\rho > 0$ with respect to $\bar{d}$ if $H_B(\epsilon, \mathcal{G}, \bar{d}) \asymp \epsilon^{-\rho}$ as $\epsilon \to 0$. Note that $H(\epsilon, \mathcal{G}, \bar{d}) \preccurlyeq H_B(\epsilon, \mathcal{G}, \bar{d})$.

Mammen and Tsybakov (1999) cite several examples of $\mathcal{G}$ with known complexities. 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 not exceeding 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 (1974) 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 $\gamma \geq 1$.

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

A third complexity assumption 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. It is also convenient for approximation by DDTs. 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* $\mathcal{G}_{\text{BOX}}(c_1)$ 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 $\gamma \leq 1$ 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 $g \in \Sigma(\gamma, c_1)$ and $\gamma \leq 1$. 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 $\mathcal{P}_m$. □

2.4.2 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

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

[‡]Similar generalizations for $\gamma < 1$ are possible but unwieldy.

To illuminate this condition, note

$$\begin{aligned} R(f) - R^* &= \int_{\Delta(f,f^*)} 2|\eta(x) - 1/2| d\mathbb{P}_X \\ &\leq \int_{\Delta(f,f^*)} d\mathbb{P}_X \\ &= \mathbb{P}_X(\Delta(f, f^*)) \end{aligned}$$

(see Devroye et al., 1996, p. 10). Thus, Tsybakov's noise condition constrains the regularity of $\eta(x)$ near the decision boundary. The case $\kappa = \infty$ is the high noise case and imposes no constraint on the distribution (provided $c_2 \geq 1$), and $\kappa = 1$ is the "low noise" case and implies a jump of $\eta(x)$ at the Bayes decision boundary (Bartlett et al., 2003). See Audibert (2004); Bartlett et al. (2003); Bousquet et al. (2004) for further discussion.

Mammen and Tsybakov (1999) and Tsybakov (2004) provide a lower bound for classification under boundary fragment and noise assumptions. 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*

$$\inf_{\widehat{f}_n} \sup_{\mathcal{D}_{\text{BF}}(\gamma,\kappa)} \left[\mathbb{E}^n \{ R(\widehat{f}_n) \} - R^* \right] \succcurlyeq n^{-\kappa/(2\kappa+\rho-1)}. \tag{2.13}$$

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 probability measures $\mathbb{P} \in \mathcal{D}_{\text{BF}}(\gamma, \kappa)$.

Mammen and Tsybakov (1999) demonstrate that empirical risk minimization (ERM) over $\mathcal{G}_{\text{BF}}(\gamma, c_1)$ yields a classifier achieving this rate when $\rho < 1$. Tsybakov

(2004) shows that ERM over a suitable ϵ-bracketing net of $\mathcal{G}_{\text{BF}}(\gamma, c_1)$ also achieves the minimax rate for $\rho < 1$. Tsybakov and van de Geer (2004) 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 (2004) 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 likely that such algorithms exist.

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

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

In particular we have

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

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

2.4.3 Excluding Low Noise Levels

It is important to understand how the minimax rate changes for different complexity and noise parameters. Clearly, for fixed ρ, the lower bound in (2.13) tends to $n^{-1/2}$ as $\kappa \to \infty$. If $\rho < 1$, the lower bound increases as $\kappa \to \infty$, and the "easiest" case, meaning the smallest lower bound, occurs when $\kappa = 1$. In contrast, if $\rho > 1$, the lower bound *decreases* as $\kappa \to \infty$, and the easiest case occurs when $\kappa = \infty$. Thus, when the smoothness γ exceeds $d - 1$, convergence to the Bayes error is faster when there is less noise near the decision boundary. When $\gamma < d - 1$, on the other hand, high noise actually increases the rate. Of course, high noise does not make learning easier in an absolute sense, only relative to Bayes error, which is larger in high noise settings.

In light of the above, care must be taken when speaking about the minimax rate when $\rho > 1$. From the definition, $\mathcal{D}_{\text{BF}}(\gamma, 1) \subset \mathcal{D}_{\text{BF}}(\gamma, \kappa)$ for any $\kappa > 1$. Hence the

minimax rate for $\mathcal{D}_{\text{BF}}(\gamma, \kappa)$ can be no faster than $n^{-1/(1+\rho)}$, the minimax rate for $\mathcal{D}_{\text{BF}}(\gamma, 1)$. To allow the possibility of a discrimination rule actually achieving rates faster than $n^{-1/(1+\rho)}$ when $\rho > 1$, clearly an additional assumption must be made. In particular, we must exclude those distributions with small κ that cause slower rates when $\rho > 1$.

2.4.4 Excluding Low Noise for the Box-Counting Class

Since recursive dyadic partitions can well approximate G^* with smoothness $\gamma \leq 1$ ($\rho \geq (d-1)/\gamma \geq 1$) we are interested in excluding low noise levels. In this subsection we introduce a new condition that excludes low noise under a concrete complexity assumption, namely, the box-counting assumption.

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 $\mathcal{T}_m(c_1) = \{T \in \mathcal{T}_m : K_j(T) \leq 2c_1 2^{\lceil j/d \rceil (d-1)} \ \ \forall j = 1, \ldots, d \log_2 m\}$. 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*

$$\lambda(\Delta(G, G_T)) \leq \frac{c_1}{m}$$

where $G_T = \{x \in [0,1]^d : T(x) = 1\}$.

See Section 2.8.4 for the proof. The lemma says that $\mathcal{T}_m(c_1)$ is an ϵ-net (without bracketing and with respect to Lebesgue measure) for $\mathcal{G}_{\text{BOX}}(c_1)$, with $\epsilon = c_1/m$.

Define $\mathcal{D}^m_{\text{BOX}}(\kappa) = \mathcal{D}^m_{\text{BOX}}(\kappa, c_0, c_1, c_2)$ 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 $A \subseteq \mathcal{X}$

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

2B There exists $T' \in \mathcal{T}_m(c_1)$ such that

$$(R(T') - R^*)^{1/\kappa} \leq \frac{c_2}{m}.$$

In a sense, **2B** is almost the negation of **2A**, with the main difference being that we require the classifier violating **2A** to also approximate f^* well. To illuminate this further, we remark that our proofs would still hold if **2B** was replaced by

2B' There exists $T'_m \in \mathcal{T}_m(c_1)$ such that

$$(R(T'_m) - R^*)^{1/\kappa} \leq c'_2 \mathbb{P}_X(\Delta(T'_m, f^*)) \leq \frac{c''_2}{m}.$$

This modified formulation provides an intuitive explanation of how low noise levels are excluded. To see this, suppose $\mathbb{P}$ has Tsybakov noise exponent $\kappa' < \kappa$. By **2A** and **2B'** there exists C such that

$$(R(T'_m) - R^*)^{1/\kappa} \leq C(R(T'_m) - R^*)^{1/\kappa'}$$

which implies

$$R(T'_m) - R^* \geq C^{-\frac{1}{1/\kappa' - 1/\kappa}} > 0.$$

However, the left hand side goes to zero as $m \to \infty$, a contradiction.

Theorem 5. *Let $d \geq 2$ and $\kappa > 1$. If $m \asymp n^{1/(2\kappa+d-2)}$ then*

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

The proof relies on ideas from Audibert (2004) and is given in Section 2.8.5. 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, but that extends beyond the scope of this paper.

We remark that Tsybakov and van de Geer (2004) and Audibert (2004) introduce a different condition, what we call a *two-sided noise exponent* κ, to exclude low

noise levels. Namely, let $\mathcal{F}_n$ be a collection of candidate classifiers, and consider all distributions such that for some constant $c_2 > 0$,

$$\frac{1}{c_2}(R(f) - R^*)^{1/\kappa} \leq \mathbb{P}_X(\Delta(f, f^*)) \leq c_2(R(f) - R^*)^{1/\kappa} \text{ for all } f \in \mathcal{F}_n. \tag{2.14}$$

Such a condition does eliminate "low noise" distributions, but it also eliminates "high noise" (in fact, forcing $\eta(x)$ to behave very uniformly near ∂G^*), and is stronger than we need.

2.5 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, effective 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.

2.5.1 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 (2.1) with Φ*

as in (2.9). If $\kappa > 1$ *and* $m \asymp (n/\log n)^{1/(2\kappa+d-2)}$ *then*

$$\sup_{\mathcal{D}^m_{\mathrm{BOX}}(\kappa)} \left[\mathbb{E}^n\{R(\widehat{T}_n)\} - R^*\right] \preccurlyeq \left(\frac{\log n}{n}\right)^{\frac{\kappa}{2\kappa+d-2}}. \tag{2.15}$$

The complexity penalized DDT $\widehat{T}_n$ is adaptive in the sense that it is constructed without knowledge of m, 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 (2.15) is achieved. See Section 2.8.6 for the proof.

The reader may notice that the choice of m in the lower bound (Theorem 5) and upper bound (Theorem 6) differ by a log factor, and hence two different sequences are in fact being bounded. Taking $m \asymp n^{1/(2\kappa+d-2)}$ in Theorem 6 leads to an upper bound of

$$\begin{aligned}\sup_{\mathcal{D}^m_{\mathrm{BOX}}(\kappa)} \left[\mathbb{E}^n\{R(\widehat{T}_n)\} - R^*\right] &\preccurlyeq n^{-\frac{\kappa}{2\kappa+d-2}} \log n \\ &= \left(\frac{\log n}{n}\right)^{\frac{\kappa}{2\kappa+d-2}} (\log n)^{\frac{\kappa+d-2}{2\kappa+d-2}},\end{aligned}$$

which is within a log factor of the minimax rate.

2.5.2 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 2.3 (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 $1 \le d' \le d$. 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 $A \in \mathcal{P}_m$, $\mathbb{P}_X(A) \le c_0 m^{-d'}$.

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

We refer to d' as the *effective data dimension*. In practice, it may be more likely that data "almost" lie on a d'-dimensional manifold. Nonetheless, we feel the adaptivity of

DDTs to data dimension depicted in the theorem below reflects a similar capability in less ideal settings.

Proposition 2. *Let $d \geq 2$. 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. 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 $\zeta \in [0,1]^d$ 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.

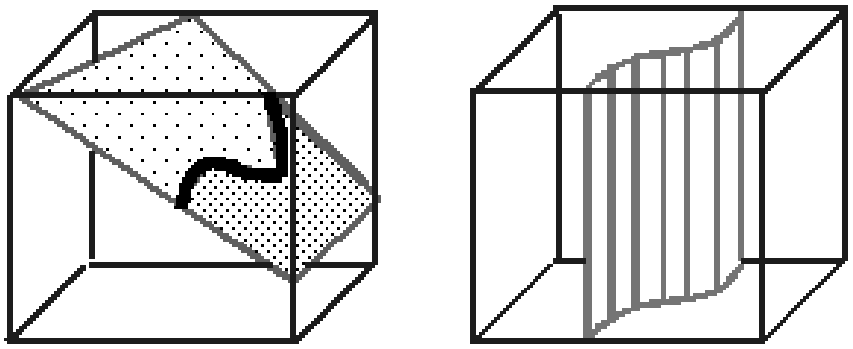
Theorem 7. *Choose M such that $M \succcurlyeq n/\log n$. Define $\widehat{T}_n$ as in (2.1) with Φ as in (2.9). If assumptions* **0B** *and* **1C** *hold and $d' \geq 2$ then*

$$\mathbb{E}^n \{ R(\widehat{T}_n) \} - R^* \preccurlyeq \left(\frac{\log n}{n} \right)^{\frac{1}{d'}}. \tag{2.16}$$

Again, $\widehat{T}_n$ is adaptive in that it does not require knowledge of the effective dimension d' or the constants c_0, c_1. The proof may be found in Section 2.8.7.

2.5.3 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 2.3 (b)).



(a) (b)

Figure 2.3 : Cartoons illustrating effective 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.

Proposition 3. *Let $d \geq 2$. 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''. 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 (2.1) with Φ as in (2.9). If all but $d'' \geq 2$ of the features X^i are independent of Y, and* **0A** *and* **1B** *hold, then*

$$\mathbb{E}^n \{ R(\widehat{T}_n) \} - R^* \preccurlyeq \left(\frac{\log n}{n} \right)^{\frac{1}{d''}}. \tag{2.17}$$

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 to DDTs' feature rejection capability.

2.5.4 Adapting to Bayes Decision Boundary Smoothness

Thus far in this section we have assumed G^* satisfies a box-couting (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). Our result applies under assumptions **0A** and

1D One coordinate of ∂G^* is a Hölder-γ function of the others.

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 Tsybakov and van de Geer (2004) 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 (2.1) with Φ as in (2.9). If assumptions* **0A** *and* **1D** *(with $\gamma \leq 1$) hold, then*

$$\mathbb{E}^n\{R(\widehat{T}_n)\} - R^* \preccurlyeq \left(\frac{\log n}{n}\right)^{\frac{\gamma}{\gamma+d-1}}. \tag{2.18}$$

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.

2.6 Computational Considerations

The data-dependent, spatially adaptive penalty in (2.9) 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. (2004) show that an algorithm of Donoho (1997), 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. First, we introduce some notation. 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 $\widehat{R}_n(T_A) + \Phi(T_A)$, 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

$$\begin{aligned} T_A^* \;=\; & \arg\min\{\widehat{R}_n(T_A) + \Phi(T_A) \mid T_A = \{A\} \text{ or} \\ & T_A = \textsc{merge}(A, T^*_{A^{s,1}}, T^*_{A^{s,2}}), s = 1, \ldots, d\}. \end{aligned}$$

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 Donoho (1997) 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). This was pointed out by Blanchard et al. (2004), who derive an implementation with $O(ndL^d \log(nL^d))$ complexity.

Although we are primarily concerned with theoretical properties of DDTs, we note that a recent experimental study by Schäfer, Blanchard, Rozenholc, and Müller

(2004) 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. Schäfer et al. (2004) report that when $d > 15$, memory and processor limitations necessitate preprocessing in the form of dimensionality reduction.

2.6.1 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 2.2.1) is substantially easier. The optimization in (2.1) 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 $O(nd \log n)$.

2.7 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 simulta-

neously. 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 (2.9)) 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 rule 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 leafs 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" (Bennett, Cristianini, Shawe-Taylor, and Wu, 2000) 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 (Korostelev and Tsybakov, 1993). Multiresolution methods known as "wedgelets" or "curvelets" (Candes and Donoho, 2000; Donoho, 1999) 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 (Scott and Nowak, 2003). Such trees yield faster rates, up to $(\log n/n)^{1/2}$, but in their current state do not give optimal rates and are computationally prohibitive. Recent risk bounds for polynomial-kernal support vector machines may offer a computationally tractable alternative to this approach (Zhou and Jetter, 2004). One way or another, we feel that dyadic decision trees, or possibly new variants thereof, hold promise to address these issues.

2.8 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 excercise 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 Hagerup and Rüb, 1990), the second is a standard (additive) Chernoff bound (Chernoff, 1952; Okamoto, 1958), and the last two were proved by Okamoto (1958).

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 $\epsilon > 0$*

$$\mathbb{P}^n \left\{ \hat{p} \leq (1-\epsilon)p \right\} \leq e^{-np\epsilon^2/2}, \tag{2.19}$$

$$\mathbb{P}^n \left\{ \hat{p} \geq p + \epsilon \right\} \leq e^{-2n\epsilon^2}, \tag{2.20}$$

$$\mathbb{P}^n \left\{ \sqrt{\hat{p}} \geq \sqrt{p} + \epsilon \right\} \leq e^{-2n\epsilon^2}, \tag{2.21}$$

$$\mathbb{P}^n \left\{ \sqrt{p} \geq \sqrt{\hat{p}} + \epsilon \right\} \leq e^{-n\epsilon^2}. \tag{2.22}$$

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 (2.19) with $\epsilon = \sqrt{\frac{2\log(1/\delta)}{pn}}$.

2.8.1 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 (2.6) 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 (2.4).

2.8.2 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 (2.21) 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 (2.4).

2.8.3 Proof of Theorem 3

Recall that in this and subsequent proofs we take $\delta = 1/n$ in the definition of Φ, p', and p'.

Let $T' \in \mathcal{T}$ be the tree minimizing the expression on the right-hand side of (2.12). By the additive Chernoff bound (2.20), with probability at least $1 - 1/n$,

$$\widehat{R}_n(T') \leq R(T') + \sqrt{\frac{\log n}{2n}}. \tag{2.23}$$

Take Ω to be the set of all Z^n such that the events in (2.8), (2.10), and (2.23) hold.

Then $\mathbb{P}(\Omega) \geq 1 - 4/n$. Given $Z^n \in \Omega$, we know

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

where the first inequality follows from (2.10), the second from (2.1), the third from (2.8), and the fourth from (2.23). 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

$$\begin{aligned} \mathbb{E}^n\{R(\widehat{T}_n)\} &= \mathbb{P}^n(\Omega)\mathbb{E}^n\{R(\widehat{T}_n) \,|\, \Omega\} + \mathbb{P}^n(\Omega^c)\mathbb{E}^n\{R(\widehat{T}_n) \,|\, \Omega^c\} \\ &\leq \mathbb{E}^n\{R(\widehat{T}_n) \,|\, \Omega\} + \frac{4}{n} \end{aligned}$$

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

2.8.4 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

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

2.8.5 Proof of Theorem 5

Audibert (2004) 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 behing 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 Yu (1997).

According to Audibert (2004, chap. 3, sec. 6.2), Assouad's lemma is inadequate for excluding low noise levels 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 Audibert (2004, chap. 3), with some variations, including a different version of Fano's lemma.

Our strategy is to construct a set of probability measures $\mathcal{D}_m \subset \mathcal{D}^m_{\text{BOX}}(\kappa)$ for which the lower bound holds. We proceed as follows. 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 of Huber (1997, 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 $b = an^{-(\kappa-1)/(2\kappa+d-2)}$. Let $\mathcal{G}'$ be as in Lemma 5 and define $\mathcal{D}'_m$ to be the set of all product measures $\mathbb{P}^n$ on $\mathcal{Z}^n$ 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. Since $m \asymp n^{1/(2\kappa+d-2)}$, there exist constants c and C such that

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

for m sufficiently large. By Lemma 3 $\mathcal{T}_m(c_1)$, there exists $T' \in \mathcal{T}_m(c_1)$ such that $\lambda(\Delta(T', f^*)) \leq \frac{c_1}{m}$. Now

$$\begin{aligned} R(T') - R^* &= \int_{\Delta(T', f^*)} 2|\eta(x) - 1/2| d\mathbb{P}_X \\ &= b\lambda(\Delta(T', f^*)) \\ &\leq \frac{bc_1}{m} \\ &\leq \frac{ac_1}{c} n^{-\kappa/(2\kappa+d-2)} \\ &\leq \frac{ac_1 C^\kappa}{c} m^{-\kappa}. \end{aligned}$$

Thus $(R(T') - R^*)^{1/\kappa} \leq \frac{c_2}{m}$ provided $a \leq c_2^\kappa c/(c_1 C^\kappa)$.

It remains to derive a lower bound for the expected excess risk. We employ the following generalization of Fano's lemma due to Yu (1997). 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 of $\mathbb{P}$.

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 seting 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 $\mathbb{P}_j^n, \mathbb{P}_{j'}^n \in \mathcal{D}_m$. We have

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

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

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

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

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

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

2.8.6 Proof of Theorem 6

Let $\mathbb{P}^n \in \mathcal{D}_{\text{BOX}}^m(\kappa, c_0, c_1, c_2)$. 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 bound on 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,

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

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

$$\tilde{\Phi}_n^2(T) = \sum_{A\in\pi(T)} \sqrt{\frac{\log n}{n}\cdot\frac{\log n}{n}}. \tag{2.24}$$

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 plugging $m \asymp (n/\log n)^{1/(2\kappa+d-2)}$ into the above bounds on approximation and estimation error.

2.8.7 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,*

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

Proof. We have

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

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 $p_A \le c_0 2^{-\lfloor j/d \rfloor d'}$ 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 (2.24). 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.

2.8.8 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, $1 \le \ell \le L$, 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,*

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

Proof. We have

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

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

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

2.8.9 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, $1 \le \ell \le L$, 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*

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

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 $\mathcal{P}^\gamma_m(j)$ to be the partition of $[0,1]^d$ formed by the set of cells in T'_{INIT} having depth j. Define $\mathcal{B}^\gamma_m(j)$ to be the set of cells in $\mathcal{P}^\gamma_m(j)$ that intersect ∂G^*. By construction of T' we have $K_j \le 2|\mathcal{B}^\gamma_m(j)|$. From the fact $|\mathcal{B}^\gamma_m(j)| \le |\mathcal{B}^\gamma_m(j+1)|$, we conclude

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

Thus it remains to show $2|\mathcal{B}^\gamma_m((\ell - \tilde{\ell})(d-1) + pd)| \le C 2^{(\ell - \tilde{\ell} + p)(d-1)}$ for each p. Each cell in $\mathcal{B}^\gamma_m((\ell - \tilde{\ell})(d-1) + pd)$ 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 sidelenth $2^{-(\ell - \tilde{\ell} + p)}$, 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 $\mathcal{B}^\gamma_m(p, \sigma) = \{U_\sigma \times V_\tau \in \mathcal{B}^\gamma_m((\ell - \tilde{\ell})(d-1) + pd) \mid \tau = 1, \ldots, 2^p\}$.

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

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

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 $|g(s) - g(s')| \le c_1 |s - s'|^\gamma$ 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

$$\begin{aligned}
R(T') - R^* &\leq 2c_0|\mathcal{B}_m^\gamma(J)|m^{-(d+\gamma-1)} \\
&\leq C2^{\ell(d-1)}m^{-(d+\gamma-1)} \\
&= Cm^{d-1}m^{-(d+\gamma-1)} \\
&= Cm^{-\gamma}.
\end{aligned}$$

□

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.

Chapter 3

Neyman-Pearson Learning with Application to Dyadic Decision Trees

The Neyman-Pearson (NP) criterion permits the design of classifiers in situations where errors for different classes carry different weights or a priori class probabilities are unknown. This chapter develops the theory of learning classifiers from training data in the NP setting. Building on a "fundamental lemma" of Cannon, Howse, Hush, and Scovel (2002) (while improving upon one of their main results), we demonstrate that several concepts from learning with the probability of error criterion have counterparts in the NP context. Thus, we consider constrained versions of empirical risk minimization (NP-ERM) and structural risk minimization (NP-SRM), proving performance guarantees for both. We also provide a general condition under which NP-SRM leads to strong universal consistency. Finally, we apply NP-SRM to decision trees, deriving rates of convergence and providing an explicit algorithm to implement NP-SRM in this setting.

3.1 Introduction

The Neyman-Pearson (NP) criterion allows for the design of decision rules in situations where a priori class probabilities are unknown or when it is far more costly to mislabel one class than another. Applications where such situations exist include fraud detection, spam filtering, machine monitoring, and target recognition. When the class conditional densities of the observed data (signal) are known, the optimal detector with respect to the Neyman-Pearson criterion leads to a likelihood ratio test. When the class-conditional densities are known to belong to a certain parametric family, generalized likelihood ratios or UMP tests often suffice. When no distributional assumption can be made, it may be necessary to learn a decision rule from training data.

In this paper we study the problem of learning decision rules from training data using the Neyman-Pearson criterion. In learning theory, classifiers are typically de-

signed with the goal of minimizing the (expected) probability of error which assumes errors from each class are given equal weight. The NP criterion provides a setting for learning when this assumption does not hold.

A recent body of work known as cost-sensitive learning has addressed this issue by modifying the standard '0-1' loss function to a weighted Bayes cost (see Domingos, 1999; Elkan, 2001; Margineantu, 2002; Zadrozny, Langford, and Abe, 2003, and references therein). Cost-sensitive learners assume the relative costs for different classes are known. Moreover, these algorithms rely on training data to estimate (either explicitly or implicitly) the a priori class probabilities.

Cost-sensitive and NP learning are fundamentally different approaches that have differing pros and cons. In some situations it may be difficult to (objectively) assign costs. For instance, how much greater is the cost of failing to detect a malignant tumor compared to the cost of erroneously flagging a benign tumor? In addition, a priori class probabilities may not accurately be reflected by their sample-based estimates. Consider, for example, the case where one class has very few representatives in the training set simply because it is very expensive to gather that kind of data. In these scenarios it may be more practical to take a Neyman-Pearson approach by constraining the error of the more costly class. On the other hand, when costs and a priori probabilities can be accurately estimated, cost-sensitive learning may be more appropriate.

Neyman-Pearson learning has been studied previously by Cannon et al. (2002). There an analysis was given of a constrained form of empirical risk minimization (ERM) that we call NP-ERM. The present work builds on their theoretical foundations in several respects. First, using different bounding techniques, we derive predictive error bounds for NP-ERM that are non-trivial for substantially smaller sample sizes. Second, while Cannon et al. consider only learning from fixed Vapnik-Chervonenkis (VC) classes, we introduce a constrained form of structural risk minimization, NP-SRM, that automatically balances model complexity and training error, a feature necessary for consistency. Third, assuming mild regularity conditions on the underlying distribution, we derive rates of convergence for a certain family of decision trees called dyadic decision trees. Finally, we present an exact and computationally efficient algorithm for implementing Neyman-Pearson learning using dyadic decision trees. The algorithm is demonstrated on synthetic data. To our knowledge this is only

the second study to consider learning from training data with the NP criterion, and the first to consider model selection, consistency, rates of convergence, and practical algorithms.

3.1.1 Notation

We focus exclusively on *binary* classification, although extensions to multi-class settings are possible. Let $\mathcal{X}$ be a set and let $Z = (X, Y)$ be a random variable taking values in $\mathcal{Z} = \mathcal{X} \times \{0, 1\}$. The variable X corresponds to the observed signal and Y is the class label associated with X. In settings where the Neyman-Pearson criterion is appropriate, $Y = 0$ typically corresponds to the null hypothesis.

A decision rule is a Borel measurable function $h : \mathcal{X} \to \{0, 1\}$ mapping signals to class labels. The terms "decision rule" and "classifier" will be used interchangeably. In the standard learning theory setup, the performance of h is measured by the probability error (or Bayes error), $R(h) = \mathbb{P}(h(X) \neq Y)$. Here $\mathbb{P}$ denotes the probability measure for Z. We will focus instead on the false alarm and miss probabilities denoted by

$$R_j(h) = \mathbb{P}_{X|Y=j}(h(X) \neq j \mid Y = j)$$

for $j = 0$ and $j = 1$, respectively. Note that $R(h) = \pi_0 R_0(h) + \pi_1 R_1(h)$ where $\pi_j = \mathbb{P}_Y(Y = j)$ is the (unknown) a priori probability of class j.

Let $\alpha \in [0, 1]$ be a user-specified false alarm threshold. The Neyman-Pearson criterion seeks the classifier g^* minimizing $R_1(h)$ over all h such that $R_0(h) \leq \alpha$. When X is univariate ($d = 1$), if $F^j(x)$ denotes the cumulative distribution function for X given $Y = j$, then the Neyman-Pearson lemma (Trees, 2001) states that $g^*(x) = \mathbb{I}_{\{\Lambda(x) > \eta\}}$. Here $\mathbb{I}$ denotes the indicator function, $\Lambda(x) = dF^1(x)/dF^0(x)$ is the likelihood ratio, and η is defined implicitly by the integral equation $\int_{\Lambda(x)>\eta} dF^0(x) = \alpha$.

In this paper we are interested in the case where our only information about $F^j(x)$ is a finite training sample. Let $Z^n = \{(X_i, Y_i)\}_{i=1}^n$ be a collection of n iid samples of $Z = (X, Y)$. Since Z^n is merely a finite sampling of the underlying distribution, we cannot hope to reconstruct g^* exactly. However, as we will see, it is possible to construct a rule $\widehat{h}_n$ that always does "about as well" as g^*. Let $\mathbb{P}^n$ denote the product measure on $\mathcal{Z}^n$ induced by $\mathbb{P}$. Let $\mathbb{E}^n$ denote expectation with respect to $\mathbb{P}^n$. Let $\widehat{h}_n$ denote a decision rule constructed from a training sample.

Our learning strategies will involving the following sample-dependent quantities. For $j = 0, 1$, let

$$n_j = \sum_{i=1}^{n} \mathbb{I}_{\{Y_i = j\}}$$

be the number of samples from class j. Let

$$\widehat{R}_j(h) = \frac{1}{n_j} \sum_{i:Y_i=j} \mathbb{I}_{\{h(X_i) \neq Y_i\}}$$

denote the empirical false alarm and miss probabilities, corresponding to $j = 0$ and $j = 1$, respectively. Given a class of decision rules $\mathcal{H}$, define $\mathcal{H}_0 = \{h \in \mathcal{H} : R_0(h) \leq \alpha\}$, and $h^* = \arg\min\{R_1(h) : h \in \mathcal{H}_0\}$. Finally, set $R_1^* = R_1(g^*)$ to be the miss probability of the optimal classifier g^*.

3.1.2 Problem Statement

Our goal is to find decision rules $\widehat{h}_n \in \mathcal{H}$ satisfying perfomance guarantees characteristic of learning theory:

PAC bounds : $\widehat{h}_n$ is "probably approximately correct" in the sense that given $\epsilon_0, \epsilon_1 > 0$, there exist δ_0, δ_1 such that for all n,

$$\mathbb{P}^n(R_0(\widehat{h}_n) - \alpha > \epsilon_0) \leq \delta_0$$

and

$$\mathbb{P}^n(R_1(\widehat{h}_n) - R_1(h^*) > \epsilon_1) \leq \delta_1.$$

Moreover, δ_0, δ_1 decay exponentially fast as functions of increasing ϵ_0, ϵ_1.

Oracle inequalities : $\widehat{h}_n$ does about as well as the best classifier in $\mathcal{H} = \cup_{k=1}^{K} \mathcal{H}^k$. In particular we will show that with high probability, both

$$R_1(\widehat{h}_n) - R_1^* \leq \inf_{1 \leq k \leq K} \left(\epsilon_1(n_1, k) + \inf_{h \in \mathcal{H}_0^k} R_1(h) - R_1^* \right)$$

and

$$R_0(\widehat{h}_n) - \alpha \leq \epsilon_0(n_0, K)$$

hold, where $\epsilon_j(n_j, k)$ tends to zero at a certain rate depending on the choice of $\mathcal{H}^k$ (see Section 3.3 for details).

Consistency : If $\mathcal{H}$ grows (as a function of n) in a suitable way, then $\widehat{h}_n$ is *strongly universally consistent* (Devroye et al., 1996, chap. 6) in the sense that

$$\lim_{n\to\infty} R_0(\widehat{h}_n) \leq \alpha \quad \text{with probability 1}$$

and

$$\lim_{n\to\infty} R_1(\widehat{h}_n) \leq R_1^* \quad \text{with probability 1}$$

for all distributions of Z.

Rates of Convergence : Under mild regularity conditions, there exist functions $r_0(n)$ and $r_1(n)$ tending to zero at a polynomial rate such that

$$\mathbb{E}^n\{R_0(h)\} - \alpha \preccurlyeq r_0(n)$$

and

$$\mathbb{E}^n\{R_1(h)\} - R_1^* \preccurlyeq r_1(n).$$

We write $a_n \preccurlyeq b_n$ when $a_n = O(b_n)$ and $a_n \asymp b_n$ if both $a_n \preccurlyeq b_n$ and $b_n \preccurlyeq a_n$.

Finally, we would like to find rules satisfying the above that can be implemented efficiently.

3.2 Neyman-Pearson and Empirical Risk Minimization

In this section we review the work of Cannon et al. who study Neyman-Pearson learning in the context of fixed Vapnik-Chervonenkis (VC) classes (Cannon et al., 2002). We also apply a different bounding technique that leads to substantially tighter upper bounds. For a review of VC theory see Devroye et al. (1996).

For the moment let $\mathcal{H}$ be an arbitrary, fixed collection of classifiers and let $\epsilon_0 > 0$. Cannon et al. propose the decision rule

$$\begin{aligned} \widehat{h}_n \quad &= \quad \arg\min_{h\in\mathcal{H}} \ \widehat{R}_1(h) \\ &\quad \text{s.}t. \quad \widehat{R}_0(h) \leq \alpha + \frac{1}{2}\epsilon_0. \end{aligned} \tag{3.1}$$

We call this procedure Neyman-Pearson empirical risk minimization (NP-ERM). Cannon et al. demonstrate that NP-ERM enjoys properties similar to standard ERM (Devroye et al., 1996; Vapnik and Chervonenkis, 1971) translated to the Neyman-Pearson context. We now recall their analysis.

To state the theoretical properties of NP-ERM, introduce the following notation.* Let $\epsilon_1 > 0$. Recall $\mathcal{H}_0 = \{h \in \mathcal{H} : R_0(h) \leq \alpha\}$ and $h^* = \arg\min\{R_1(h) : h \in \mathcal{H}_0\}$. Define

$$\begin{aligned}
\Theta_0 &= \{Z^n : R_0(\widehat{h}_n) > \alpha + \epsilon_0\} \\
\Theta_1 &= \{Z^n : R_1(\widehat{h}_n) > R_1(h^*) + \epsilon_1\} \\
\Omega_0 &= \{Z^n : \sup_{h \in \mathcal{H}_0} |R_0(h) - \widehat{R}_0(h)| > \epsilon_0/2\} \\
\Omega_1 &= \{Z^n : \sup_{h \in \mathcal{H}_0} |R_1(h) - \widehat{R}_1(h)| > \epsilon_1/2\}.
\end{aligned}$$

The key result of Cannon et al. (2002) is the following lemma.

Lemma 16 (Cannon, Howse, Hush and Scovel). *With Θ_j and Ω_j as defined above, we have*

$$\Theta_0 \subset \Omega_0,$$

$$\Theta_1 \subset \Omega_0 \cup \Omega_1$$

and in particular,

$$\Theta_0 \cup \Theta_1 \subset \Omega_0 \cup \Omega_1.$$

An immediate corollary is the following.

Proposition 4 (Cannon, Howse, Hush and Scovel). *Let $\epsilon_1, \epsilon_0 > 0$ and take $\widehat{h}_n$ as in (3.1). Then*

$$\begin{aligned}
&\mathbb{P}^n\left(\left(R_0(\widehat{h}_n) - \alpha > \epsilon_0\right) \text{ or } \left(R_1(\widehat{h}_n) - R_1(h^*) > \epsilon_1\right)\right) \\
&\quad \leq \mathbb{P}^n\left(\sup_{h \in \mathcal{H}} |R_1(h) - \widehat{R}_1(h)| > \epsilon_1/2\right) + \mathbb{P}^n\left(\sup_{h \in \mathcal{H}} |R_0(h) - \widehat{R}_0(h)| > \epsilon_0/2\right).
\end{aligned}$$

*We interchange the meanings of the subscripts 0 and 1 used by Cannon et al. (2002), prefering to associate class 0 with the null hypothesis.

This result is termed a "fundamental lemma" for its pivotal role in relating the performance of $\widehat{h}_n$ to bounds on the error deviance. Below this result is used to derive PAC bounds by applying results for convergence of empirical processes such as VC inequalities.

We make one observation that is not mentioned by Cannon et al. (2002). In both of the above results, the tolerance parameters ϵ_0 and ϵ_1 need not be constants, in the sense that they may depend on the sample or certain other parameters. This will be a key to our improved bounds and extension to structural risk minimization. In particular, we will choose ϵ_j to depend on n_j, a confidence parameter δ_j, and a measure of the capacity of $\mathcal{H}$ such as the cardinality (if $\mathcal{H}$ is finite) or VC dimension (if $\mathcal{H}$ is infinite).

3.2.1 NP Learning with VC Classes

Suppose $\mathcal{H}$ has VC dimension $V < \infty$. Cannon et al. consider two viewpoints for NP learning from $\mathcal{H}$. First they consider *retrospective sampling* where n_0 and n_1 are known before the sample is gathered. Applying the VC inequality as discussed in Devroye et al. (1996), together with Proposition 4, they obtain

Theorem 10 (Cannon, Howse, Hush and Scovel). *Let $\epsilon_1, \epsilon_0 > 0$ and take $\widehat{h}_n$ as in (3.1). Then*

$$\mathbb{P}^n \left(\left(R_0(\widehat{h}_n) - \alpha > \epsilon_0 \right) or \left(R_1(\widehat{h}_n) - R_1^* > \epsilon_1 \right) \right) \leq 8 n_0^V e^{-n_0 \epsilon_0^2/128} + 8 n_1^V e^{-n_1 \epsilon_1^2/128}.$$

The retrospective sampling plan is often not realistic. In *i.i.d sampling*, n_0 and n_1 are unknown a priori. Unfortunately, application of the VC inequality is not so straightforward in this setting because n_0 and n_1 are now random variables. To circumvent this problem, Cannon et al. arrive at the following result by arguing that with high probability n_0 and n_1 are concentrated near their expected values.

Theorem 11 (Cannon, Howse, Hush and Scovel). *Let $\epsilon_1, \epsilon_0 > 0$ and take $\widehat{h}_n$ as in (3.1). If $n \geq \frac{10\sqrt{5}}{\pi_j^2 \epsilon_j^2}, j = 0, 1$, then*

$$\mathbb{P}^n \left(\left(R_1(\widehat{h}_n) - R_1^* > \epsilon_1 \right) or \left(R_0(\widehat{h}_n) - \alpha > \epsilon_0 \right) \right) \leq 10(2n)^V (e^{-\frac{n\pi_0^2\epsilon_0^2}{640\sqrt{5}}} + e^{-\frac{n\pi_1^2\epsilon_1^2}{640\sqrt{5}}}).$$

Owing to the larger constants out front and in the exponents, their bound for i.i.d. sampling is substantially larger than for retrospective sampling. In addition, the bound does not hold for small n, and since the a priori class probabilities are unknown in the NP setting, it is not known for which n the bound does hold.

We propose an alternate PAC bound for NP learning from VC classes under i.i.d. sampling that improves upon the preceding result. In particular, our bound is as tight as the bound in Theorem 10 for retrospective sampling and it holds for all values of n. As mentioned previously, the key idea is to let the tolerances not be fixed. Thus let $\delta_0, \delta_1 > 0$ and define

$$\epsilon_j = \epsilon_j(n_j, \delta_j) = \sqrt{128 \frac{V \log n_j + \log(8/\delta_j)}{n_j}} \tag{3.2}$$

for $j = 0, 1$. Let $\widehat{h}_n$ be defined in the same way as before, but with the new definition of ϵ_0 (which now depends on n_0 and δ_0). We have the following.

Theorem 12. *For NP-ERM over a VC class $\mathcal{H}$ with tolerances given by (3.2),*

$$\mathbb{P}^n \left(\left(R_0(\widehat{h}_n) - \alpha > \epsilon_0(n_0, \delta_0) \right) \text{ or } \left(R_1(\widehat{h}_n) - R_1(h^*) > \epsilon_1(n_1, \delta_1) \right) \right) \leq \delta_0 + \delta_1.$$

Proof. By Proposition 4, it suffices to show for $j = 0, 1$

$$\mathbb{P}^n \left(\sup_{h \in \mathcal{H}} |R_j(h) - \widehat{R}_j(h)| > \frac{1}{2} \epsilon_j(n_j, \delta_j) \right) \leq \delta_j.$$

Without loss of generality take $j = 0$. Then

$$\begin{aligned}
&\mathbb{P}^n \left(\sup_{h \in \mathcal{H}} |R_0(h) - \widehat{R}_0(h)| > \frac{1}{2} \epsilon_0(n_0, \delta_0) \right) \\
&= \sum_{n_0=0}^{n} \mathbb{P}^n \left(\sup_{h \in \mathcal{H}} |R_0(h) - \widehat{R}_0(h)| > \frac{1}{2} \epsilon_0(n_0, \delta_0) \mid n_0 \right) \mathbb{P}(n_0) \\
&\leq \sum_{n_0=0}^{n} \delta_0 \mathbb{P}(n_0) \\
&= \delta_0,
\end{aligned}$$

where the inequality follows from the VC theorem of Devroye et al. (1996, chap. 12).

This completes the proof. □

The new bound is substantially tighter than that of Theorem 11. For the purpose of comparison, assume $V = 1$, $n_0 = n_1 = \frac{1}{2}n$, and $\pi_0 = \pi_1 = \frac{1}{2}$. How large should n be so that we are guaranteed that with at least .8 probability, both bounds hold with $\epsilon_0 = \epsilon_1 = 0.2$? For the new bound of Theorem 12 to hold we need

$$\sqrt{128\frac{\log(\frac{1}{2}n) + \log(8/(0.1))}{\frac{1}{2}n}} \leq 0.2$$

which implies $n \geq 97104$. In contrast, Theorem 11 requires

$$10ne^{-\frac{n(0.5)^2(0.2)^2}{640\sqrt{5}}} \leq 0.1$$

which implies $n \geq 2782609$. Note also that the bound of Cannon et al. for retrospective sampling requires as many samples as our bound for i.i.d sampling.

The main difference between the rule of Cannon et al. and the rule proposed here is that in their forumulation the tolerance $\frac{1}{2}\epsilon_0$ constraining the empirical false alarm probability is independent of the sample. In contrast, our tolerance is smaller for larger values of n_0. When more training data is available for class 0, a higher level of accuracy is required. We argue that this is a desirable property. Theoretically, we have shown that it leads to a substantially tighter bound. Intuitively, when n_0 is larger, $\widehat{R}_0$ should more accurately estimate R_0, and therefore a suitable decision rule should be available from among those rules approximating α to within the smaller tolerance.

3.2.2 NP Learning with Finite Classes

The VC inequality is so general that in most practical settings it is too loose owing to large constants. Fortunately, Proposition 4 allows for the application of many of the error deviance bounds that have appeared in the machine learning literature in recent years (Bousquet et al., 2004). One very simple example occurs when $\mathcal{H}$ is finite. In this case, bounds with substantially smaller constants may be derived even when $\mathcal{H}$ is obtained by quantizing all elements of some VC class to machine precision.

Let $\mathcal{H}$ be finite and define the NP-ERM estimator $\widehat{h}_n$ as before. Redefine the

tolerances ϵ_0 and ϵ_1 by

$$\epsilon_j = \epsilon_j(n_j, \delta_j) = \sqrt{2\frac{\log|\mathcal{H}| + \log(2/\delta_j)}{n_j}}. \tag{3.3}$$

We have the following analogue of Theorem 12.

Theorem 13. *For NP-ERM over a finite class $\mathcal{H}$ with tolerances given by (3.3),*

$$\mathbb{P}^n\left(\left(R_0(\widehat{h}_n) - \alpha > \epsilon_0(n_0, \delta_0)\right) \text{ or } \left(R_1(\widehat{h}_n) - R_1(h^*) > \epsilon_1(n_1, \delta_1)\right)\right) \leq \delta_0 + \delta_1.$$

The proof is identical to the proof of Theorem 12 except that the VC inequality is replaced by a simpler bound (with smaller constants) derived from Hoeffding's inequality and the union bound (see Devroye et al., 1996, thm. 8.3).

To illustrate the significance of the new bound, consider the scenario described above (just after Theorem 12), but assume the VC class $\mathcal{H}$ is quantized so that $|\mathcal{H}| = 2^{16}$. For the bound of Theorem 13 to hold, we need

$$\sqrt{2\frac{16\log 2 + \log(2/(0.1))}{\frac{1}{2}n}} \leq 0.2$$

which implies $n \geq 1409$, a significant improvement.

Another example of the improved bounds available for finite $\mathcal{H}$ is given in Section 3.5.2 where we apply NP-SRM to dyadic decision trees.

3.3 Neyman-Pearson and Structural Risk Minimization

One limitation of NP-ERM over fixed $\mathcal{H}$ is that most possibilities for the optimal rule g^* cannot be approximated arbitrarily well. Such rules will never be universally consistent. A solution to this problem is known as structural risk minimization (SRM) (Vapnik, 1982), whereby a classifier is selected from a family $\mathcal{H}^k, k = 1, 2, \ldots$, of increasingly rich classes of decision rules. In this section we present NP-SRM, a version of SRM adapted to the NP setting, in the two cases where the $\mathcal{H}^k$ are either all VC classes or all finite.

3.3.1 SRM over VC Classes

Let $\mathcal{H}^k, k = 1, 2, \ldots$, be given, with $\mathcal{H}^k$ having VC dimension V_k. Assume $V_1 < V_2 < \cdots$. Define the tolerances ϵ_0 and ϵ_1 by

$$\epsilon_j = \epsilon_j(n_j, \delta_j, k) = \sqrt{128 \frac{V_k \log n_j + k \log 2 + \log(8/\delta_j)}{n_j}}. \tag{3.4}$$

Remark 3. The new value for ϵ_j is equal to the value of ϵ_j in the previous section with δ_j replaced by $\delta_j 2^{-k}$. The choice of the scaling factor 2^{-k} stems from the fact $\sum_{k=1}^{\infty} 2^{-k} = 1$, which is used to show (by the union bound) that the VC inequalities hold for all k uniformly with probability at least $1 - \delta_j$ (see the proof of Theorem 14).

NP-SRM produces a decision rule $\widehat{h}_n$ according to the following two-step process. Let $K(n)$ be a nondecreasing integer valued function of n with $K(1) = 1$.

1. For each $k = 1, 2, \ldots, K(n)$, set

$$\begin{aligned} \widehat{h}_n^k &= \arg\min_{h \in \mathcal{H}^k} \widehat{R}_1(h) \\ &\text{s.}t. \quad \widehat{R}_0(h) \leq \alpha + \frac{1}{2}\epsilon_0(n_0, \delta_0, k). \end{aligned} \tag{3.5}$$

2. Set

$$\widehat{h}_n = \arg\min\{\widehat{R}_1(\widehat{h}_n^k) + \frac{1}{2}\epsilon_1(n_1, \delta_1, k) \mid k = 1, 2, \ldots K(n)\}.$$

The term $\frac{1}{2}\epsilon_1(n_1, \delta_1, k)$ may be viewed as a *penalty* that measures the complexity of class $\mathcal{H}^k$. In words, NP-SRM uses NP-ERM to select a candidate from each VC class, and then selects the best candidate by balancing empirical miss probability with classifier complexity.

Remark 4. If $\mathcal{H}^1 \subset \mathcal{H}^2 \subset \cdots$ and $\mathcal{H}(n) = \cup_{k=1}^{K(n)} \mathcal{H}^k$, then NP-SRM may equivalently be viewed as the solution to a single-step optimization problem:

$$\begin{aligned} \widehat{h}_n &= \arg\min_{h \in \mathcal{H}(n)} \widehat{R}_1(h) + \frac{1}{2}\epsilon_1(n_1, \delta_1, k(h)) \\ &\text{s.}t. \quad \widehat{R}_0(h) \leq \alpha + \frac{1}{2}\epsilon_0(n_0, \delta_0, k(h)) \end{aligned}$$

where $k(h)$ is the smallest k such that $h \in \mathcal{H}^k$.

We have the following PAC bound for NP-SRM. For a similar result in the context of Bayes error learning see Lugosi and Zeger (1996).

Theorem 14. *With probability at least* $1-(\delta_0+\delta_1)$ *over the training sample* Z^n*, both*

$$R_0(\widehat{h}_n) - \alpha \leq \epsilon_0(n_0, \delta_0, K(n)) \tag{3.6}$$

and

$$R_1(\widehat{h}_n) - R_1^* \leq \inf_{1\leq k\leq K(n)} \left(\epsilon_1(n_1, \delta_1, k) + \inf_{h\in\mathcal{H}_0^k} R_1(h) - R_1^* \right) \tag{3.7}$$

hold.

To interpret the second inequality, observe that for any k we may write

$$R_1(\widehat{h}_n) - R_1^* = \left(R_1(\widehat{h}_n) - \inf_{h\in\mathcal{H}_0^k} R_1(h) \right) + \left(\inf_{h\in\mathcal{H}_0^k} R_1(h) - R_1^* \right).$$

The two terms on the right are refered to as *estimation error* and *approximation*[†] error, respectively. If k is such that $\widehat{h}_n = \widehat{h}_n^k$, then Theorem 12 implies that the estimation error is bounded above by $\epsilon_1(n_1, \delta_1, k)$. Thus (3.7) says that $\widehat{h}_n$ performs as well as an *oracle* that clairvoyantly selects k to minimize this upper bound. As we will soon see, this result leads to strong universal consistency of $\widehat{h}_n$ under mild conditions. For further discussion of oracle inequalities for penalized ERM see Bartlett et al. (2002); Bousquet et al. (2004).

The inequality in (3.6) implies that the "excess false alarm probability" decays no slower than $O(\sqrt{V_{K(n)} \log n/n})$. Unfortunatley no oracle inequality is available for this error, which may limit the ability of NP-SRM to adapt to different conditions on the underlying distribution.

3.3.2 SRM over Finite Classes

The developments of the preceding section have counterparts in the context of SRM over a family of finite classes. The rule for NP-SRM is defined in the same way, but

[†]Note that, in constrast to the standard approximation error for Bayes error learning, here the optimal classifier g^* must be approximated by classifiers satisfying the constraint on the false alarm probability.

now with penalties

$$\epsilon_j = \epsilon_j(n_j, \delta_j, k) = \sqrt{2\frac{\log|\mathcal{H}^k| + k\log 2 + \log(2/\delta_j)}{n_j}}. \tag{3.8}$$

Theorem 14 holds in this setting as well. The proof is an easy modification of the proof of that theorem, substituting Theorem 13 for Theorem 12, and is omitted.

3.4 Consistency

The inequalities above for NP-SRM over VC and finite classes may be used to prove strong universal consistency of NP-SRM provided the sets $\mathcal{H}^k$ are sufficiently rich as $k \to \infty$ and provided $\delta_j = \delta_j(n)$ and $K(n)$ are callibrated appropriately. Note that the conditions on δ hold if $\delta_j(n) \asymp n^{-\beta}$ for some $\beta > 1$.

Theorem 15. *Let $\widehat{h}_n$ be the classifier given by (3.5), with $\epsilon_0(n_0, \delta_0, k)$ defined by (3.4) for NP-SRM over VC classes, or (3.8) for NP-SRM over finite classes. Specify $\delta_j(n)$ and $K(n)$ such that*

1. *$\delta_j(n)$ satisfies $\log(1/\delta_j(n)) = O(n)$*
2. *$\delta_j(n)$ is summable, i.e., for each $j = 0, 1$*

$$\sum_{n=1}^{\infty} \delta_j(n) < \infty$$

3. *$K(n) \to \infty$ as $n \to \infty$*
4. *If $\mathcal{H}^k$ are VC classes, then $V_{K(n)} = o(n/\log n)$. If $\mathcal{H}^k$ are finite, then $\log|\mathcal{H}^{K(n)}| = o(n)$.*

Assume that for any distribution of Z there exists a sequence $h_k \in \mathcal{H}_0^k$ such that

$$\liminf_{k\to\infty} R_1(h_k) = R_1^*.$$

Then $\widehat{h}_n$ is strongly universally consistent, i.e.,

$$\lim_{n\to\infty} R_0(\widehat{h}_n) \leq \alpha \quad \textit{with probability 1}$$

and

$$\lim_{n\to\infty} R_1(\widehat{h}_n) \leq R_1^* \quad \textit{with probability 1}$$

for all distributions.

To illustrate the theorem, suppose $\mathcal{X} = [0,1]^d$ and $\mathcal{H}^k$ is the family of regular histogram classifiers based on cells of bin-width $1/k$. Then $|\mathcal{H}^k| = 2^{k^d}$ and NP-SRM is consistent provided $k \to \infty$ and $k^d = o(n)$, in analogy to the requirement for strong universal consistency of the regular histogram rule under probability of error (Devroye et al., 1996, chap. 9).

3.5 Rates of Convergence

In this section assume $\mathcal{X} = [0,1]^d$. Our goal is to derive rates of convergence to zero for the expected[‡] *excess false alarm probability*

$$\mathbb{E}^n\{R_0(\widehat{h}_n)\} - \alpha$$

and expected *excess miss probability*

$$\mathbb{E}^n\{R_1(\widehat{h}_n)\} - R_1^*$$

where $\widehat{h}_n$ is the decision rule produced by NP-SRM over a certain family of classes. Moreover, we are interested in rates that hold independent of α.

Several recent studies (Blanchard et al., 2003; Scovel and Steinwart, 2004; Tsybakov, 2004; Tsybakov and van de Geer, 2004, see also chap. 2 of this thesis) have derived rates of convergence for the expected *excess probability of error*, $\mathbb{E}^n\{R(\widehat{\phi}_n)\} - R(\phi^*)$, where ϕ^* is the (optimal) Bayes classifier. Observe that

$$R(\phi) - R(\phi^*) = \pi_0(R_0(\phi) - R_0(\phi^*)) + \pi_1(R_1(\phi) - R_1(\phi^*)).$$

Hence, rates of convergence for Neyman-Pearson learning with $\alpha = R_0(\phi^*)$ imply rates of convergence for standard Bayes error learning. We summarize this fact as follows.

[‡]It is also possible to prove rates involving probability inequalities. While more general, this approach is slightly more cumbersome so we prefer to work with expected values.

Proposition 5. *Fix a distribution of the data Z. Let $\widehat{h}_n$ be a decision rule, and let $r_j(n), j = 0, 1$, be non-negative functions tending toward zero. If for each $\alpha \in [0, 1]$*

$$\mathbb{E}^n\{R_0(\widehat{h}_n)\} - \alpha \preccurlyeq r_0(n)$$

and

$$\mathbb{E}^n\{R_1(\widehat{h}_n)\} - R_1^* \preccurlyeq r_1(n),$$

then

$$\mathbb{E}^n\{R(\widehat{h}_n)\} - R(\phi^*) \preccurlyeq \max\{r_0(n), r_1(n)\}.$$

Devroye (1982) has shown that for any decision rule $\widehat{\phi}_n$ there exists a distribution of Z such that $R(\widehat{\phi}_n) - R(\phi^*)$ decays at an arbitrarily slow rate. In other words, to prove rates of convergence one must impose some kind of assumption on the distribution. In light of Proposition 5, the same must be true of NP learning. Thus, let $\mathcal{D}$ be some class of distributions. Proposition 5 also informs us about lower bounds for learning from distributions in $\mathcal{D}$.

Proposition 6. *Assume that Bayes error learning satisfies the minimax lower bound*

$$\inf_{\widehat{\phi}} \sup_{\mathcal{D}} \left(\mathbb{E}^n\{R(\widehat{\phi})\} - R(\phi^*) \right) \succcurlyeq r(n).$$

If $r_0(n), r_1(n)$ are upper bounds on the rate of convergence for NP learning (that hold independent of α), then either $r_0(n) \succcurlyeq r(n)$ or $r_1(n) \succcurlyeq r(n)$.

In other words, minimax lower bounds for Bayes error learning translate to minimax lower bounds for NP learning.

3.5.1 The Box-Counting Class

It is our goal to illustrate an example of how one may derive rates of convergence using NP-SRM combined with an appropriate analysis of approximation error. Before introducing a class $\mathcal{D}$ we need some additional notation. Let m denote a positive integer, and define $\mathcal{P}_m$ to be the collection of m^d cells formed by the regular partition of $[0, 1]^d$ into hypercubes of sidelength $1/m$. Let $c_0, c_1 > 0$ be positive real numbers. Let $G^* = \{x \in [0, 1]^d : g^*(x) = 1\}$ be the optimal decision set, and let ∂G^* be the

topological boundary of G^*. Finally, let $N_m(\partial G^*)$ denote the number of cells in $\mathcal{P}_m$ that intersect ∂G^*.

We define the *box-counting* class to be the set $\mathcal{D}$ of all distributions satisfying

A0 : The marginal density $f_1(x)$ of X given $Y = 1$ is essentially bounded by c_0.

A1 : $N_m(\partial G^*) \leq c_1 m^{d-1}$ for all m.

The first assumption[§] is equivalent to requiring $\mathbb{P}_{X|Y=1}(A) \leq c_0 \lambda(A)$ for all measurable sets A, where λ denotes the Lebesgue measure on $[0,1]^d$. The second assumption essentially requires the optimal decision boundary ∂G^* to have Lipschitz smoothness. See Chapter 2 for further discussion. A theorem of Tsybakov (2004) implies that the minimax rate for Bayes error learning from this class is $n^{-1/d}$ when $d \geq 2$. By Proposition 6, we can hope for both errors to decay at this rate. Below we prove that such optimality is almost attained using dyadic decision trees and NP-SRM.

3.5.2 NP-SRM for Dyadic Decision Trees

In Chapter 2 we demonstrate that a certain family of decision trees, *dyadic decision trees* (DDTs), offer a computationally feasible classifier that also achieves optimal rates of convergence (with respect to Bayes error) under a wide range of conditions (Scott and Nowak, 2002, 2003, see also). DDTs are especially well suited for rate of convergence studies. Indeed, bounding the approximation error is handled by the restriction to dyadic splits, which allows us to take advantage of recent insights from multiresolution analysis and nonlinear approximations (Cohen et al., 2001; DeVore, 1998; Donoho, 1999). We now show that an analysis similar to that of Scott and Nowak (2002) applies to NP-SRM for DDTs, leading to similar results: rates of convergence and a computationally efficient algorithm.

A dyadic decision tree is a decision tree that divides the input space by means of axis-orthogonal dyadic splits. More precisely, a dyadic decision tree T is a binary tree (with a distinguished root node) specified by assigning an integer $s(v) \in \{1, \ldots, d\}$ to each internal node v of T (corresponding to the coordinate this is split at that

[§]When proving rates for Bayes error, it is often necessary to place a similar restriction on the *unconditional* density $f(x)$ of X. In our formulation of NP learning, it is only necessary to bound $f_1(x)$ because only the excess *miss* probability requires an analysis of approximation error.

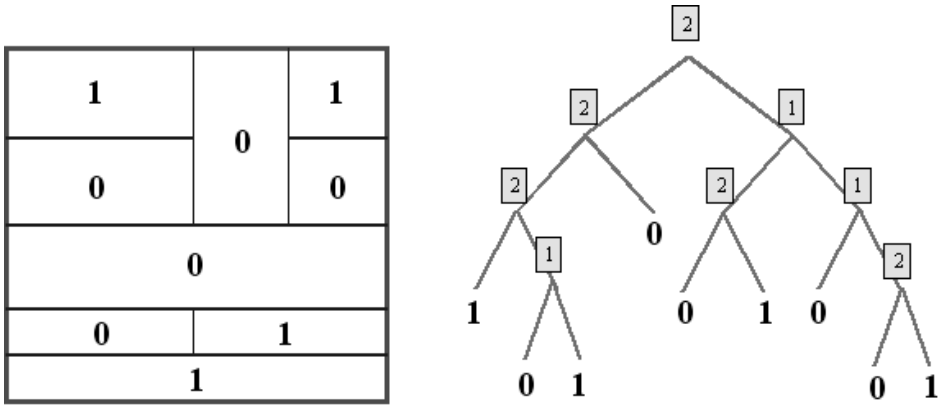


Figure 3.1 : A dyadic decision tree (right) with the associated recursive dyadic parition (left) assuming $d = 2$. Each internal node of the tree is labeled with an integer from 1 to d indicating the coordinate being split at that node. The leaf nodes are decorated with class labels.

node), and a binary label 0 or 1 to each terminal (or leaf) node of T. The nodes of DDTs correspond to hyperrectangles (cells) in $[0,1]^d$. 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 T is associated with a cell according to the following rules: (1) The root node is associated with $[0,1]^d$; (2) If v is an internal node associated with the cell A, then the children of v are associated with $A^{s(v),1}$ and $A^{s(v),2}$. See Figure 3.1.

Let $L = L(n)$ be a nonnegative integer and define $\mathcal{T}_L$ to be the collection of all DDTs such that no leaf cell has a sidelength smaller than 2^{-L}. In other words, when traversing a path from the root to a leaf cell no coordinate is split more than L times. Finally, define $\mathcal{T}_L^k$ to be the collection of all trees in $\mathcal{T}_L$ having k leaf nodes.

We study NP-SRM over the family $\mathcal{H}^k = \mathcal{T}_L^k$. Since $\mathcal{H}^k$ is both finite and a VC class, we may define penalties via (3.8) or (3.4). The VC dimension of $\mathcal{H}^k$ is simply k, while $|\mathcal{H}^k|$ may be bounded as follows: The number of binary trees with $k+1$ leaves is given by the Catalan number[¶] $C_k = (k+1)^{-1}\binom{2k}{k}$. The leaves of such trees may be labeled in 2^{k+1} ways, while the internal splits $s(v)$ may be assigned in d^k

[¶]See http://mathworld.wolfram.com/CatalanNumber.html.

ways. Asymptotically, it is known that $C_k \sim 4^k/(\sqrt{\pi}k^{3/2})$. Thus, for n sufficiently large, $\log|\mathcal{H}^k| \leq k(\log 8 + \log d)$. If $\epsilon_j(n_j, \delta_j, k)$ is defined by (3.8) for finite classes, it behaves like $\sqrt{2k(\log 8 + \log d)/n_j}$, while the penalty defined by (3.4) for VC classes behaves like $\sqrt{128k\log(n_j)/n_j}$. In conclusion, viewing $\mathcal{H}^k$ as a finite class rather than a VC class leads to bounds with smaller constants and without an additional log term.

By applying NP-SRM to DDTs[‖], with parameters $L(n), K(n)$, and $\delta_j(n)$ chosen appropriately, we obtain the following result. Note that the condition on $\delta_j(n)$ holds whenever $\delta_j(n) \asymp n^{-\beta}$, $\beta > 1/2$. The proof is found in Section 3.7.3.

Theorem 16. *Let $\widehat{h}_n$ be the decision rule given by (3.5), with $\epsilon_0(n_0, \delta_0, k)$ defined by (3.8). Specify $L(n), K(n)$, and $\delta_j(n)$ such that*

1. $2^{L(n)} \succcurlyeq n^{1/(d+1)}$
2. $K(n) \asymp n^{(d-1)/(d+1)}$
3. $\delta_j(n) = O(1/\sqrt{n})$ *and* $\log(1/\delta_j(n)) = O(\log n)$

If $d \geq 2$ and the distribution of Z belongs to the box-counting class, then

$$\mathbb{E}^n\{R_0(\widehat{h}_n)\} - \alpha \preccurlyeq n^{-1/(d+1)}$$

and

$$\mathbb{E}^n\{R_1(\widehat{h}_n)\} - R_1^* \preccurlyeq n^{-1/(d+1)}.$$

In light of the earlier discussion of minimax lower bounds for the box-counting class, either the lower bound of $n^{-1/d}$ is not tight, or the rates obtained for DDTs in the previous section are suboptimal. Assuming the lower bound is tight, does suboptimality stem from the use of DDTs or is the limitation a function of the NP-SRM learning procedure? Given previous experience, our guess is the latter. In Scott and Nowak (2002), we showed that DDTs and standard SRM yield suboptimal rates like those in Theorem 16 for Bayes error learning. Subsequently, we were able to obtain the optimal rate with DDTs using a spatially adaptive penalty (which depends

[‖]Since $L(n)$ changes with n, the classes $\mathcal{H}^k$ are not independent of n as they are in the development of Section 3.3. However, a quick inspection of the proofs of Theorems 14 and 15 reveals that those theorems also hold in this slightly more general setting.

on more than the size of the tree) and a penalized empirical risk procedure (Scott and Nowak, 2003, see chap. 2). Our conjecture is that a similar phenomenon occurs here. In other words, with a spatially adaptive penalty we should be able to achieve the optimal rate. However, since this would require a significant amount of additional space and detract from the focus of the paper, we do not pursue the matter further here.

3.5.3 Implementing Dyadic Decision Trees

The importance of DDTs stems not only from their theoretical properties but also from the fact that for DDTs, NP-SRM may be implemented exactly in polynomial time. In this subsection we provide an explicit algorithm to accomplish this task. The algorithm is inspired the work of Blanchard et al. (2004) who extend an algorithm of Donoho (1997) to perform penalized empirical (Bayes) risk minimization for DDTs.

Our theorems on consistency and rates of convergence tell us how to specify the asymptotic behavior of K and L, but in a practical setting these guidelines are less helpful. Assume L is selected by the user to be some maximum resolution of interest and take $K = 2^{dL}$, the largest possible meaningful value. We replace the symbol h for a generic decision rule by the notation T for trees. We seek an algorithm implementing

$$\begin{aligned} \widehat{T} &= \arg\min_{T \in \mathcal{T}_L} \widehat{R}_1(T) + \frac{1}{2}\epsilon_1(n_1, \delta_1, |T|) \\ &\quad \text{s.}t. \quad \widehat{R}_0(T) \leq \alpha + \frac{1}{2}\epsilon_0(n_0, \delta_0, |T|). \end{aligned}$$

Let $\mathcal{A}_L$ be the set of all cells corresponding to nodes of trees in $\mathcal{T}_L$. In other words, every $A \in \mathcal{A}_L$ is obtained by applying no more than L dyadic splits to each coordinate. For $A \in \mathcal{A}_L$, let $\mathcal{T}_L(A)$ denote the set of all subtrees of trees $T \in \mathcal{T}_L$ rooted at A. Let $\mathcal{I}_A$ be the set of all (k, ℓ) such that $|T_A| = k$ and $\widehat{R}_0(T_A) = \ell/n_0$ for some $T_A \in \mathcal{T}_L(A)$. For all $(k, \ell) \in \mathcal{I}_A$ define

$$T_A^{k,\ell} = \arg\min\{\widehat{R}_1(T_A) \mid T_A \in \mathcal{T}_L(A), |T_A| = k, \widehat{R}_0(T_A) = \ell/n_0\}.$$

When $A = [0,1]^d$ write $T^{k,\ell}$ for $T_A^{k,\ell}$ and $\mathcal{I}$ for $\mathcal{I}_A$. We refer to these trees $T^{k,\ell}$ as minimum empirical risk trees, or MERTs for short. They may be computed in a

Input: Minimum empirical risk trees $T^{k,\ell}$ and $\mathcal{I}$
Initialize: $C^{\min} = \infty$
For $(k,\ell) \in \mathcal{I}$
  If $\ell/n_0 \leq \alpha + \frac{1}{2}\epsilon_0(n_0, \delta_0, k)$
    $C^{\text{temp}} = \widehat{R}_1(T^{k,\ell}) + \frac{1}{2}\epsilon_1(n_1, \delta_1, k)\}$
    If $C^{\text{temp}} < C^{\min}$
      $C^{\min} \leftarrow C^{\text{temp}}$
      $\widehat{T} \leftarrow T^{k,\ell}$
    End
  End
End
Output: $\widehat{T}$

Figure 3.2 : Algorithm for NP-SRM with dyadic decision trees.

recursive fashion (described below) and used to determine $\widehat{T}$:

$$\widehat{T} = \arg\min\{\widehat{R}_1(T) + \frac{1}{2}\epsilon(n_1, \delta_1, |T|) \mid T = T^{k,\ell}, (k,\ell) \in \mathcal{I}\}.$$

The algorithm is stated formally in Figure 3.2.

The MERTs may be computed as follows. The idea is, for each cell $A \in \mathcal{A}_L$, to compute $T_A^{k,\ell}$ recursively in terms of $T_{A^{s,1}}^{k,\ell}$ and $T_{A^{s,2}}^{k,\ell}$, $s = 1, \ldots, d$, starting from the bottom of the tree and working up. The procedure for computing $T_A^{k,\ell}$ is as follows. When A is a cell at maximum depth $J = dL$, $\mathcal{I}_A = \{(1,0), (1, n_{0A})\}$ where $n_{jA} := |\{(X_i, Y_i) \mid X_i \in A, Y_i = j\}|$. Furthermore, then $T_A^{1,0} = \{A\}$ (labeled with class 0) and $T_A^{1,n_{0A}} = \{A\}$ (labeled with class 1).

Some additional notation is necessary to state the recursion: Denote by $\text{MERGE}(A, T_{A^{s,1}}, T_{A^{s,2}})$ the element of $\mathcal{T}_L(A)$ having $T_{A^{s,1}}$ and $T_{A^{s,2}}$ as its left and right branches. Now observe that for any cell at depth $j < J$,

$$\begin{aligned} T_A^{k,\ell} \;=\; & \arg\min\{\widehat{R}_1(T_A) \mid T_A = \text{MERGE}(A, T_{A^{s,1}}^{k',\ell'}, T_{A^{s,1}}^{k'',\ell''}), \\ & \qquad k' + k'' = k, \ell' + \ell'' = \ell, s = 1, \ldots, d\}. \end{aligned}$$

This follows by additivity of the empirical miss probability $\widehat{R}_1$. Note that this recursive relationship leads to a recursive algorithm for computing $T^{k,\ell}$.

At first glance, the algorithm appears to involve visiting all $A \in \mathcal{A}_L$, a potentially huge number of cells. However, given a fixed training sample, most of those cells will be empty. If A is empty, then $T_A^{k,\ell}$ is the degenerate tree consisting only of A. Thus it is only necessary to perform the recursive update at nonempty cells. This observation was made by Blanchard et al. (2004) to derive an algorithm for penalized empirical risk minimization over $\mathcal{T}_L$ for DDTs (using an additive penalty). They employ a *dictionary-based* approach which uses a dictionary $\mathcal{R}$ to keep track of the cells that need to be considered. Let $\mathcal{R}_j$, $j = 0, \ldots, J$ denote the cells in $\mathcal{R}$ at depth j. Our algorithm is inspired by their formulation, and is summarized in Figure 3.3.

Proposition 7. *The algorithm in Figure 3.3 requires $O(n^2 n_0 d^2 L^{d+1} \log(nL^d))$ operations.*

Proof. The proof is a minor variation on an argument given by Blanchard et al. (2004). For each training point X_i there are exactly $(L+1)^d$ cells in $\mathcal{A}_L$ containing the point (see Blanchard et al., 2004). Thus the total number of dictionary elements is $O(nL^d)$. For each cell $A' \in \mathcal{R}_j$ there are at most d parents $A \in \mathcal{R}_{j-1}$ to consider. For each such A a loop over $(k, \ell) \in \mathcal{I}_A$ is required. The size of $\mathcal{I}_A$ is $O(n_0 n(J-j))$. This follows because there are $O(n_0)$ possibilities for ℓ and $O(n(J-j))$ for k. To see this last assertion note that each element of $\mathcal{R}_J$ has $O(J-j)$ ancestors up to depth j. Using $J - j \leq J = Ld$ and combining the above observations it follows that each $A' \in \mathcal{R}_j$ requires $O(nd^2L)$ operations. Assuming that dictionary operations (searches and inserts) can be implemented in $O(\log |\mathcal{R}|)$ operations the result follows. □

Unfortunately the computational complexity has an exponential dependence on d. Schäfer et al. (2004) report that computational and memory constraints limit the algorithm to problems for which $d < 15$. However, if one desires a computationally efficient algorithm that achieves the rates in Theorem 16 for all d, there is an alternative. As shown in the proof of Theorem 16, it suffices to consider *cyclic* DDTs (defined in the proof). For NP-SRM with cyclic DDTs the algorithm of Figure 3.3 can be can be simplified so that it requires $O(n_0 n^2 d^2 L^2)$ operations. We opted to present the more general algorithm because it should perform much better in practice.

Input: Training sample Z^n
Initialize: Let $\mathcal{R}_J$, $J = dL$, denote the set of all nonempty dyadic hypercubes of sidelength 2^{-L}.
For all $A \in \mathcal{R}_J$ set $\mathcal{I}_A = \{(1,0),(1,n_{0A})\}$
For $j = J$ downto 1
  Initialize $\mathcal{R}_{j-1} = \emptyset$.
  For all $A' \in \mathcal{R}_j$
    For $s = 1, \ldots, d$
      Let A'' = sibling of A' along coordinate s
      Let A = parent of A' and A''
      If $A \notin \mathcal{R}_{j-1}$
        Add A to $\mathcal{R}_{j-1}$
        Initialize $\mathcal{I}_A = \{(1,0),(1,n_{0A})\}$
        Set $T_A^{1,0} = \{A\}$ (label = 0)
        Set $T_A^{1,n_{0A}} = \{A\}$ (label = 1)
      End
      For $(k',\ell') \in \mathcal{I}_{A'}$ and $(k'',\ell'') \in \mathcal{I}_{A''}$
        Set $k = k' + k''$ and $\ell = \ell' + \ell''$
        If $(k,\ell) \notin \mathcal{I}_A$
          Add (k,ℓ) to $\mathcal{I}_A$
          $T_A^{k,\ell} \leftarrow \text{MERGE}(A, T_{A'}^{k',\ell'}, T_{A''}^{k'',\ell''})$
        Elseif $\widehat{R}_1(T_{A'}^{k',\ell'}) + \widehat{R}_1(T_{A''}^{k'',\ell''}) < \widehat{R}_1(T_A^{k,\ell})$
          $T_A^{k,\ell} \leftarrow \text{MERGE}(A, T_{A'}^{k',\ell'}, T_{A''}^{k'',\ell''})$
        End
      End
    End
  End
End
Output: $T^{k,\ell}$ and $\mathcal{I}$

Figure 3.3 : Algorithm for computing minimum empirical risk trees for dyadic decision trees.

3.6 Conclusion

We have made an effort to extend standard results for learning classifiers from training data to the Neyman-Pearnson setting. Familiar concepts such as empirical and structural risk minimization have counterparts with analogous performance guarantees. In particular, NP-SRM, under mild assumptions, allows one to deduce strong univeral consistency and rates of convergence. Furthermore, dyadic decision trees allow for not only rates of convergence but also efficient and exact implementation of NP-SRM.

This work should be viewed as an initial step in translating the ever growing field of supervised learning for classification into the NP setting. An important next step is to evaluate the potential impact of NP learning in practical settings where different class-errors are valued differently. Toward this end it will be necessary to translate the theoretical framework established here into practical learning paradigms beyond decision trees, such as boosting and support vector machines (SVMs). In boosting, for example, it is conceivable that the procedure for "reweighting" the training data could be controlled to constrain the false alarm error. With SVMs or other margin-based classifiers, one could imagine a margin on each side of the decision boundary, with the class 0 margin constrained in some manner to control the false alarm error. If the results of this study are any indication, the performance of such NP algorithms should resemble that of their more familiar counterparts.

3.7 Proofs

3.7.1 Proof of Theorem 14

Define the sets

$$\begin{aligned}
\Theta_0 &= \{Z^n : R_0(\widehat{h}_n) - \alpha > \epsilon_0(n_0, \delta_0, K(n))\} \\
\Theta_1 &= \{Z^n : R_1(\widehat{h}_n) - R_1^* > \inf_{1 \le k \le K(n)} \left(\epsilon_1(n_1, \delta_1, k) + \inf_{h \in \mathcal{H}_0^k} R_1(h) - R_1^* \right)\} \\
\Omega_0^k &= \{Z^n : \sup_{h \in \mathcal{H}_0^k} |R_0(h) - \widehat{R}_0(h)| > \frac{1}{2}\epsilon_0(n_0, \delta_0, k)\} \\
\Omega_1^k &= \{Z^n : \sup_{h \in \mathcal{H}_0^k} |R_1(h) - \widehat{R}_1(h)| > \frac{1}{2}\epsilon_1(n_1, \delta_1, k)\} \\
\Theta_0^k &= \{Z^n : R_0(\widehat{h}_n^k) - \alpha > \epsilon_0(n_0, \delta_0, k)\} \\
\Theta_1^k &= \{Z^n : R_1(\widehat{h}_n^k) - \inf_{h \in \mathcal{H}_0^k} R_1(h) > \epsilon_1(n_1, \delta_1, k)\}
\end{aligned}$$

Our goal is to show

$$\mathbb{P}^n(\Theta_0 \cup \Theta_1) \le \delta_0 + \delta_1.$$

Lemma 17.

$$\Theta_0 \cup \Theta_1 \subset \cup_{k=1}^{\infty}(\Omega_0^k \cup \Omega_1^k)$$

Proof. We show the contrapositive, $\cap_{k=1}^{\infty}(\overline{\Omega_0^k} \cap \overline{\Omega_1^k}) \subset \overline{\Theta_0} \cap \overline{\Theta_1}$. So suppose $Z^n \in \cap_{k=1}^{\infty}(\overline{\Omega_0^k} \cap \overline{\Omega_1^k})$. By Lemma 16, $Z^n \in \cap_{k=1}^{\infty}(\overline{\Theta_0^k} \cap \overline{\Theta_1^k})$. In particular, $Z^n \in \cap_{k=1}^{\infty}\overline{\Theta_0^k}$. Since $\widehat{h}_n = \widehat{h}_n^{\hat{k}}$ for some $\hat{k} \le K(n)$ it follows that $Z^n \in \overline{\Theta_0}$.

To show $Z^n \in \overline{\Theta_1}$, first note that

$$R_1(\widehat{h}_n) \le \widehat{R}_1(\widehat{h}_n) + \frac{1}{2}\epsilon_1(n_1, \delta_1, \hat{k})$$

since $Z^n \in \cap_{k=1}^{\infty} \overline{\Omega_1^k}$. By the definition of NP-SRM,

$$\begin{aligned}
\widehat{R}_1(\widehat{h}_n) + \frac{1}{2}\epsilon_1(n_1, \delta_1, \hat{k}) &= \min_{1 \le k \le K(n)} \left(\widehat{R}_1(\widehat{h}_n^k) + \frac{1}{2}\epsilon_1(n_1, \delta_1, k) \right) \\
&= \min_{1 \le k \le K(n)} \left(\inf_{h \in \widehat{\mathcal{H}}_0^k} \widehat{R}_1(h) + \frac{1}{2}\epsilon_1(n_1, \delta_1, k) \right) \\
&\le \min_{1 \le k \le K(n)} \left(\inf_{h \in \widehat{\mathcal{H}}_0^k} R_1(h) + \epsilon_1(n_1, \delta_1, k) \right)
\end{aligned}$$

where $\widehat{\mathcal{H}}_0^k = \{h \in \mathcal{H}^k : \widehat{R}_0(h) \le \alpha + \frac{1}{2}\epsilon_0(n_0, \delta_0, k)\}$, and in the last step we use $Z^n \in \cap_{k=1}^{\infty} \overline{\Omega_1^k}$ again. Since $Z^n \in \cap_{k=1}^{\infty} \overline{\Omega_0^k}$, it follows that $\mathcal{H}_0^k \subset \widehat{\mathcal{H}}_0^k$ from which we conclude

$$R_1(\widehat{h}_n) \le \min_{1 \le k \le K(n)} \left(\epsilon_1(n_1, \delta_1, k) + \inf_{h \in \mathcal{H}_0^k} R_1(h) \right).$$

The lemma now follows by subtracting R_1^* from both sides. □

The theorem is proved by observing

$$\begin{aligned}
\mathbb{P}(\Theta_0 \cup \Theta_1) &\le \sum_{k=1}^{\infty} \mathbb{P}(\Omega_0^k) + \mathbb{P}(\Omega_1^k) \\
&\le \sum_{k=1}^{\infty} \delta_0 2^{-k} + \delta_1 2^{-k} \\
&= \delta_0 + \delta_1
\end{aligned}$$

where the second inequality comes from Remark 3 in Section 3.3 and a repetition of the argument in the proof of Theorem 12.

3.7.2 Proof of Theorem 15

We prove the theorem in the case of VC classes, the case of finite classes being entirely analogous. First consider the convergence of $R_0(\widehat{h}_n)$ to α. By the Borel-Cantelli lemma (Devroye et al., 1996, sect. A.6), it suffices to show that for each

$\epsilon > 0$, $\sum_{n=1}^{\infty} \mathbb{P}^n(R_0(\widehat{h}_n) - \alpha > \epsilon) < \infty$. So let $\epsilon > 0$. Define the events

$$\begin{aligned}
\Phi_0^n &= \{Z^n : R_0(\widehat{h}_n) - \alpha > \epsilon\} \\
\Psi_0^n &= \{Z^n : n_0 \leq \frac{1}{2}\pi_0 n\} \\
\Theta_0^n &= \{Z^n : R_0(\widehat{h}_n) - \alpha > \epsilon_0(n_0, \delta_0(n), K(n))\}.
\end{aligned}$$

Since $\Phi_0^n = (\Phi_0^n \cap \overline{\Psi_0^n}) \cup (\Phi_0^n \cap \Psi_0^n) \subset (\Phi_0^n \cap \overline{\Psi_0^n}) \cup \Psi_0^n$, we have

$$\sum_{n=1}^{\infty} \mathbb{P}^n(\Phi_0^n) \leq \sum_{n=1}^{\infty} \mathbb{P}^n(\Phi_0^n \cap \overline{\Psi_0^n}) + \sum_{n=1}^{\infty} \mathbb{P}^n(\Psi_0^n). \tag{3.9}$$

To bound the second term we use the following lemma.

Lemma 18.

$$\mathbb{P}^n(\Psi_0^n) \leq e^{-n\pi_0/8}$$

Proof. The relative Chernoff bound (Hagerup and Rüb, 1990) states that if $U \sim$ Binomial(n, p), then for all $\gamma > 0$, $\mathbb{P}(U/n \leq (1-\gamma)p) \leq e^{-np\gamma^2/2}$. Since $n_0 \sim$ Binomial(n, π_0), the lemma follows by applying the relative Chernoff bound with $\gamma = \frac{1}{2}$. □

It now follows that

$$\sum_{n=1}^{\infty} \mathbb{P}^n(\Psi_0^n) \leq \sum_{n=1}^{\infty} e^{-n\pi_0/8} < \infty.$$

To bound the first term in (3.9) we use the following lemma.

Lemma 19. *There exists N such that for all $n > N$, $\Phi_0^n \cap \overline{\Psi_0^n} \subset \Theta_0^n$.*

Proof. Define

$$\epsilon'(n) = \sqrt{128 \frac{V_{K(n)} \log n + K(n) \log 2 + \log(8/\delta_0(n))}{\frac{1}{2}\pi_0 n}}.$$

Since $V_{K(n)} = o(n/\log n)$ and $\log(1/\delta_0(n)) = o(n)$, we may choose N such that $n > N$ implies $\epsilon'(n) \leq \epsilon$. Suppose $n > N$ and consider $Z^n \in \Phi_0^n \cap \overline{\Psi_0^n}$. Since $Z^n \in \overline{\Psi_0^n}$ we have $\epsilon_0(n_0, \delta_0(n), K(n)) \leq \epsilon'(n) \leq \epsilon$, and since $Z^n \in \Phi_0^n$ we conclude $Z^n \in \Theta_0^n$. □

It now follows that, for the integer N provided by the lemma,

$$\begin{aligned}\sum_{n=1}^{\infty}\mathbb{P}^n(\Phi_0^n\cap\overline{\Psi_0^n}) &\leq N+\sum_{n>N}\mathbb{P}^n(\Phi_0^n\cap\overline{\Psi_0^n})\\ &\leq N+\sum_{n>N}\mathbb{P}^n(\Theta_0^n)\\ &\leq N+\sum_{n>N}(\delta_0(n)+\delta_1(n))<\infty,\end{aligned}$$

where in the last line we use Theorem 14.

Now consider the convergence of $R_1(\widehat{h}_n)$ to R_1^*. As before, it suffices to show that for each $\epsilon>0$, $\sum_{n=1}^{\infty}\mathbb{P}^n(R_1(\widehat{h}_n)-R_1^*>\epsilon)<\infty$. Let $\epsilon>0$ and define the sets

$$\begin{aligned}\Phi_1^n &= \{Z^n : R_1(\widehat{h}_n)-R_1^*>\epsilon\}\\ \Psi_1^n &= \{Z^n : n_1\leq\frac{1}{2}\pi_1 n\}\\ \Theta_1^n &= \{Z^n : R_1(\widehat{h}_n)-R_1^*>\inf_{1\leq k\leq K(n)}\left(\epsilon_1(n_1,\delta_1(n),k)+\inf_{h\in\mathcal{H}_0^k}R_1(h)-R_1^*\right)\}\end{aligned}$$

Arguing as before, it suffices to show

$$\sum_{n=1}^{\infty}\mathbb{P}^n(\Psi_1^n)<\infty$$

and

$$\sum_{n=1}^{\infty}\mathbb{P}^n(\Phi_1^n\cap\overline{\Psi_1^n})<\infty.$$

The first expression is bounded using an analogue of Lemma 18. To bound the second expression we employ the following analogue of Lemma 19.

Lemma 20. *There exists N such that $n>N$ implies $\Phi_1^n\cap\overline{\Psi_1^n}\subset\Theta_1^n$.*

Proof. Define

$$\epsilon'(n,k)=\sqrt{128\frac{V_k\log n+k\log 2+\log(8/\delta_1(n))}{\frac{1}{2}\pi_1 n}}.$$

Since $V_{K(n)}=o(n/\log n)$ and $\log(1/\delta_1(n))=o(n)$, we may choose N such that there exists $k^*\leq K(N)$ satisfying (i) $\epsilon'(n,k^*)\leq\epsilon/2$ and (ii) $\inf_{h\in\mathcal{H}_0^{k^*}}R_1(h)-R_1^*<\epsilon/2$.

Suppose $n > N$ and $Z^n \in \Phi_1^n \cap \overline{\Psi_0^n}$. Since $Z^n \in \overline{\Psi_1^n}$ we have

$$\begin{aligned}
&\inf_{1\le k\le K(n)} \left(\epsilon_1(n_1,\delta_1(n),k) + \inf_{h\in\mathcal{H}_0^k} R_1(h) - R_1^*\right) \\
&\quad\le \epsilon_1(n_1,\delta_1(n),k^*) + \inf_{h\in\mathcal{H}_0^{k^*}} R_1(h) - R_1^* \\
&\quad\le \epsilon'(n,k^*) + \inf_{h\in\mathcal{H}_0^{k^*}} R_1(h) - R_1^* \\
&\quad\le \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon.
\end{aligned}$$

Since $Z^n \in \Phi_1^n$ we conclude $Z^n \in \Theta_1^n$. □

The remainder of the proof now proceeds exactly as in the case of the false alarm error.

3.7.3 Proof of Theorem 16

Define the sets

$$\begin{aligned}
\Theta_0 &= \{Z^n : R_0(\widehat{h}_n) - \alpha > \epsilon_0(n_0,\delta_0,K(n))\} \\
\Theta_1 &= \{Z^n : R_1(\widehat{h}_n) - R_1^* > \inf_{1\le k\le K(n)} \left(\epsilon_1(n_1,\delta_1,k) + \inf_{h\in\mathcal{H}_0^k} R_1(h) - R_1^*\right)\} \\
\Psi_j &= \{Z^n : n_j \le \frac{1}{2}\pi_j n\},\ j = 0,1.
\end{aligned}$$

Observe

$$\begin{aligned}
\mathbb{E}^n\{R_0(\widehat{h}_n)\} - \alpha &= \mathbb{P}^n(\overline{\Theta_0}\cap\overline{\Psi_0})\mathbb{E}^n\{R_0(\widehat{h}_n) - \alpha \mid \overline{\Theta_0}\cap\overline{\Psi_0}\} \\
&\quad + \mathbb{P}^n(\Theta_0\cup\Psi_0)\mathbb{E}^n\{R_0(\widehat{h}_n) - \alpha \mid \Theta_0\cup\Psi_0\} \\
&\le \mathbb{E}^n\{R_0(\widehat{h}_n) - \alpha \mid \overline{\Theta_0}\cap\overline{\Psi_0}\} + \mathbb{P}^n(\Theta_0\cup\Psi_0) \\
&\le \mathbb{E}^n\{R_0(\widehat{h}_n) - \alpha \mid \overline{\Theta_0}\cap\overline{\Psi_0}\} + \delta_0(n) + \delta_1(n) + e^{-n\pi_0/8} \\
&= \mathbb{E}^n\{R_0(\widehat{h}_n) - \alpha \mid \overline{\Theta_0}\cap\overline{\Psi_0}\} + O\left(1/\sqrt{n}\right),
\end{aligned}$$

where the next to last step follows from Theorem 14 and Lemma 18, and the last step follows from the assumption $\delta_j(n) = O(1/\sqrt{n})$.

Thus it suffices to show $R_0(\widehat{h}_n) - \alpha$ decays at the desired rate whenever $Z^n \in$

$\overline{\Theta_0} \cap \overline{\Psi_0}$. For such Z^n we have

$$\begin{aligned} R_0(\widehat{h}_n) - \alpha &\leq \epsilon_0(n_0, \delta_0(n), K(n)) \\ &= \sqrt{2\frac{\log|\mathcal{H}^{K(n)}| + K(n)\log 2 + \log(2/\delta_0(n))}{n_0}} \end{aligned}$$

Since $Z^n \in \overline{\Psi_0}$, we know $\frac{1}{n_0} \leq \frac{2}{\pi_0 n}$. By assumption, $\log(1/\delta_0(n)) = O(\log n)$. Furthermore, from the discussion prior to the statement of Theorem 16, we know $\log|\mathcal{H}^{K(n)}| \leq K(n)(\log 8 + \log d)$ for n sufficiently large. Combining these facts yields

$$R_0(\widehat{h}_n) - \alpha \preccurlyeq \sqrt{\frac{K(n)}{n}}.$$

Plugging in $K(n) \asymp n^{\frac{d-1}{d+1}}$ gives

$$R_0(\widehat{h}_n) - \alpha \preccurlyeq n^{-1/(d+1)} \tag{3.10}$$

as desired.

For the miss probability, it can similarly be shown that

$$\mathbb{E}^n\{R_1(\widehat{h}_n)\} - R_1^* \leq \mathbb{E}^n\{R_1(\widehat{h}_n) - R_1^* \mid \overline{\Theta_1} \cap \overline{\Psi_1}\} + O\left(1/\sqrt{n}\right).$$

Thus, it suffices to consider $Z^n \in \overline{\Theta_1} \cap \overline{\Psi_1}$. Our strategy is to find a tree $\tilde{h} \in \mathcal{H}_0^{\tilde{k}}$ for some $\tilde{k} \leq K(n)$ such that

$$\epsilon_1(n_1, \delta_1(n), \tilde{k})$$

and

$$R_1(\tilde{h}) - R_1^*$$

both decay at the desired rate. The result will then follow by the oracle inequality implied by $Z \in \overline{\Theta_1}$.

Let m be a dyadic integer (a power of two) such that $4dc_1m^{d-1} \leq K(n)$ and $m \leq 2^L$ and $m \asymp n^{1/(d+1)}$. Note that this is always possible by the assumptions $K(n) \asymp n^{(d-1)/(d+1)}$ and $2^L \succcurlyeq n^{1/(d+1)}$. Recall that $\mathcal{P}_m$ denotes the partition of $[0,1]^d$ into hypercubes of sidelength $1/m$. Define $\mathcal{B}_m$ to be the collection of all cells in $\mathcal{P}_m$ that intersect the optimal decision boundary ∂G^*. By the box counting hypothesis

(**A1**), $|\mathcal{B}_m| \leq c_1 m^{d-1}$ for all m.

Construct $\tilde{h}$ as follows. We will take $\tilde{h}$ to be a *cyclic* DDT. A cyclic DDT is a DDT such that $s(A) = 1$ when A is the root node and if A is a cell with child A', then $s(A') \equiv s(A) + 1 \pmod{d}$. Thus cyclic DDTs may be "grown" by cycling through the coordinates and splitting at the midpoint. Define $\tilde{h}$ to be the cyclic DDT consisting of all the cells in $\mathcal{B}_m$, together with their ancestors, and their ancestors' children. In other words, $\tilde{h}$ is the smallest cyclic DDT containing all cells in $\mathcal{B}_m$ among its leaves. Finally, label the leaves of $\tilde{h}$ so that they agree with the optimal classifier g^* on cells not intersecting ∂G^*, and label cells intersecting ∂G^* with class 0. By this construction, $\tilde{h}$ satisfies $R_0(\tilde{h}) \leq \alpha$. Note that $\tilde{h}$ has depth $J = d \log_2 m$.

By the following lemma we know $\tilde{h} \in \mathcal{H}^{\tilde{k}}$ for some $\tilde{k} \leq K(n)$.

Lemma 21. *Let $\tilde{k}$ denote the number of leaf nodes of $\tilde{h}$. Then $\tilde{k} \leq K(n)$.*

Proof. Observe that only those nodes in $\tilde{h}$ that intersect ∂G^* can be ancestors of nodes in $\mathcal{B}_m$. By the box-counting hypothesis, there are at most $c_1 2^{\lceil j/d \rceil (d-1)}$ nodes of $\tilde{h}$ at depth j that can intersect ∂G^*. Hence, there are at most

$$\sum_{j=0}^{J} c_1 2^{\lceil j/d \rceil (d-1)} \leq \sum_{\ell=0}^{J/d} d c_1 2^{\ell (d-1)} \leq 2 d c_1 2^{(J/d)(d-1)} = 2 d c_1 m^{d-1}$$

ancestors of cells in $\mathcal{B}_m$. Since the leaf nodes of $\tilde{h}$ are the children of ancestors of cells in $\mathcal{B}_m$, it follows that $\tilde{k} \leq 4 d c_1 m^{d-1} \leq K(n)$. □

Applying the lemma we have

$$\begin{aligned} \epsilon_1(n_1, \delta_1(n), \tilde{k}) &\leq \epsilon_1(n_1, \delta_1(n), K(n)) \\ &\preccurlyeq n^{-1/(d+1)}, \end{aligned}$$

where the last step follows by the same argument that produced Equation (3.10).

To bound the approximation error, observe

$$\begin{aligned} R_1(\tilde{h}) - R_1^* &\leq \sum_{A \in \mathcal{B}_m} \mathbb{P}_{X|Y=1}(A) \\ &\leq |\mathcal{B}_m| c_0 m^{-d} \\ &= \tilde{k} c_0 m^{-d} \\ &\leq 4 d c_0 c_1 m^{d-1} m^{-d} \\ &\preccurlyeq m^{-1} \\ &\asymp n^{-1/(d+1)}, \end{aligned}$$

where the second inequality follows from hypothesis **A0**. This completes the proof.

Chapter 4

Tree Pruning with Sub-Additive Penalties

In this chapter we study the problem of pruning a binary tree by minimizing, over all pruned subtrees of the given tree, an objective function that sums an additive cost with a non-additive penalty that depends only on tree size. In particular, we focus on sub-additive penalties (roughly, penalties that grow more slowly than an additive penalty) which are motivated by recent results in statistical learning theory. Consider the family of optimally prunings generated by varying the scalar multiplier of a sub-additive penalty. We show that this family is a subset of the analogous family produced by an additive penality. This implies (by known results about additive penalties) that the trees generated by a sub-additive penalty (1) are nested; (2) are unique; and (3) can be computed efficiently. It also implies that, when a single tree is to be selected by cross-validation from the family of prunings, sub-additive penalties will never present a richer set of options than an additive penalty.

4.1 Introduction

Tree-based methods are one of the most widely applied techniques in all of applied mathematics and engineering, from nonparametric statistics to machine learning to multiscale signal and image processing. In this paper we focus on *pruning* trees via complexity regularization, a task that often occurs in the design of tree-based methods. Rather than focusing on a specific application or pruning problem, we present general results that apply in a number of different settings.

The formal statement of our problem is as follows. In graph theory a tree is simply a connected graph without cycles. We consider a specific kind of tree that we call a *rooted binary tree* which has the following properties:

- there exists a unique node with degree 2
- all other nodes have degree 1 or 3.

The degree of a node is the number of edges linking that node to other nodes (called neighbors). The node with degree 2 is called the *root* node. Nodes with degree greater

than 1 are called *internal* nodes and nodes with degree 1 are called *terminal* or *leaf* nodes. The *depth* of a node t is the length of the path between t and the root node. Every node t except the root has a *parent* which is the unique neighbor of t whose depth is one less than t's. Every internal node t has two *children* which are the two neighbors of t having depth one more than t's. Rooted binary trees are readily envisioned by picturing the root node at the top of the graph and the remaining nodes "dangling" down in such a way that parents are above their children and one child always branches left while the other branches right.

The set of leaf nodes of T is denoted $\tilde{T}$. The *size* of a tree T is the number of leaf nodes and denoted $|T|$. A *subtree* of T is a subgraph $S \subseteq T$ that is a rooted binary tree in its own right. If S is a subtree that contains the root of T we say S is a *pruned subtree* of T and write $S \preccurlyeq T$.

For the remainder of the paper let T be a fixed rooted binary tree. Let ρ be a functional mapping subtrees of T to the positive reals. Let Φ be a mapping from the positive integers to the positive reals. We make the following assumptions on ρ and Φ.

- ρ is monotonically nonincreasing, that is, $S_1 \preccurlyeq S_2 \Rightarrow \rho(S_1) \geq \rho(S_2)$
- Φ is monotonically increasing, that is, $k_1 < k_2 \Rightarrow \Phi(k_1) < \Phi(k_2)$
- ρ is *additive*, that is,
$$\rho(S) = \sum_{t \in \tilde{S}} \rho(t).$$

We are interested in algorithms computing and theorems describing two kinds of pruning problems. The first is

$$T^* = \arg\min_{S \preccurlyeq T} \ \rho(S) + \Phi(|S|). \tag{4.1}$$

If multiple trees achieve the minimum, choose T^* to be the smallest. Note that T^* is still not necessarily unique. The problem of solving (4.1) is called *single pruning*, in contrast with *family pruning* described below.

We refer to $\rho(S)$ and $\Phi(|S|)$ as the *cost* and *penalty* of S, respectively. Conceptually, every $S \preccurlyeq T$ is a model that explains some observed phenomenon. Typically $S = T$ is the most complicated model while the root node is the simplest. The idea

behind pruning is to find a model that appropriately balances the the complexity of S with the fidelity of S to an observed phenomenon.

One of the earliest and perhaps most widely known examples of this kind of pruning problem comes from the method of Classification and Regression Trees (CART) of Breiman et al. (1984). In CART, a training dataset $(X_i, Y_i)_{i=1}^n$ is given, where the X_i are feature vectors and the labels $Y_i \in \{1, \ldots, M\}$ for classification and $Y_i \in \mathbb{R}$ for regression. The training data is used to construct an initial tree T that "overfits" the training data (for example, classifying every training sample correctly), and the purpose of pruning is to select a tree $S \preccurlyeq T$ that generalizes to accurately predict the correct Y for unlabeled X observed in the future.

For classification (aka decision) trees, each node $t \in T$ is assigned a class label y_t by majority vote over the training samples reaching t, and $\rho(S)$ is taken to be the empirical error

$$\rho(S) = \frac{1}{n} \sum_{t \in \tilde{S}} \sum_{i: X_i \in t} \mathbb{I}_{\{Y_i \neq y_t\}}.$$

Here $\mathbb{I}$ denotes the indicator function. For regression trees each $t \in T$ is assigned the empirical average

$$y_t = \frac{1}{|\{i : X_i \in t\}|} \sum_{i: X_i \in t} Y_i$$

and $\rho(S)$ is the average empirical squared error

$$\rho(S) = \frac{1}{n} \sum_{t \in \tilde{S}} \sum_{i: X_i \in t} (Y_i - y_t)^2.$$

For a penalty CART uses $\Phi(|S|) = \lambda |S|$ where $\lambda > 0$ is some constant. Additional examples of costs and penalties for tree structured source coding may be found in Chou, Lookabaugh, and Gray (1989).

In many applications it is not known precisely how to calibrate ρ with respect to Φ so as to achieve an optimal pruning. In such cases it is customary to introduce a tuning parameter α, solve

$$T^*(\alpha) = \arg\min_{S \preccurlyeq T} \ \rho(S) + \alpha \Phi(|S|). \tag{4.2}$$

for several different values of α, and choose the best α by cross-validation or some

other means. This is the second pruning problem we consider in this paper.

Since T is in general finite it follows that there exist constants $0 = \alpha_0 < \alpha_1 < \ldots < \alpha_m = \infty$ and pruned subtrees $R_1, \ldots, R_m$ such that

$$\alpha \in [\alpha_{\ell-1}, \alpha_\ell) \Rightarrow T(\alpha) = R_\ell.$$

Moreover, since Φ is increasing it follows that $|R_1| > \cdots > |R_m| = 1$ (see Section 4.2 for further discussion). We refer to $R_1, \ldots, R_m$ as the *family of prunings* of T with respect to ρ and Φ. The problem of computing these subtrees and thresholds is called *family pruning*.

4.1.1 Motivation

Single and family pruning have been studied extensively in the case where Φ is *additive*, by which we mean $\Phi(|S|) = \lambda|S|$ for some $\lambda > 0$ (for family pruning it suffices to take $\lambda = 1$). Additive penalties are by far the most popular choice for Φ, owing in large part to the existence of computationally efficient algorithms (discussed later) for computing T^* and the family of prunings of T. Moreover, the family of prunings satisfies the desirable properties that the trees R_ℓ are unique and nested. In many cases, however, the choice of an additive penalty appears to have no other grounding besides computational convenience.

Rroughly speaking, sub-additive penalties are penalties that grow more slowly than additive penalties as a function of tree size (a precise definition is given in Section 4.4). For example, $\Phi(k) \propto k^\tau$ for $0 < \tau \leq 1$ defines a sub-additive penalty. Several theoretical results, many of them recent, suggest that sub-additive penalties may be more appropriate than additive penalties for certain applications. Barron (1991) demonstrates risk bounds for bounded loss functions which, when applied to classification or regression trees, imply a penalty of $\Phi(k) \propto \sqrt{k}$. Mansour and McAllester (2000); Nobel (2002); Scott and Nowak (2002) also derive risk bounds for classification trees with $\Phi(k) \propto \sqrt{k}$. Mansour and McAllester (2000) and Langford (2002) derive penalties for classification trees that vary between $\Phi(k) \propto k$ and $\Phi(k) \propto \sqrt{k}$. Meanwhile, classification risk bounds for additive penalties are only known for the special "zero error" case (when the optimal classifier is correct with probability one) and under the more general but still quite restrictive "identifiability" assumption

of Blanchard et al. (2004). In summary, sub-additive penalties appear to have a much stronger theoretical foundation than additive ones in certain settings, especially classification.

4.1.2 Overview

The purpose of this paper is to present algorithms and relevant properties for single and family pruning with non-additive, and in particular sub-additive penalties. One of our main results is that the family of prunings generated by a sub-additive penalty is a subset of the family of prunings generated by the additive penalty. Positive implications of this fact are that sub-additive families are nested and unique. It also leads to a simple algorithm for generating the family. A negative implication, however, is that, when a tree is to be selected from the family of prunings by cross-validation, sub-additive penalties never provide a richer class of options than the additive penalty.

The paper is organized as follows. In Section 4.2 we study pruning with general size-based penalties. We give explicit algorithms for single pruning and family pruning and provide a geometric framework for interpreting the family of prunings. This section brings together several known results and perspectives, adds a few new insights, and sets the stage for our later discussion of sub-additive penalties. In Section 4.3 we review algorithms and properties related to pruning with additive penalties. In Section 4.4 we define dominating and sub-additive penalties and prove a general theorem about nested families of prunings. We also explore in more detail the implications of this theorem as outlined above. Section 4.5 contains discussion and conclusions.

4.2 General Sized-Based Penalties

We first present a general algorithm for single pruning when $\Phi(k)$ is arbitrary. This first algorithm applies even if Φ is not necessarily increasing. The algorithm should not be considered novel; its key components have appeared previously in other guises as discussed below.

For each $k = 1, 2, \ldots, |T|$, define T^k to be a pruned subtree $S \preccurlyeq T$ (there may be more than one) minimizing $\rho(S)$ subject to $|S| = k$. These trees are referred to as

minimum cost trees. Observe that

$$T^* = \arg\min\{\rho(T^k) + \Phi(k) \,:\, k = 1, 2, \ldots, |T|\} \tag{4.3}$$

is a solution to (4.1). In other words, it suffices to construct the sequence T^k and minimize the objective function over this collection. For the remainder of the paper, fix choices of T^k whenever T^k is not unique.

4.2.1 Computing Minimum Cost Trees

Let the nodes of T be indexed 1 through $2|T| - 1$ in such a way that children have a larger index than their parents. (We refer to nodes and their indices interchangeably.) Let T_t denote the subtree rooted at node t and containing all of its descendants in T (thus $T = T_1$). Let $l(t)$ and $r(t)$ denote the left and right children of node t, respectively. If U and V are pruned subtrees of $T_{l(t)}$ and $T_{r(t)}$, let $\textsc{merge}(t, U, V)$ denote the pruned subtree of T_t having U and V as its left and right subtrees, respectively. Finally, let T_t^i denote the pruned subtree of T_t having minimum cost among all pruned subtrees of T_t with i leaf nodes.

The algorithm for computing minimum cost trees is based on the following fact: If we know $T_{l(t)}^i$ and $T_{r(t)}^j$ for $i = 1, 2, \ldots, |T_{l(t)}|$ and $j = 1, 2, \ldots, |T_{r(t)}|$, it is a simple matter to find T_t^k, $k = 1, 2, \ldots, |T_t|$. For each $k = 1, 2, \ldots, |T_t|$, there exist i, j with $i + j = k$ such that $T_t^k = \textsc{merge}(t, T_{l(t)}^i, T_{r(t)}^j)$. This follows from additivity of ρ. Moreover, if $T_t^k = \textsc{merge}(t, T_{l(t)}^i, T_{r(t)}^j)$, then $\rho(T_t^k) = \rho(T_{l(t)}^i) + \rho(T_{r(t)}^j)$. We may then set $T_t^k = \textsc{merge}(t, T_{l(t)}^{i^*}, T_{r(t)}^{j^*})$, where i^*, j^* minimize $\rho(T_{l(t)}^i) + \rho(T_{r(t)}^j)$ over all i, j such that $i + j = k$, $1 \le i \le |T_{l(t)}|$, and $1 \le j \le |T_{r(t)}|$. Note that i^*, j^* are determined by exhaustive search.

This step may be applied at each level of T, working from the bottom up, and leads to an algorithm for computing the minimum cost trees, and hence for determining T^*. The complete algorithm is presented in Figure 4.1. The computational complexity of computing the minimum cost trees is $O(|T|^2)$. This was proved by Bohanec and Bratko (1994). The algorithm takes longer to run when T is more balanced. If T is maximally lopsided, e.g., all right children are terminal nodes, the algorithm computes the minimum cost trees in $O(|T|)$ operations.

The procedure described above for determining minimum cost trees is essentially

Input:
Initial tree T
Main Loop:
For $t = 2|T| - 1$ downto 1
Set $T_t^1 = \{t\}$;
If t is not a terminal node, Then
For $k = 2$ to $|T_t|$
Set mincost $= \infty$;
For $i = \max(1, k - |T_{r(t)}|)$ to $\min(|T_{l(t)}|, k - 1)$
Set $j = k - i$;
Set cost $= \rho(T_{l(t)}^i) + \rho(T_{r(t)}^j)$;
If cost $<$ mincost, Then
Set mincost $=$ cost;
Set $T_t^k = \text{MERGE}(t, T_{l(t)}^i, T_{r(t)}^j)$;
End If
End For
End For
End If
End For
Output:
The minimum cost trees $T^k = T_1^k, k = 1, 2, \ldots, |T|$

Figure 4.1 : An algorithm for computing minimum cost trees. The limits for the innermost "For" loop in Figure 4.1 ensure that i, j satisfy $i + j = k$, $1 \leq i \leq |T_{l(t)}|$, and $1 \leq j \leq |T_{r(t)}|$.

the dual of an algorithm first described by Bohanec and Bratko (1994). They considered the problem of finding the pruned subtree with smallest size among all pruned subtrees with empirical error below a certain threshold. This procedure was apparently known to the CART authors. As reported by Bohanec and Bratko (1994), "Leo Breiman (Private Communication, December 1990) did implement such an algorithm for optimal pruning; he was satisfied that it worked, but no futher development was done, and the algorithm was not published." As far as we know, the present work is the first to point out the use of minimum cost trees for single pruning with general size-based penalties.

4.2.2 Geometric aspects of family pruning

To each $S \preccurlyeq T$ associate the function $f_S : [0, \infty) \to \mathbb{R}$, $\alpha \mapsto f_S(\alpha) = \rho(S) + \alpha\Phi(|S|)$. In this way each pruned subtree maps to a line in the plane, as shown in Figure 4.2. Define

$$f^*(\alpha) = \min_{S \preccurlyeq T} f_S(\alpha).$$

Clearly f^* has the form

$$f^*(\alpha) = f_{R_\ell}(\alpha), \quad \alpha \in [\alpha_{\ell-1}, \alpha_\ell)$$

for some constants $0 = \alpha_0 < \alpha_1 < \cdots < \alpha_m = \infty$ and subtrees $R_\ell \preccurlyeq T, \ell = 1, 2, \ldots m$. Moreover, since Φ is monotonically increasing, we conclude $\Phi(R_{\ell-1}) > \Phi(R_\ell)$ which implies $|R_{\ell-1}| > |R_\ell|$. These observations are summarized as follows.

Proposition 8. *If $\Phi(k)$ is monotonically increasing in k, then there exist constants $0 = \alpha_0 < \alpha_1 < \cdots < \alpha_m = \infty$ and pruned subtrees $R_\ell \preccurlyeq T, \ell = 1, 2, \ldots m$, with $|R_1| > \cdots > |R_m| = 1$, such that $T(\alpha) = R_\ell$ whenever $\alpha \in [\alpha_{\ell-1}, \alpha_\ell)$.*

This picture also provides us with an algorithm for determining the α_ℓ and R_ℓ. Observe that each R_ℓ must be a minimum cost tree T^k. Therefore

$$f^*(\alpha) = \min_k f_{T^k}(\alpha).$$

Clearly $R_1 = T^{k_1}$ where k_1 is the smallest k such that $\rho(T^k) = \rho(T)$. Now observe that, assuming $i < j$, f_{T^i} and f_{T^j} intersect at the point

$$\gamma_{i,j} = \frac{\rho(T^i) - \rho(T^j)}{\Phi(j) - \Phi(i)}.$$

Therefore, if $R_\ell = T^{k_\ell}$, then $R_{\ell+1} = T^{k_{\ell+1}}$, where

$$k_{\ell+1} = \arg\min_{k < k_\ell} \gamma_{k,k_\ell}.$$

If multiple k minimize the right-hand side, let $k_{\ell+1}$ be the smallest. Furthermore, we have

$$\alpha_\ell = \gamma_{k_{\ell+1},k_\ell}.$$

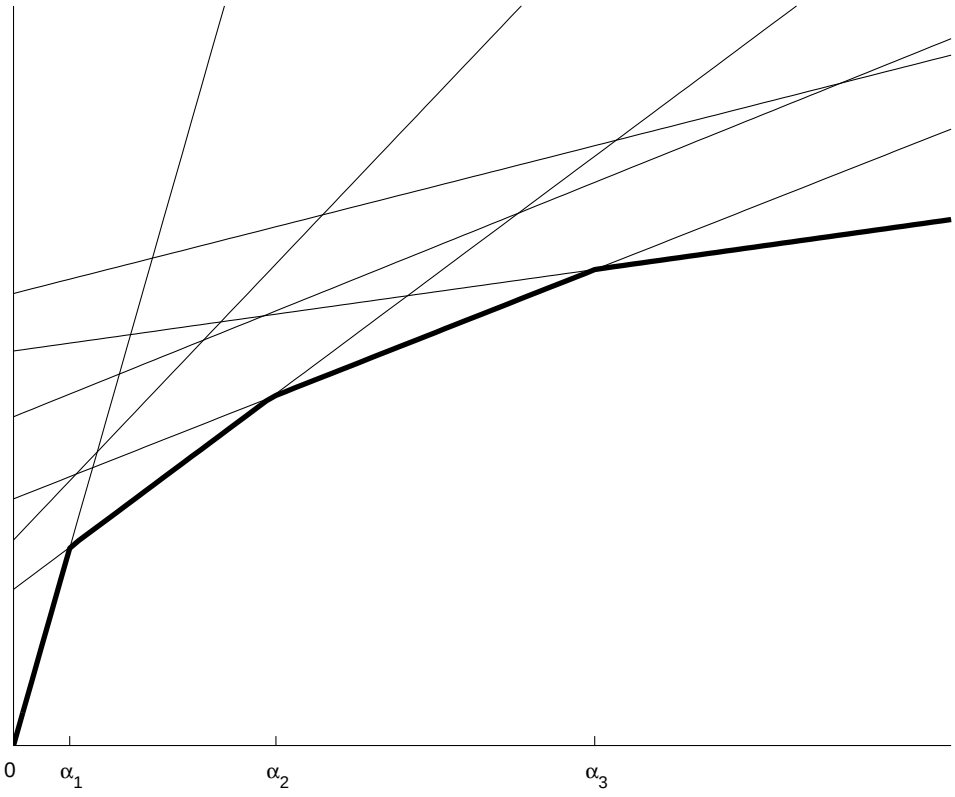


Figure 4.2 : Hypothetical plots of $\rho(S) + \alpha\Phi(|S|)$ as a function of α for all $S \preccurlyeq T$. Pruned subtrees coinciding with the minimum of these functions (shown in bold) over a range of α minimize the pruning criterion for those α.

Input:
Minimum cost trees $T^k, k = 1, \ldots, |T|$
Initialization:
Set $k_1 = \arg\min\{k : \rho(T^k) = \rho(T)\}$;
Set $R_1 = T^{k_1}$;
Set $\ell = 1$;
Main Loop:
While $k_\ell > 1$
Set $\alpha_\ell = \infty$;
For $k = k_\ell - 1$ downto 1
Set $\gamma_{k,k_\ell} = (\rho(T^k) - \rho(T^{k_\ell}))/(\Phi(k_\ell) - \Phi(k))$;
If $\gamma_{k,k_\ell} \leq \alpha_\ell$
Set $\alpha_\ell = \gamma_{k,k_\ell}$;
Set $k_{\ell+1} = k$;
End If
End For
Set $\ell = \ell + 1$;
Set $R_\ell = T^{k_\ell}$;
End While
Output:
The family of prunings R_ℓ and thresholds α_ℓ.

Figure 4.3 : An algorithm generating the family of prunings and associated thresholds for an arbitrary increasing penalty.

This algorithm is summarized in Figure 4.3.

We also highlight a property inherent in the definition of k_ℓ that will be of use later.

Lemma 22. *If $k < k_\ell$, then $\gamma_{k_{\ell+1},k_\ell} \leq \gamma_{k,k_\ell}$. If $k < k_{\ell+1}$, then $\gamma_{k_{\ell+1},k_\ell} < \gamma_{k,k_\ell}$.*

A second geometric picture due to Chou et al. (1989) offers essentially equivalent insights into the family of prunings of T. Consider the set of points $\mathcal{P} = \{p(S) = (\rho(S), \Phi(|S|)) \mid S \preccurlyeq T\} \subset \mathbb{R}^2$, as depicted in Figure 4.4. The point corresponding to R_m (= the root of T) is furthest down and to the right. The point corresponding to R_1 is furthest up and to the left (assuming $R_1 = T$). Moreover, the points corresponding to $R_\ell, \ell = 1, \ldots, m$, are the vertices of the lower boundary of the convex hull of $\mathcal{P}$,

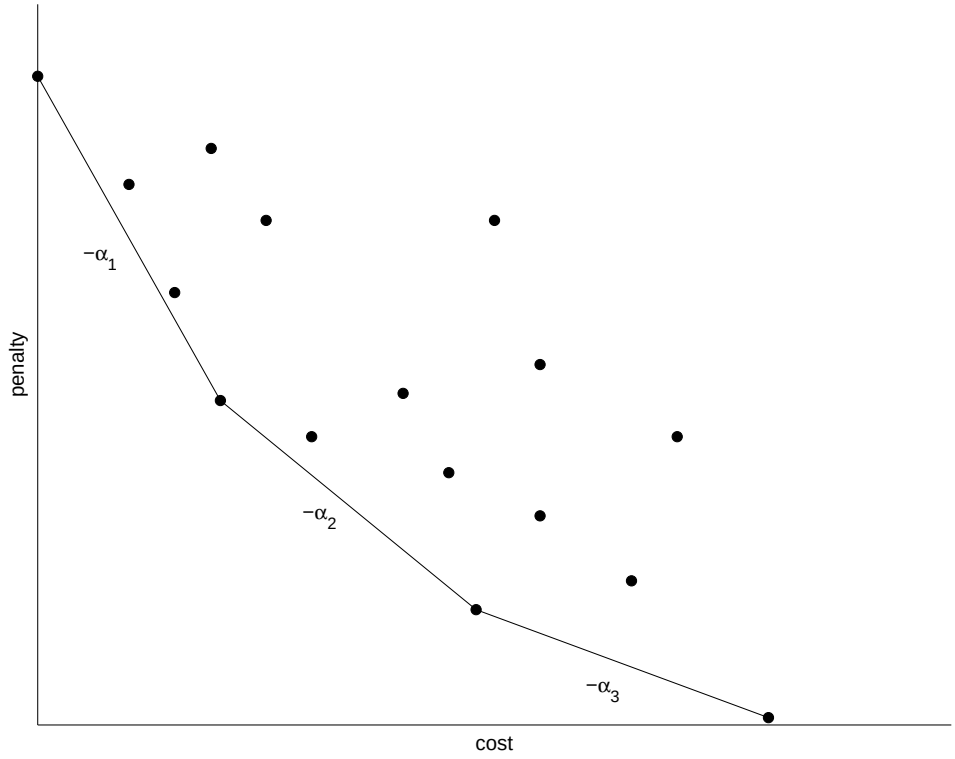


Figure 4.4 : Hypothetical plot of points $(\rho(S), \Phi(S))$ for all $S \preccurlyeq T$. The family of prunings consists of points on the lower boundary of the convex hull of these points, and the (negative) slopes between vertices correspond to the thresholds α_ℓ.

listed counterclockwise. Thus, α_ℓ is the negative of the slope of the line segment connecting $p(R_\ell)$ to $p(R_{\ell+1})$. The algorithm described above for generating α_ℓ and R_ℓ can now be rederived in this setting by starting with R_1 and successively learning faces of the lower boundary of the convex hull of $\mathcal{P}$ in a counterclockwise fashion.

4.3 Additive Penalties

When $\Phi(|T|) = \lambda|T|$ for some $\lambda > 0$, there exist faster algorithms for single and family pruning than those described in the previous section. Moreover, the optimally pruned trees satisfy certain nice properties. The material in this section is taken from Breiman et al. (1984, chap. 10).

When λ is known, T^* may be computed by a simple bottom-up procedure. In

particular, denoting

$$T_t^* = \arg\min_{S \preccurlyeq T_t} \ \rho(S) + \lambda|S|,$$

we have $T_t^* = \{t\}$ for leaf nodes and for internal nodes

$$T_t^* = \arg\min\{\rho(S) + \lambda|S| : S = \{t\} \text{ or } S = \textsc{merge}(t, T_{l(t)}^*, T_{r(t)}^*\}.$$

This last fact follows easily by additivity of ρ and Φ and by induction on t. This is an $O(|T|)$ algorithm for computing T^*, much faster than the more general $O(|T|^2)$ algorithm described previously. Moreover, Breiman et al. (1984) show that T^* is unique.

Breiman et al. (1984) also prove the following theorem about the family of prunings generated by an additive penalty.

Theorem 17 (Breiman, Friedman, Olshen and Stone). *If $\Phi(k) = k$, then there exist weights $0 = \alpha_0 < \alpha_1 < \cdots < \alpha_m = \infty$ and subtrees $T \succcurlyeq R_1 \succcurlyeq \cdots \succcurlyeq R_m = \{root\}$ such that $T(\alpha) = R_\ell$ whenever $\alpha \in [\alpha_{\ell-1}, \alpha_\ell)$.*

In particular, the result is an improvement over Proposition 8 because the family of prunings is *nested*. We refer to the family $\{R_\ell\}$ in Theorem 17 as the CART trees.

Breiman et al. (1984) further demonstrate how the nesting property leads to an algorithm for finding these weights and subtrees. The algorithm has a worst case running time of $O(|T|^2)$, the worst case being when T is unbalanced. However, when T is balanced (i.e., when the depth of T is proportional to $\log|T|$, then the algorithm runs in time $O(|T|\log|T|)$. This yields an improvement (relative to the algorithm in Figure 4.3) for many problems, such as those in signal and image processing, where the initial tree is balanced.

The proof of Theorem 17 given by Breiman et al. (1984) is algebraic; an alternative geometric proof is given by Chou et al. (1989). These authors also extend the theorem to *affine* costs and penalties. An separate algebraic account may be found in Ripley (1996).

4.4 Sub-additive penalties

In this, the main section of the paper, we introduce sub-additive penalties and show that for such penalties, the family of prunings is a subset of the CART trees. Thus,

these trees are also unique, nested, and may be computed using the CART trees.

Definition 1. Let Φ^1 and Φ^2 be two increasing penalties. We say Φ^1 *dominates* Φ^2, denoted $\Phi^1 \gg \Phi^2$, if, for all positive integers $a > b > c$, we have

$$\frac{\Phi^2(a) - \Phi^2(b)}{\Phi^2(a) - \Phi^2(c)} \leq \frac{\Phi^1(a) - \Phi^1(b)}{\Phi^1(a) - \Phi^1(c)}. \tag{4.4}$$

If $\Phi^1(k) = k$ and $\Phi^1 \gg \Phi^2$, we say Φ^2 is *sub-additive*.

An important example of a sub-additive penalty is the square root penalty $\Phi^2(k) = \sqrt{k}$. To see that this is indeed sub-additive, observe that for $a > b > c$,

$$\begin{aligned}
\frac{\Phi^2(a) - \Phi^2(b)}{\Phi^2(a) - \Phi^2(c)} &= \frac{\sqrt{a} - \sqrt{b}}{\sqrt{a} - \sqrt{c}} \\
&< \frac{\sqrt{a} - \sqrt{b}}{\sqrt{a} - \sqrt{c}} \cdot \frac{\sqrt{a} + \sqrt{b}}{\sqrt{a} + \sqrt{c}} \\
&= \frac{a - b}{a - c} \\
&= \frac{\Phi^1(a) - \Phi^1(b)}{\Phi^1(a) - \Phi^1(c)}
\end{aligned}$$

where $\Phi^1(k) = k$.

More generally, the following result characterizes a large class of penalties with $\Phi^1 \gg \Phi^2$.

Proposition 9. *Let f, g be real valued, twice differentiable functions on $(0, \infty)$, and for $k = 1, 2, \ldots$, set $\Phi^1(k) = f(k)$ and $\Phi^2(k) = g(k)$. Let x_0 be a positive real number. If $0 < g'(x) \leq f'(x)$ and $g''(x) \leq f''(x)[g'(x)/f'(x)]$ for all $x \geq x_0$, then the inequality (4.4) is satisfied for all real numbers $a > b > c \geq x_0$. Therefore, if $x_0 \leq 1$, then $\Phi^1 \gg \Phi^2$.*

Proof. The proof has two main steps: First, we prove the lemma for the special case $f(x) = x$, and then we use the first step to establish the general case by reparametrizing $\mathbb{R}$.

Assume for now $f(x) = x$. Then, by assumption, $g'(x) \leq 1$ and $g''(x) \leq 0$ for

$x \geq x_0$. Let $a > b > c \geq x_0$. Note that inequality (4.4) is equivalent to

$$\frac{\Phi^2(a) - \Phi^2(b)}{\Phi^2(b) - \Phi^2(c)} \leq \frac{\Phi^1(a) - \Phi^1(b)}{\Phi^1(b) - \Phi^1(c)}, \tag{4.5}$$

which can be seen by writing $\Phi^\kappa(a) - \Phi^\kappa(c) = (\Phi^\kappa(a) - \Phi^\kappa(b)) + (\Phi^\kappa(b) - \Phi^\kappa(c))$, for $\kappa = 1, 2$, and simplifying. Also note that Φ^1 and Φ^2 are monotonically increasing from the assumption on the first derivative. By the fundamental theorem of calculus, $\Phi^2(a) - \Phi^2(b) = \int_b^a g'(x)\,dx \leq g'(b)(a-b)$, where we use the concavity of g in the last step. Similarly, $\Phi^2(b) - \Phi^2(c) = \int_c^b g'(x) \geq g'(b)(b-c)$. Summarizing, we have shown

$$\frac{\Phi^2(a) - \Phi^2(b)}{a - b} \leq g'(b) \leq \frac{\Phi^2(b) - \Phi^2(c)}{b - c},$$

which by (4.5) implies the theorem for this special case.

Now consider the general case. Define $\tilde{f}(x) = x$ and $\tilde{g}(x) = g(f^{-1}(x))$. Now $\tilde{f}'(x) = 1$, while $\tilde{g}'(x) = g'(f^{-1}(x))/f'(f^{-1}(x))$, which is ≤ 1 provided $x \geq \tilde{x}_0 := f(x_0)$. In addition, $\tilde{f}''(x) = 0$ and

$$\tilde{g}''(x) = \frac{g''(f^{-1}(x)) - f''(f^{-1}(x))[g'(f^{-1}(x))/f'(f^{-1}(x))]}{(f'(f^{-1}(x)))^2},$$

which is ≤ 0 if $x \geq \tilde{x}_0$. Thus, we may apply the previous case to $\tilde{f}$ and $\tilde{g}$. For all real numbers $x > y > z \geq f(x_0)$, we have

$$\frac{g(f^{-1}(x)) - g(f^{-1}(y))}{g(f^{-1}(y)) - g(f^{-1}(z))} \leq \frac{x - y}{y - z}.$$

By taking $x = f(a), y = f(b)$, and $z = f(c)$, and by monotonicity of f, we conclude

$$\frac{g(a) - g(b)}{g(b) - g(c)} \leq \frac{f(a) - f(b)}{f(b) - f(c)},$$

which is what we wanted to show. □

The following corollary gives a concrete example of a family of penalties to which Proposition 9 applies.

Corollary 4. *Let $\sigma \geq \tau > 0$, and set $\Phi^1(k) = k^\sigma$ and $\Phi^2(k) = k^\tau$. Then $\Phi^1 \gg \Phi^2$.*

Hence, if $0 < \tau \leq 1$, *then* Φ^2 *is sub-additive.*

Proof. Define $f(x) = x^\sigma$ and $g(x) = x^\tau$. For $x \geq 1 = x_0$

$$g'(x) = \tau x^{\tau-1} \leq \sigma x^{\sigma-1} = f'(x).$$

Furthermore, for $x \geq 1$,

$$\begin{aligned} g''(x) &= \tau(\tau-1)x^{\tau-2} \\ &\leq \tau(\sigma-1)x^{\tau-2} \\ &= \left(\sigma(\sigma-1)x^{\sigma-2}\right)\left(\frac{\tau x^{\tau-1}}{\sigma x^{\sigma-1}}\right) \\ &= f''(x)\left(\frac{g'(x)}{f'(x)}\right). \end{aligned}$$

Now apply Proposition 9. □

Further examples of dominating and sub-additive penalties may be derived from the following result.

Proposition 10. *If* f, g, *and* x_0 *satisfy the hypothesis of Proposition 9, then so do* $\tilde{f}(x) = f(h(x))$, $\tilde{g}(x) = g(h(x))$, *and* $\tilde{x}_0 = h^{-1}(x_0)$ *where* $h : (0,\infty) \to (0,\infty)$ *is any twice differentiable function such that* $h'(x) > 0$ *for all* $x > 0$.

Proof. Observe that for any $x \geq \tilde{x}_0$

$$0 < \tilde{g}'(x) = g'(h(x)) \cdot h'(x) \leq f'(h(x)) \cdot h'(x) = \tilde{f}'(x)$$

and

$$\begin{aligned} \tilde{g}''(x) &= g''(h(x)) \cdot (h'(x))^2 + g'(h(x)) \cdot h''(x) \\ &\leq f''(x)\left(\frac{g'(x)}{f'(x)}\right) \cdot (h'(x))^2 + f'(h(x))\left(\frac{g'(x)}{f'(x)}\right) \cdot h''(x) \\ &= \left(f''(h(x)) \cdot (h'(x))^2 + f'(h(x)) \cdot h''(x)\right)\left(\frac{g'(x)}{f'(x)}\right) \\ &= \tilde{f}''(x)\left(\frac{\tilde{g}'(x)}{\tilde{f}'(x)}\right). \end{aligned}$$

□

4.4.1 Main result

For $\kappa = 1, 2$, and $\alpha \in \mathbb{R}$, define

$$T^{\kappa}(\alpha) = \underset{S \preccurlyeq T}{\arg\min}\ \rho(S) + \alpha\Phi^{\kappa}(|S|).$$

By Proposition 8, there exist scalars $\alpha_0 < \alpha_1 < \ldots < \alpha_m$, and $\beta_0 < \beta_1 < \ldots < \beta_n$, and subtrees $U_1, \ldots, U_m$, and $V_1, \ldots, V_n$, such that

- $\alpha \in [\alpha_{\ell-1}, \alpha_\ell) \Rightarrow T^1(\alpha) = U_\ell$
- $|U_1| > \cdots > |U_m| = 1$
- $\beta \in [\beta_{\ell-1}, \beta_\ell) \Rightarrow T^2(\beta) = V_\ell$
- $|V_1| > \cdots > |V_n| = 1$.

Theorem 18. *With the notation defined above, if Φ^1 and Φ^2 are two increasing penalties such that $\Phi^1 \gg \Phi^2$, then $\{V_1, \ldots, V_n\} \subseteq \{U_1, \ldots, U_m\}$. In other words, for each β, there exists α such that $T^2(\beta) = T^1(\alpha)$.*

An immediate application of the theorem is an alternate method for pruning using a sub-additive penalty. Let $\Phi^1(k) = k$ and let $R_1, \ldots, R_m$ denote the CART trees. These may be computed efficiently by the algorithm of Breiman et al. (1984) or Chou et al. (1989). By Theorem 18, if Φ^2 is sub-additive, then $T^2(\beta)$ is one of these R_ℓ. Therefore,

$$T^2(\beta) = \underset{S \in \{R_1, \ldots, R_m\}}{\arg\min}\ \rho(S) + \beta\Phi^2(|S|).$$

This last minimization may be solved by exhaustive search over the $m \leq |T|$ CART trees.

The theorem also implies a new algorithm for family pruning when Φ^2 is sub-additive. The procedure is exactly like the one described in Figure 4.3, except that one only needs to consider k (see line 3 of the main loop) such that

$$k \in \{i : i = |R_j| \text{ for some } j > \ell\ \}.$$

Thus it is not necessary to compute all minimum cost trees, only the CART trees, which can often be done more efficiently.

We have two distinct algorithms for computing the family of prunings induced by a sub-additive penalty. Both algorithms have worst case running time $O(|T|^2)$. The first algorithm, discussed in Section 4.2, takes longer when T is more balanced, but prunes totally lop-sided trees in $O(|T|)$ time. The second algorithm, just discussed, takes longer when T is unbalanced, and runs in $O(|T| \log |T|)$ time when T is balanced. Conceivably, one could devise a test that determines how balanced a tree is in order to choose which of the two algorithms would be faster on a given tree.

Other properties for pruning with sub-additive penalties follow from Theorem 18 and known results about the CART trees. For example, pruning with a sub-additive penalty always produces unique pruned subtrees, and the family of pruned subtrees is nested.

Families of pruning are useful when the appropriate family member needs to be chosen by cross-validation. When this is the case, Theorem 18 implies that sub-additive penalties will never provide a richer class of options than an additive penalty.

Finally, we remark that the proof of Theorem 18 only requires ρ to be nonincreasing, not necessarily additive. The theorem may also be of practical use in this more general setting.

4.4.2 Proof of Theorem 18

We require the following lemma. Recall that for $i < j$ we define

$$\gamma_{i,j} = \frac{\rho(T^i) - \rho(T^j)}{\Phi(j) - \Phi(i)}.$$

Lemma 23. *Let a, b, c be positive integers with $a > b > c$. The following are equivalent:*

(i) $\gamma_{c,a} < \gamma_{c,b}$

(ii) $\gamma_{b,a} < \gamma_{c,b}$

(iii) $\gamma_{b,a} < \gamma_{c,a}$.

The three statements are also equivalent if we replace $<$ by $\leq, >, \geq$, or $=$.

Proof. A straightforward calculation establishes

$$\begin{aligned}(\Phi(a)-\Phi(c))(\Phi(b)-\Phi(c))\left[\gamma_{c,a}-\gamma_{c,b}\right] &= \\ (\Phi(a)-\Phi(b))(\Phi(b)-\Phi(c))\left[\gamma_{b,a}-\gamma_{c,b}\right] &= \\ (\Phi(a)-\Phi(b))(\Phi(a)-\Phi(c))\left[\gamma_{b,a}-\gamma_{c,a}\right].\end{aligned}$$

The lemma follows from these identities and the fact that Φ is increasing.

The lemma may also be established by geometric considerations. Consider the three points $p_a, p_b, p_c \in \mathcal{P}$ defined by T^a, T^b, T^c respectively (see Section 4.2.2). Note that p_a is above and to the left of p_b, which is above and left of p_c. Then $\gamma_{a,b}$ is the negative slope of the line segment connecting p_a and p_b, and similarly for the other two combinations of points. Then the statements in (i), (ii), and (iii) are all true if and only if p_b is strictly *above* the line connecting p_a and p_c. Similarly, all three statements hold with equality if and only if b lies *on* the line joining p_a and p_c, and so on. □

To prove the theorem, first notice that $U_1 = V_1 =$ the smallest tree $S \preccurlyeq T$ such that $\rho(S) = \rho(T)$. The theorem follows by induction if we can show $V_j \in \{U_1, \ldots, U_m\} \Rightarrow V_{j+1} \in \{U_1, \ldots, U_m\}$. To show this, we suppose it is not true and arrive at a contradiction. Assume there exists j such that $V_j \in \{U_1, \ldots, U_m\}$ but $V_{j+1} \notin \{U_1, \ldots, U_m\}$. Then $V_j = U_i$ for some i. Moreover, there exists $k \geq 0$ such that $|U_{i+k}| > |V_{j+1}| > |U_{i+k+1}|$.

Introduce the notation

$$\gamma^{\kappa}_{i,j} = \frac{\rho(T^i) - \rho(T^j)}{\Phi^{\kappa}(j) - \Phi^{\kappa}(i)},$$

for $\kappa = 1, 2$, and $i < j$. Define $p = |U_i| = |V_j|, q = |U_{i+k}|, r = |V_{j+1}|$, and $s = |U_{i+k+1}|$. Then $p \geq q > r > s$. We will show

$$\gamma^2_{s,r} \leq \gamma^2_{r,q} \leq \gamma^2_{r,p} < \gamma^2_{s,r},$$

thus arriving at our desired contradiction. Denote these inequalities by I1, I2, and I3, respectively.

To establish I1, observe

$$\begin{aligned}
\gamma^2_{s,q} &= \frac{\rho(T^s) - \rho(T^q)}{\Phi^2(q) - \Phi^2(s)} \\
&= \frac{\rho(T^s) - \rho(T^q)}{\Phi^1(q) - \Phi^1(s)} \cdot \frac{\Phi^1(q) - \Phi^1(s)}{\Phi^2(q) - \Phi^2(s)} \\
&\leq \frac{\rho(T^r) - \rho(T^q)}{\Phi^1(q) - \Phi^1(r)} \cdot \frac{\Phi^1(q) - \Phi^1(s)}{\Phi^2(q) - \Phi^2(s)} \\
&= \frac{\rho(T^r) - \rho(T^q)}{\Phi^2(q) - \Phi^2(r)} \cdot \frac{\Phi^2(q) - \Phi^2(r)}{\Phi^1(q) - \Phi^1(r)} \cdot \frac{\Phi^1(q) - \Phi^1(s)}{\Phi^2(q) - \Phi^2(s)} \\
&\leq \frac{\rho(T^r) - \rho(T^q)}{\Phi^2(q) - \Phi^2(r)} \\
&= \gamma^2_{r,q},
\end{aligned}$$

where the first inequality follows from Lemma 22 and the second inequality comes from the definition of $\Phi^1 \gg \Phi^2$. Since $\gamma^2_{s,q} \leq \gamma^2_{r,q}$, Lemma 23 (iii $\Rightarrow$ ii) implies $\gamma^2_{s,r} \leq \gamma^2_{r,q}$, establishing I1.

To show I2, assume $p \neq q$ (otherwise the inequality is trivial). Note that Lemma 22 implies $\gamma^2_{r,p} \leq \gamma^2_{q,p}$. Lemma 23 (iii $\Rightarrow$ i) then implies I2.

Finally, by Lemma 22, we have $\gamma^2_{r,p} < \gamma^2_{s,p}$. I3 follows from Lemma 23 (iii $\Rightarrow$ ii). □

4.5 Conclusion

We have presented two polynomial time algorithms for pruning and generating families of prunings using non-additive penalties. The first algorithm applies to arbitrary penalties, while the second algorithm applies to sub-additive penalties. Both algorithms have a worst-case run time of $O(|T|^2)$. The first algorithm achieves the worst-case for balanced trees (i.e., when $\text{depth}(T) \propto \log |T|$), and only requires $O(|T|)$ operations for lop-sided trees (e.g., when every left descendant is a leaf node). The second algorithm has its worst case when T is unbalanced, and runs in $O(|T| \log |T|)$ time for balanced trees.

The second algorithm is based on a general theorem that, as a special case, implies that the family of prunings induced by a sub-additive penalty is a subset of the family induced by an additive penalty. This implies that sub-additive families are unique

and nested. It also implies a negative result: when cross-validation is to be used to select the best member from a family of prunings, sub-additive penalties will never offer a richer set of options than an additive penalty.

An immediate impact of this work is in the area of classification tree design. It has recently been shown that sub-additive penalties are more appropriate than an additive penalty for pruning classification trees (as discussed in the introduction). The work presented here provides for efficient implementation of such strategies and characterizes the resulting pruned subtrees.

It is quite possible that other machine learning and signal processing tree-based methodologies employ an additive penalty simply for convenience, when perhaps a non-additive penalty would be more appropriate. We hope the present work might lead to a reassessment of such problems.

Chapter 5

Conclusion

Since the emergence of wavelets in the early 1990s, methods based on recursive dyadic partitions (RDPs) have repeatedly offered computationally tractable yet theoretically optimal solutions to a number of problems in nonparametric statistics. This thesis is the first work to apply dyadic thinking to the classification problem, and the aforementioned trend is seen to extend to this domain as well.

While the application of RDPs to classification has much in common with RDPs applied to regression or density estimation, some important differences also exist. Complexity penalties for regression with a squared error loss are proportional to partition size. In classification a 0-1 loss is used leading to penalties proportional to (sums of) the square root of tree size. The fundamentally different nature of the "target" function in classification (the Bayes decision boundary) also requires regularity conditions specified in terms of the complexity of classes of sets, while regression or density estimation often involve regularity assumptions stated in terms of standard function spaces.

This thesis opens several avenues for future research. As discussed in the conclusion of Chapter 2, DDTs decorated with polynomial kernel SVMs should achieve faster rates for very smooth decision boundaries, and may also provide a competitive practical classifier. DDTs will also serve as a platform for comparing and contrasting a number of generalization error bounds that have appeared recently in the machine learning literature. Indeed, the algorithm of Blanchard et al. (2004) discussed in section 2.6 can be extended to implement exact penalized empirical risk minimization for a number of different penalties.

Neyman Pearson learning (Chapter 3) also holds many promising possibilities. As discussed in the conclusion to that chapter, it would be interesting to devise boosting or kernel methods based on Neyman-Pearson and compare them to cost-sensitive learners. Another need is to implement the algorithm for NP-SRM with DDTs and evaluate its performance. Furthermore, a nestedness result (parallel to that in Chapter 4) for family pruning may exist in this setting.

In conclusion, dyadic decision trees yield practical yet theoretically sound discrimination rules and extend the success of dyadic methods from other areas of nonparametric estimation to classification. As in these other areas, RDPs for classification are sufficiently rich to be optimal estimators, but sufficiently sparse to allow globally exact, computationally efficient implementation.

Bibliography

J.-Y. Audibert. *PAC-Bayesian Statistical Learning Theory*. PhD thesis, Université Paris 6, June 2004.

A. Barron. Complexity regularization with application to artificial neural networks. In G. Roussas, editor, *Nonparametric functional estimation and related topics*, pages 561–576. NATO ASI series, Kluwer Academic Publishers, Dordrecht, 1991.

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

P. Bartlett, M. Jordan, and J. McAuliffe. Convexity, classification, and risk bounds. Technical Report 638, Department of Statistics, U.C. Berkeley, 2003.

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

N. Berkman and T. Sandholm. What should be minimized in a decision tree: A re-examination. Technical Report TR 95-20, University of Massachusetts at Amherst, 1995.

G. Blanchard, G. Lugosi, and N. Vayatis. On the rate of convergence of regularized boosting classifiers. *J. Machine Learning Research*, 4:861–894, 2003.

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

M. Bohanec and I. Bratko. Trading accuracy for simplicity in decision trees. *Machine Learning*, 15:223–250, 1994.

O. Bousquet, S. Boucheron, and G. Lugosi. Theory of classification: a survey of recent advances. *Preprint*, 2004. URL `http://www.econ.upf.es/~lugosi`.

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

E. Candes and D. Donoho. Curvelets and curvilinear integrals. *J. Approx. Theory*, 113:59–90, 2000.

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

H. Chernoff. A measure of asymptotic efficiency of tests of a hypothesis based on the sum of observations. *Annals of Mathematical Statistics*, 23:493–507, 1952.

P. Chou, T. Lookabaugh, and R. Gray. Optimal pruning with applications to tree-structured source coding and modeling. *IEEE Trans. Inform. Theory*, 35(2):299–315, 1989.

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

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

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

L. Devroye. Any discrimination rule can have an arbitrarily bad probability of error for finite sample size. *IEEE Trans. Patt. Anal. Mach. Intell.*, 4:154–157, 1982.

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

P. Domingos. MetaCost: A general method for making classifiers cost sensitive. In *Proceedings of the 5th International Conference on Knowledge Discovery and Data Mining*, pages 155–164, San Diego, CA, 1999. ACM Press.

D. Donoho. CART and best-ortho-basis selection: A connection. *Ann. Stat.*, 25: 1870–1911, 1997.

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

D. Donoho and I. Johnstone. Ideal adaptation via wavelet shrinkage. *Biometrika*, 81: 425–455, 1994.

D. Donoho and I. Johnstone. Adapting to unknown smoothness via wavelet shrinkage. *J. Amer. Statist. Assoc.*, 90:1200–1224, 1995.

D. L. Donoho, I. M. Johnstone, G. Kerkyacharian, and D. Picard. Wavelet shrinkage: Asymptopia? *J. Roy. Statist. Soc. B*, 57(432):301–369, 1995.

R. Dudley. Metric entropy of some classes of sets with differentiable boundaries. *J. Approximation Theory*, 10:227–236, 1974.

C. Elkan. The foundations of cost-sensitive learning. In *Proceedings of the 17th International Joint Conference on Artificial Intelligence*, pages 973–978, Seattle, Washington, USA, 2001.

F. Esposito, D. Malerba, and G. Semeraro. A comparitive analysis of methods for pruning decision trees. *IEEE Trans. Patt. Anal. Mach. Intell.*, 19(5):476–491, 1997.

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

M. Golea, P. Bartlett, W. S. Lee, and L. Mason. Generalization in decision trees and DNF: Does size matter? In S. Thrun, L. Saul, and B. Schölkopf, editors, *Advances in Neural Information Processing Systems 16*, Cambridge, MA, 2003. MIT Press.

L. Gordon and R. Olshen. Asymptotically efficient solutions to the classification problem. *Ann. Stat.*, 6(3):515–533, 1978.

T. Hagerup and C. Rüb. A guided tour of Chernoff bounds. *Inform. Process. Lett.*, 33(6):305–308, 1990.

C. Huber. Lower bounds for function estimation. In D. Pollard, E. Torgersen, and G. Yang, editors, *Festschrift for Lucien Le Cam*, pages 245–258. Springer-Verlag, 1997.

I. Johnstone. Wavelets and the theory of nonparametric function estimation. *Phil. Trans. Roy. Soc. Lond. A.*, 357:2475–2494, 1999.

M. Kearns and Y. Mansour. A fast, bottom-up decision tree pruning algorithm with near-optimal generalization. In Jude W. Shavlik, editor, *Proc. 15th Int. Conf. Machine Learning*, pages 269–277, Madison, WI, 1998. Morgan Kaufmann.

M. Kearns and Y. Mansour. On the boosting ability of top-down decision tree learning algorithms. *Journal of Computer and Systems Sciences*, 58(1):109–128, 1999.

E. Kolaczyk and R. Nowak. Multiscale generalized linear models for nonparametric function estimation. To appear in *Biometrika*, 91(4), December 2004a.

E. Kolaczyk and R. Nowak. Multiscale likelihood analysis and complexity penalized estimation. *Ann. Stat.*, 32(2):500–527, 2004b.

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

J. Langford. *Quantitatively Tight Sample Complexity Bounds*. PhD thesis, Carnegie Mellon University, 2002.

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

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

E. Mammen and A. B. Tsybakov. Smooth discrimination analysis. *Ann. Stat.*, 27: 1808–1829, 1999.

Y. Mansour. Pessimistic decision tree pruning. In D. H. Fisher, editor, *Proc. 14th Int. Conf. Machine Learning*, pages 195–201, Nashville, TN, 1997. Morgan Kaufmann.

Y. Mansour and D. McAllester. Generalization bounds for decision trees. In N. Cesa-Bianchi and S. Goldman, editors, *Proceedings of the Thirteenth Annual Conference on Computational Learning Theory*, pages 69–74, Palo Alto, CA, 2000.

D. Margineantu. Class probability estimation and cost-sensitive classification decisions. In *Proceedings of the 13th European Conference on Machine Learning*, pages 270–281, Helsinki, Finland, 2002.

J. S. Marron. Optimal rates of convergence to Bayes risk in nonparametric discrimination. *Ann. Stat.*, 11(4):1142–1155, 1983.

S. Murthy. Automatic construction of decision trees from data: A multi-disciplinary survey. *Data Mining and Knowledge Discovery*, 2(4):345–389, 1998.

A. Nobel. Analysis of a complexity based pruning scheme for classification trees. *IEEE Trans. Inform. Theory*, 48(8):2362–2368, 2002.

M. Okamoto. Some inequalities relating to the partial sum of binomial probabilites. *Annals of the Institute of Statistical Mathematics*, 10:29–35, 1958.

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

B. Ripley. *Pattern Recognition and Neural Networks*. Cambridge University Press, Cambridge, UK, 1996.

C. Schäfer, G. Blanchard, Y. Rozenholc, and K.-R. Müller. An algorithm for optimal dyadic tree classification. *Submitted to NIPS 2004*, 2004.

C. Scott and R. Nowak. Dyadic classification trees via structural risk minimization. In S. Becker, S. Thrun, and K. Obermayer, editors, *Advances in Neural Information Processing Systems 15*, Cambridge, MA, 2002. MIT Press.

C. Scott and R. Nowak. Near-minimax optimal classification with dyadic classification trees. In S. Thrun, L. Saul, and B. Schölkopf, editors, *Advances in Neural Information Processing Systems 16*, Cambridge, MA, 2003. MIT Press.

J. C. Scovel and I. Steinwart. Fast rates for support vector machines. Technical Report LA-UR 03-9117, Los Alamos National Laboratory, 2004.

H. Van Trees. *Detection, Estimation, and Modulation Theory: Part I*. John Wiley & Sons, New York, 2001.

A. B. Tsybakov. Optimal aggregation of classifiers in statistical learning. *Ann. Stat.*, 32(1):135–166, 2004.

A. B. Tsybakov and S. A. van de Geer. Square root penalty: adaptation to the margin in classification and in edge estimation. *Preprint*, 2004.

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

V. Vapnik and C. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and Its Applications*, 16(2): 264–280, 1971.

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

Y. Yang. Minimax nonparametric classification–Part I: Rates of convergence. *IEEE Trans. Inform. Theory*, 45(7):2271–2284, 1999.

B. Yu. Assouad, Fano, and Le Cam. In D. Pollard, E. Torgersen, and G. Yang, editors, *Festschrift for Lucien Le Cam*, pages 423–435. Springer-Verlag, 1997.

B. Zadrozny, J. Langford, and N. Abe. Cost sensitive learning by cost-proportionate example weighting. In *Proceedings of the 3rd International Conference on Data Mining*, Melbourne, FA, USA, 2003. IEEE Computer Society Press.

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