A Neyman–Pearson Approach to Statistical Learning

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

***Abstract*—The Neyman–Pearson (NP) approach to hypothesis testing is useful in situations where different types of error have different consequences or *a priori* probabilities are unknown. For any $\alpha > 0$, the NP lemma specifies the most powerful test of size α, but assumes the distributions for each hypothesis are known or (in some cases) the likelihood ratio is monotonic in an unknown parameter. This paper investigates an extension of NP theory to situations in which one has no knowledge of the underlying distributions except for a collection of independent and identically distributed (i.i.d.) *training examples* from each hypothesis. Building on a "fundamental lemma" of Cannon *et al.*, we demonstrate that several concepts from statistical learning theory have counterparts in the NP context. Specifically, we consider constrained versions of empirical risk minimization (NP-ERM) and structural risk minimization (NP-SRM), and prove performance guarantees for both. General conditions are given under which NP-SRM leads to strong universal consistency. We also apply NP-SRM to (dyadic) decision trees to derive rates of convergence. Finally, we present explicit algorithms to implement NP-SRM for histograms and dyadic decision trees.**

***Index Terms*—Generalization error bounds, Neyman–Pearson (NP) classification, statistical learning theory.**

I. Introduction

In most approaches to binary classification, classifiers are designed to minimize the probability of error. However, in many applications it is more important to avoid one kind of error than the other. Applications where this situation arises include fraud detection, spam filtering, machine monitoring, target recognition, and disease diagnosis, to name a few. In this paper, we investigate one approach to classification in this context inspired by classical Neyman–Pearson (NP) hypothesis testing.

In the spirit of statistical learning theory, we develop the theoretical foundations of an NP approach to learning classifiers from labeled training data. We show that several results and concepts from standard learning theory have counterparts in the NP setting. Specifically, we consider constrained versions of empirical risk minimization (NP-ERM) and structural risk minimization (NP-SRM), and prove performance guarantees for both. General conditions are given under which NP-SRM leads to strong universal consistency. Here consistency entails the (almost sure) convergence of both the learned miss and false alarm probabilities to the optimal probabilities given by the NP lemma. We also apply NP-SRM to (dyadic) decision trees to derive rates of convergence. Finally, we present explicit algorithms to implement NP-SRM for histograms and dyadic decision trees.

A. Motivation

In the NP theory of binary hypothesis testing, one must decide between a *null hypothesis* and an *alternative hypothesis*. A level of significance α (called the *size* of the test) is imposed on the false alarm (type I error) probability, and one seeks a test that satisfies this constraint while minimizing the miss (type II error) probability, or equivalently, maximizing the detection probability (power). The NP lemma specifies necessary and sufficient conditions for the most powerful test of size α, provided the distributions under the two hypotheses are known, or (in special cases) the likelihood ratio is a monotonic function of an unknown parameter [1]–[3] (see [4] for an interesting overview of the history and philosophy of NP testing). We are interested in extending the NP paradigm to the situation where one has no knowledge of the underlying distributions except for independent and identically distributed (i.i.d.) training examples drawn from each hypothesis. We use the language of classification, whereby a test is a classifier and each hypothesis corresponds to a particular class.

To motivate the NP approach to learning classifiers, consider the problem of classifying tumors using gene expression microarrays [5], which are typically used as follows: First, identify several patients whose status for a particular form of cancer is known. Next, collect cell samples from the appropriate tissue in each patient. Then, conduct a microarray experiment to assess the relative abundance of various gene transcripts in each of the subjects. Finally, use this "training data" to build a classifier that can, in principle, be used to diagnose future patients based on their gene expression profiles.

The dominant approach to classifier design in microarray studies has been to minimize the probability of error (see, for example, [6] and references therein). Yet it is clear that failing to detect a malignant tumor has drastically different consequences than erroneously flagging a benign tumor. In the NP approach, the diagnostic procedure involves setting a level of significance α, an upper bound on the fraction of healthy patients that may be unnecessarily sent for treatment or further screening, and constructing a classifier to minimize the number of missed true cancer cases.[1]

Further motivation for the NP paradigm comes from a comparison with cost-sensitive classification (CSC). CSC (also called cost-sensitive learning) is another approach to handling

Manuscript received September 19, 2004; revised February 2, 2005. This work was supported by the National Science Foundation under Grants CCR-0310889 and CCR-0325571, and by the Office of Naval Research under Grant N00014-00-1-0966.

C. Scott is with the Department of Statistics, Rice University, Houston, TX 77005 USA (e-mail: cscott@rice.edu).

R. Nowak is with the Department of Electrical and Computer Engineering, University of Wisconsin, Madison, WI 53706 USA (e-mail: nowak@engr.wisc.edu).

Communicated by P. L. Bartlett, Associate Editor for Pattern Recognition, Statistical Learning and Inference.

Digital Object Identifier 10.1109/TIT.2005.856955

[1] Obviously, the two classes may be switched if desired.

disparate kinds of errors in classification (see [7]–[10] and references therein). Following classical Bayesian decision theory, CSC modifies the standard "0–1" loss function to a weighted Bayes cost. Cost-sensitive classifiers assume the relative costs for different classes are known. Moreover, these algorithms assume estimates (either implicit or explicit) of the *a priori* class probabilities can be obtained from the training data or some other source.

CSC and NP classification 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? The two approaches are similar in that the user essentially has one free parameter to set. In CSC, this free parameter is the ratio of costs of the two class errors, while in NP classification it is the false alarm threshold α. The selection of costs does not directly provide control on α, and conversely setting α does not directly translate into costs. The lack of precise knowledge of the underlying distributions makes it impossible to precisely relate costs with α. The choice of method will thus depend on which parameter can be set more realistically, which in turn depends on the particular application. For example, consider a network intrusion detector that monitors network activity and flags events that warrant a more careful inspection. The value of α may be set so that the number of events that are flagged for further scrutiny matches the resources available for processing and analyzing them.

From another perspective, however, the NP approach seems to have a clear advantage with respect to CSC. Namely, NP classification does not assume knowledge of or about *a priori* class probabilities. CSC can be misleading if *a priori* class probabilities are not accurately 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. This situation arises, for example, in machine fault detection, where machines must often be induced to fail (usually at high cost) to gather the necessary training data. Here the fraction of "faulty" training examples is not indicative of the true probability of a fault occurring. In fact, it could be argued that in most applications of interest, class probabilities differ between training and test data. Returning to the cancer diagnosis example, class frequencies of diseased and normal patients at a cancer research institution, where training data is likely to be gathered, in all likelihood do not reflect the disease frequency in the population at large.

B. *Previous Work on NP Classification*

Although NP classification has been studied previously from an empirical perspective [11], the theoretical underpinnings were apparently first studied by Cannon, Howse, Hush, and Scovel [12]. They give an analysis 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 substantially tighter. Second, while Cannon *et al.* consider only learning from fixed Vapnik–Chervonenkis (VC) classes, we introduce a constrained form of NP-SRM that automatically balances model complexity and training error, and gives rise to strongly universally consistent rules. Third, assuming mild regularity conditions on the underlying distribution, we derive rates of convergence for NP-SRM as realized by a certain family of decision trees called dyadic decision trees. Finally, we present exact and computationally efficient algorithms for implementing NP-SRM for histograms and dyadic decision trees.

In a separate paper, Cannon *et al.* [13] consider NP-ERM over a data-dependent collection of classifiers and are able to bound the estimation error in this case as well. They also provide an algorithm for NP-ERM in some simple cases involving linear and spherical classifiers, and present an experimental evaluation. To our knowledge, the present work is the third study to consider an NP approach to statistical learning, the second to consider practical algorithms with performance guarantees, and the first to consider model selection, consistency and rates of convergence.

C. *Notation*

We focus exclusively on *binary* classification, although extensions to multiclass 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 (pattern, feature vector) and Y is the class label associated with X. In classical NP testing, $Y = 0$ corresponds to the null hypothesis.

A *classifier* is a Borel measurable function $h : \mathcal{X} \to \{0, 1\}$ mapping signals to class labels. In *standard classification*, the performance of h is measured by the probability of 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

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

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. The false-alarm probability is also known as the *size*, while one minus the miss probability is known as the *power*.

Let $\alpha \in [0, 1]$ be a user-specified level of significance or false alarm threshold. In NP testing, one seeks the classifier g^* minimizing $R_1(h)$ over all h such that $R_0(h) \leq \alpha$. In words, g^* is the most powerful test (classifier) of size α. If f_0 and f_1 are the conditional densities of X (with respect to a measure ν) corresponding to classes 0 and 1, respectively, then the NP lemma [1] states that $g^*(x) = \mathbb{I}_{\{\Lambda(x) > \eta\}}$. Here $\mathbb{I}$ denotes the indicator function, $\Lambda(x) = f_1(x)/f_0(x)$ is the likelihood ratio, and η is as small as possible such that $\int_{\Lambda > \eta} f_0 d\nu \leq \alpha$. Thus, when f_0 and f_1 are known (or in certain special cases where the likelihood ratio is a monotonic function of an unknown parameter) the NP lemma delivers the optimal classifier.[2]

[2] In certain cases, such as when X is discrete, one can introduce *randomized tests* to achieve slightly higher power. Otherwise, the false alarm constraint may not be satisfied with equality. In this paper, we do not explicitly treat randomized tests. Thus, h^* and g^* should be understood as being optimal with respect to classes of nonrandomized tests/classifiers.

In this paper, we are interested in the case where our only information about f_j is a finite training sample. Let $Z^n = \{(X_i, Y_i)\}_{i=1}^n \in \mathcal{Z}^n$ be a collection of n i.i.d. samples of $Z = (X, Y)$. A learning algorithm (or *learner* for short) is a mapping $\widehat{h}_n : \mathcal{Z}^n \to \mathcal{H}(\mathcal{X})$, where $\mathcal{H}(\mathcal{X})$ is the set of all classifiers. In words, the learner $\widehat{h}_n$ is a rule for selecting a classifier based on a training sample. When a training sample Z^n is given, we also use $\widehat{h}_n$ to denote the classifier produced by the learner.

In classical NP theory it is possible to impose absolute bounds on the false alarm and miss probabilities. When learning from training data, however, one cannot make such guarantees because of the dependence on the random training sample. Unfavorable training samples, however unlikely, can lead to arbitrarily poor performance. Therefore, bounds on R_0 and R_1 can only be shown to hold with high probability or in expectation. Formally, let $\mathbb{P}^n$ denote the product measure on $\mathcal{Z}^n$ induced by $\mathbb{P}$, and let $\mathbb{E}^n$ denote expectation with respect to $\mathbb{P}^n$. Bounds on the false alarm and miss probabilities must be stated in terms of $\mathbb{P}^n$ and $\mathbb{E}^n$ (see Section I-D).

We investigate classifiers defined in terms of 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

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

denote the empirical false alarm and miss probabilities, corresponding to $j = 0$ and $j = 1$, respectively. Given a class of classifiers $\mathcal{H}$, define

$$\mathcal{H}_0 = \{h \in \mathcal{H} : R_0(h) \leq \alpha\}$$

and

$$h^* = \arg\min\{R_1(h) : h \in \mathcal{H}_0\}.$$

That is, h^* is the most powerful test/classifier[3] in $\mathcal{H}$ of size α. Finally, set $R_1^* = R_1(g^*)$ to be the miss probability of the optimal classifier g^* provided by the NP lemma.

D. Problem Statement

The goal of NP learning is to design learning algorithms $\widehat{h}_n$ producing classifiers that perform almost as well as h^* or g^*. In particular, we desire learners with the following properties. (This section is intended as a preview; precise statement are given later.)

PAC bounds: $\widehat{h}_n$ is "probably approximately correct" (PAC) in the sense that given $\delta_0, \delta_1 > 0$, there exist $\epsilon_0 = \epsilon_0(n_0, \delta_0)$ and $\epsilon_1 = \epsilon_1(n_1, \delta_1)$ such that for any n

$$\mathbb{P}^n(R_0(\widehat{h}_n) - \alpha > \epsilon_0(n_0, \delta_0)) \leq \delta_0$$

and

$$\mathbb{P}^n\left(R_1(\widehat{h}_n) - R_1(h^*) > \epsilon_1(n_1, \delta_1)\right) \leq \delta_1.$$

Moreover, δ_0, δ_1 decay exponentially fast as functions of increasing ϵ_0, ϵ_1 (see Section II for details).

False Alarm Probability Constraints: In classical hypothesis testing, one is interested in tests/classifiers h satisfying $R_0(h) \leq \alpha$. In the learning setting, such constraints can only hold with a certain probability. It is however possible to obtain nonprobabilistic guarantees of the form (see Section II-C for details)

$$\mathbb{E}^n\{R_0(\widehat{h}_n)|n_0\} \leq \alpha'.$$

Oracle inequalities: Given a hierarchy of sets of classifiers $\mathcal{H}_1 \subset \cdots \subset \mathcal{H}_K$, $\widehat{h}_n$ does about as well as an oracle that knows which $\mathcal{H}_k$ achieves the proper balance between the estimation and approximation errors. In particular we will show that with high probability, both

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

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 III for details).

Consistency: If $\mathcal{H}$ grows (as a function of n) in a suitable way, then $\widehat{h}_n$ is strongly universally consistent [14, Ch. 6] in the sense that

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

and

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

for all distributions of Z (see Section IV for details).

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

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

and

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

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 $b_n \preccurlyeq a_n$ (see Section V for details).

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

II. Neyman–Pearson and Empirical Risk Minimization

In this section, we review the work of Cannon *et al.* [12] who study NP learning in the context of fixed VC classes. We also apply a different bounding technique that leads to substantially tighter upper bounds. For a review of VC theory see Devroye, Györfi, and Lugosi [14].

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

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

[3] In this paper, we assume that a classifier h^* achieving the minimum exists. Although not necessary, this allows us to avoid laborious approximation arguments.

We call this procedure NP-ERM. Cannon *et al.* demonstrate that NP-ERM enjoys properties similar to standard ERM [14], [15] translated to the NP context. We now recall their analysis.

To state the theoretical properties of NP-ERM, introduce the following notation.[4] Let $\epsilon_1 > 0$. Recall

$$h^* = \arg\min\{R_1(h) : h \in \mathcal{H}_0\}.$$

Define

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

A main result of Cannon *et al.* [12] is the following lemma.

Lemma 1 ([12]): With Θ_j and Ω_j defined as above and $\widehat{h}_n$ as defined in (1) we have

$$\begin{aligned}\Theta_0 &\subset \Omega_0\\ \Theta_1 &\subset \Omega_0 \cup \Omega_1\end{aligned}$$

and in particular

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

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. Vapnik and Chervonenkis introduced an analogous result to bound the performance of ERM (for standard classification) in terms of the error deviance [16] (see also [14, Ch. 8]). An immediate corollary is the following.

Proposition 1 ([12]): Let $\epsilon_1, \epsilon_0 > 0$ and take $\widehat{h}_n$ as in (1). Then for any n

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

Later, this result is used to derive PAC bounds by applying results for convergence of empirical processes such as VC inequalities.

We make an important observation that is not mentioned by Cannon *et al.* [12]. 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 bound and extension to SRM. In particular, we will choose ϵ_j to depend on n_j, a specified 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).

While we focus our discussion on VC and finite classes for concreteness and comparison with [12], other classes and bounding techniques are applicable. Proposition 1 allows for the use of many of the error deviance bounds that have appeared in the empirical process and machine learning literature in recent years. The tolerances ϵ_j may even depend on the full sample Z^n or on the individual classifier. Thus, for example, Rademacher averages [17], [18] could be used to define the tolerances in NP-ERM. However, the fact that the tolerances for VC and finite classes (defined below in (2) and (3)) are independent of the classifier, and depend on the sample only through n_0 and n_1, does simplify our extension to SRM in Section III.

A. NP-ERM With VC Classes

Suppose $\mathcal{H}$ has VC dimension $V < \infty$. Cannon *et al.* consider two viewpoints for NP classification. First they consider *retrospective sampling* where n_0 and n_1 are known before the sample is observed. Applying the VC inequality as stated by Devroye *et al.* [14], together with Proposition 1, they obtain the following.

Theorem 1 ([12]): Let $\epsilon_0, \epsilon_1 > 0$ and take $\widehat{h}_n$ as in (1). Then for any n

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

An alternative viewpoint for NP classification is *i.i.d. sampling* in which n_0 and n_1 are unknown until the training sample is observed. 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 using the fact that with high probability n_0 and n_1 are concentrated near their expected values. Recall that π_j is the *a priori* probability of class j.

Theorem 2 ([12]): Let $\epsilon_0, \epsilon_1 > 0$ and take $\widehat{h}_n$ as in (1). If $n \ge \frac{10\sqrt{5}}{\pi_j^2\epsilon_j^2}$, $j = 0, 1$, then

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

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 VC-based bound for NP classification under i.i.d. sampling that improves upon the preceding result. In particular, our bound is as tight as the bound in Theorem 1 for retrospective sampling and it holds for all values of n. Thus, it is no longer necessary to be concerned with the philosophical differences between retrospective and i.i.d. sampling,

[4] We interchange the meanings of the subscripts 0 and 1 used in [12], preferring to associate class 0 with the null hypothesis.

(ϵ, δ)	(0.2, 0.1)	(0.1, 0.1)	(0.1, 0.05)	(0.05,0.1)	(0.001, 0.001)	$(10^{-5}, 10^{-5})$
$\pi_0 = 0.5$	2.97×10^1	2.90×10^1	2.87×10^1	2.85×10^1	2.59×10^1	2.47×10^1
$\pi_0 = 0.9$	1.68×10^2	1.63×10^2	1.61×10^2	1.59×10^2	1.39×10^2	1.30×10^2
$\pi_0 = 0.99$	2.04×10^3	1.95×10^3	1.91×10^3	1.88×10^3	1.56×10^3	1.42×10^3
$\pi_0 = 0.999$	2.29×10^4	2.27×10^4	2.22×10^4	2.17×10^4	1.73×10^4	1.53×10^4

Fig. 1. Assuming $\delta = \delta_0 = \delta_1$ and $\varepsilon = \varepsilon_0 = \varepsilon_1$, this table reports, for various values of δ, ε, and π_0, the ratio of the sample sizes needed to satisfy the two bounds of Theorems 2 and 3, respectively. Clearly, the bound of Theorem 3 is tighter, especially for asymmetric *a priori* probabilities. Furthermore, the bound of Theorem 2 requires knowledge of π_0, which is typically not available in the NP setting, while the bound of Theorem 3 does not. See Example 1 for further discussion.

since the same bound is available for both. As mentioned previously, the key idea is to let the tolerances ϵ_0 and ϵ_1 be variable as follows. Let $\delta_0, \delta_1 > 0$ and define

$$\epsilon_j = \epsilon_j(n_j, \delta_j, \mathcal{H}) = \sqrt{128 \frac{V \log n_j + \log(8/\delta_j)}{n_j}} \tag{2}$$

for $j = 0, 1$. Let $\widehat{h}_n$ be defined by (1) as before, but with the new definition of ϵ_0 (which now depends on n_0, δ_0, and $\mathcal{H}$).

The main difference between the rule of Cannon *et al.* and the rule proposed here is that in their formulation the term $\alpha + \frac{1}{2}\epsilon_0$ constraining the empirical false alarm probability is independent of the sample. In contrast, our constraint 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. Intuitively, when n_0 is larger, $\widehat{R}_0$ should more accurately estimate R_0, and therefore a suitable classifier should be available from among those rules approximating α to within the smaller tolerance. Theoretically, our approach leads to a substantially tighter bound as the following theorem shows.

Theorem 3: For NP-ERM over a VC class $\mathcal{H}$ with tolerances given by (2), and for any n

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

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

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

Without loss of generality take $j = 0$. Let $\mathbb{P}(n_0)$ be the probability that there are n_0 examples of class 0 in the training sample. Then

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

where the inequality follows from the VC inequality [14, Ch. 12]. This completes the proof. □

The new bound is substantially tighter than that of Theorem 2, as illustrated with the following example.

Example 1: 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 0.8 probability (say $\delta_0 = \delta_1 = 0.1$), both bounds hold with $\epsilon_0 = \epsilon_1 = 0.2$ For the new bound of Theorem 3 to hold we need

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

which implies $n \ge 97104$. In contrast, Theorem 2 requires

$$20ne^{-\frac{n(0.5)^2(0.2)^2}{640\sqrt{5}}} \le 0.1$$

which implies $n \ge 2887079$. To verify that this phenomenon is not a consequence of the particular choice of δ_j, ϵ_j, or π_0, Fig. 1 reports the ratios of the necessary sample sizes for a range of these parameters.[5] Indeed, as π_0 tends to 1, the disparity grows significantly. Finally, we note that the bound of Cannon *et al.* for retrospective sampling requires as many samples as our bound for i.i.d. sampling.

B. NP-ERM With Finite Classes

The VC inequality is so general that in most practical settings it is too loose owing to large constants. Much tighter bounds are possible when $\mathcal{H}$ is finite. For example, $\mathcal{H}$ could be obtained by quantizing the 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

$$\epsilon_j = \epsilon_j(n_j, \delta_j, \mathcal{H}) = \sqrt{2 \frac{\log|\mathcal{H}| + \log(2/\delta_j)}{n_j}}. \tag{3}$$

We have the following analog of Theorem 3.

Theorem 4: For NP-ERM over a finite class $\mathcal{H}$ with tolerances given by (3)

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

The proof is identical to the proof of Theorem 3 except that the VC inequality is replaced by a bound derived from Hoeffding's inequality and the union bound (see [14, Ch. 8]).

Example 2: To illustrate the significance of the new bound, consider the scenario described in Example 1, but assume the

[5] When calculating (2) for this example we assume $n_j = n \cdot \pi_j$.

VC class $\mathcal{H}$ is quantized so that $|\mathcal{H}| = 2^{16}$. For the bound of Theorem 4 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 VI-B where we apply NP-SRM to dyadic decision trees.

C. Learner False Alarm Probabilities

In the classical setting, one is interested in classifiers h satisfying $R_0(h) \leq \alpha$. When learning from training data, such bounds on the false alarm probability of $\widehat{h}_n$ are not possible due to its dependence on the training sample. However, a practitioner may still wish to have an absolute upper bound of some sort. One possibility is to bound the quantity

$$\mathbb{E}^n\{R_0(\widehat{h}_n)|n_0\}$$

which we call the *learner false alarm probability*.

The learner false alarm probability can be constrained for a certain range of α depending on the class $\mathcal{H}$ and the number n_0 of class 0 training samples. Recall

$$\Theta_0 = \{Z^n : R_0(\widehat{h}_n) > \alpha + \epsilon_0\}.$$

Arguing as in the proof of Theorem 3 and using Lemma 1 ($\Theta_0 \subseteq \Omega_0$), we have

$$\begin{aligned}\mathbb{E}^n\{R_0(\widehat{h}_n)|n_0\} &= \mathbb{P}^n(\Theta_0|n_0)\mathbb{E}^n\{R_0(\widehat{h}_n)|\Theta_0, n_0\}\\ &\quad + \mathbb{P}^n(\overline{\Theta_0}|n_0)\mathbb{E}^n\{R_0(\widehat{h}_n)|\overline{\Theta_0}, n_0\}\\ &\leq \alpha + \epsilon_0(n_0, \delta_0, \mathcal{H}) + \delta_0.\end{aligned}$$

The confidence δ_0 is essentially a free parameter, so let α_0 be the minimum possible value of $\epsilon_0(n_0, \delta_0, \mathcal{H}) + \delta_0$ as δ_0 ranges over $(0, 1]$. Then the learner false alarm probability can be bounded by any desired α', $\alpha_0 \leq \alpha' \leq 1$, by choosing α appropriately. In particular, if $\widehat{h}_n$ is obtained by NP-ERM with $\alpha = \alpha' - \alpha_0$, then

$$\mathbb{E}^n\{R_0(\widehat{h}_n)|n_0\} \leq \alpha'.$$

Example 3: Suppose $\mathcal{H}$ is finite, $|\mathcal{H}| = 2^8$, and $n_0 = 1000$. A simple numerical experiment, using the formula of (3) for ϵ_0, shows that the minimum value of $\delta_0 + \epsilon_0(n_0, \delta_0, \mathcal{H})$ is $\alpha_0 = 0.157$. Thus, if a bound of $\alpha' = 0.2$ on the learner false alarm probability is desired, it suffices to perform NP-ERM with $\alpha = \alpha' - \alpha_0 = 0.043$.

III. 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) [19], whereby a classifier is selected from a family $\mathcal{H}^k$, $k = 1, 2, \ldots$, of increasingly rich classes of classifiers. In this section we present NP-SRM, a version of SRM adapted to the NP setting, in the two cases where all $\mathcal{H}^k$ are either VC classes or finite.

A. NP-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

$$\epsilon_j = \epsilon_j(n_j, \delta_j, k) = \sqrt{128\frac{V_k \log n_j + k\log 2 + \log(8/\delta_j)}{n_j}}. \tag{4}$$

Remark 1: 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 simultaneously with probability at least $1 - \delta_j$ (see the proof of Theorem 5).

NP-SRM produces a classifier $\widehat{h}_n$ 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

$$\begin{aligned}\widehat{h}_n^k = \arg\min_{h\in\mathcal{H}^k} &\ \widehat{R}_1(h)\\ \text{s.t.} &\ \widehat{R}_0(h) \leq \alpha + \frac{1}{2}\epsilon_0(n_0, \delta_0, k).\end{aligned} \tag{5}$$

2) Set

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

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

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

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

We have the following oracle bound for NP-SRM. The proof is given in Appendix I. For a similar result in the context of standard classification see Lugosi and Zeger [20].

Theorem 5: For any n, with probability at least $1 - (\delta_0 + \delta_1)$ over the training sample Z^n, both

$$R_0(\widehat{h}_n) - \alpha \leq \epsilon_0(n_0, \delta_0, K(n)) \tag{6}$$

and

$$\begin{aligned}R_1(\widehat{h}_n) - R_1^*&\\ \leq \min_{1\leq k\leq K(n)} &\left(\epsilon_1(n_1, \delta_1, k) + \inf_{h\in\mathcal{H}_0^k} R_1(h) - R_1^*\right)\end{aligned} \tag{7}$$

hold.

The inequality in (6) implies that the "excess false alarm probability" decays as $O(\sqrt{V_{K(n)} \log n/n})$. Moreover, this rate can be designed through the choice of $K(n)$ and requires no assumption about the underlying distribution.

To interpret the second inequality, observe that for any k

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

The two terms on the right are referred to as *estimation error* and *approximation error*,[6] respectively. If k is such that $\widehat{h}_n = \widehat{h}_n^k$, then Theorem 3 implies that the estimation error is bounded above by $\epsilon_1(n_1, \delta_1, k)$. Thus, (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 $\widehat{h}_n$.

Oracle inequalities are important because they indicate the ability of learning rules to *adapt* to unknown properties of the data generating distribution. Unfortunately, we have no oracle inequality for the false alarm error. Perhaps this is because, while the miss probability involves both an estimation and approximation error, the false alarm probability has only a stochastic component (no approximation is required). In other words, oracle inequalities typically reflect the learner's ability to strike a balance between estimation and approximation errors, but in the case of the false alarm probability, there is nothing to balance. For further discussion of oracle inequalities for standard classification see [18], [21].

B. NP-SRM Over Finite Classes

The developments of the preceding subsection 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 now with penalties

$$\epsilon_j = \epsilon_j(n_j, \delta_j, k) = \sqrt{2 \frac{\log |\mathcal{H}^k| + k \log 2 + \log(2/\delta_j)}{n_j}}. \quad (8)$$

Theorem 5 holds in this setting as well. The proof is an easy modification of the proof of that theorem, substituting Theorem 4 for Theorem 3, and is omitted.

IV. 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 calibrated appropriately.

Theorem 6: Let $\widehat{h}_n$ be the classifier given by (5), with $\epsilon_0(n_0, \delta_0, k)$ defined by (4) for NP-SRM over VC classes, or (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.,

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

and

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

for all distributions.

The proof is given in Appendix II. Note that the conditions on δ in the theorem hold if $\delta_j(n) \asymp n^{-\beta}$ for some $\beta > 1$.

Example 4: 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 $K(n) \to \infty$ and $K(n)^d = o(n)$, in analogy to the requirement for strong universal consistency of the regular histogram rule for standard classification (see [14, Ch. 9]). Moreover, NP-SRM with histograms can be implemented efficiently in $O(n n_0 K)$ operations as described in Appendix IV.

V. Rates of Convergence

In this section, we examine rates of convergence to zero for the expected [7] excess false alarm probability

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

and expected excess miss probability

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

Moreover, we are interested in rates that hold independent of α.

A. Rates for False Alarm Error

The rate for false alarm error can be specified by the choice $K(n)$. In the case of NP-SRM over VC classes we have the following result. A similar result holds for NP-SRM over finite classes (but without the logarithmic terms).

Proposition 2: Select $K(n)$ such that $V_{K(n)} \asymp (n/\log n)^\tau$ for some τ, $0 < \tau < 1$. Under assumptions 3) of Theorem 7, $\widehat{h}_n$ satisfies

$$\mathbb{E}^n R_0(\widehat{h}_n) - \alpha \preccurlyeq \left(\frac{\log n}{n} \right)^{\frac{1-\tau}{2}}.$$

The proof follows exactly the same lines as the proof of Theorem 7 below, and is omitted.

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

[7] 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.

B. Rates for Both Errors

A more challenging problem is establishing rates for both errors simultaneously. Several recent studies 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 [22]–[26]. Observe that

$$R(\widehat{\phi}_n) - R(\phi^*) = \pi_0 \left(R_0(\widehat{\phi}_n) - R_0(\phi^*) \right) + \pi_1 \left(R_1(\widehat{\phi}_n) - R_1(\phi^*) \right).$$

Hence, rates of convergence for NP classification with $\alpha = R_0(\phi^*)$ imply rates of convergence for standard classification. We summarize this observation as follows.

Proposition 3: Fix a distribution of the data Z. Let $\widehat{h}_n$ be a classifier, and let $r_j(n)$, $j = 0, 1$, be nonnegative 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 [27] has shown that for any classifier $\widehat{\phi}_n$ 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 3, the same must be true of NP learning. Thus, let $\mathcal{D}$ be some class of distributions. Proposition 3 also informs us about lower bounds for learning from distributions in $\mathcal{D}$.

Proposition 4: Assume that a learner for standard classification satisfies the minimax lower bound

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

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 standard classification translate to minimax lower bounds for NP classification.

VI. Rates for Dyadic Decision Trees

In this section, we provide an example of how to derive rates of convergence using NP-SRM combined with an appropriate analysis of the approximation error. We consider a special family of decision trees known as dyadic decision trees (DDTs) [28]. Before introducing DDTs, however, we first introduce the class of distributions $\mathcal{D}$ with which our study is concerned.

A. The Box-Counting Class

From this point on assume $\mathcal{X} = [0,1]^d$. Before introducing $\mathcal{D}$ we need some additional notation. Let m denote a positive integer, and define $\mathcal{P}_m$ to be the collection of m^d cells formed by the regular partition of $[0,1]^d$ into hypercubes of sidelength $1/m$. Let $c_0, c_1 > 0$ 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 the following assumptions.

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[8] 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 [28] for further discussion. A theorem of Tsybakov [22] implies that the minimax rate for standard classification for this class is $n^{-1/d}$ when $d \geq 2$[28]. By Proposition 4, both errors cannot simultaneously decay faster than this rate. In the following, we prove that this lower bound is almost attained using dyadic decision trees and NP-SRM.

B. Dyadic Decision Trees

Scott and Nowak [23], [28] demonstrate that a certain family of decision trees, DDTs, offer a computationally feasible classifier that also achieves optimal rates of convergence (for standard classification) under a wide range of conditions [23], [29], [30]. DDT's 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 [31]–[33]. We now show that an analysis similar to that of Scott and Nowak [29] applies to NP-SRM for DDTs, leading to similar results: rates of convergence for a computationally efficient learning algorithm.

A dyadic decision tree is a decision tree that divides the input space by means of axis-orthogonal dyadic splits. More precisely, a dyadic decision tree T is a binary tree (with a distinguished root node) specified by assigning 1) an integer $s(v) \in \{1, \ldots, d\}$ to each internal node v of T (corresponding to the coordinate that gets split at that node); 2) a binary label 0 or 1 to each leaf node of T. The nodes of DDTs correspond to hyperrectangles (cells) in $[0,1]^d$. Given a hyperrectangle $A = \prod_{r=1}^d [a_r, b_r]$, 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

[8] When proving rates for standard classification, it is often necessary to place a similar restriction on the *unconditional* density $f(x)$ of X. Here it is only necessary to bound $f_1(x)$ because only the excess *miss* probability requires an analysis of approximation error.

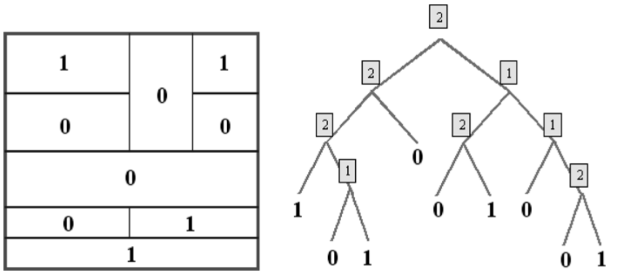

Fig. 2. A dyadic decision tree (right) with the associated recursive dyadic partition (left) in $d = 2$ dimensions. Each internal node of the tree is labeled with an integer from 1 to d indicating the coordinate being split at that node. The leaf nodes are decorated with class labels.

cell A, then the children of v are associated with $A^{s(v),1}$ and $A^{s(v),2}$. See Fig. 2.

Let $L = L(n)$ be a nonnegative integer and define $\mathcal{T}_L$ to be the collection of all DDT's such that no leaf cell has a side length smaller than 2^{-L}. In other words, when traversing a path from the root to a leaf, 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.

C. NP-SRM With DDTs

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 (4) or (8). 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[9] $C_k = (k+1)^{-1}\binom{2k}{k}$. 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 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 (8) for finite classes, it behaves like $\sqrt{2k(\log 8 + \log d)/n_j}$, while the penalty defined by (4) for VC classes behaves like $\sqrt{128k\log(n_j)/n_j}$. Therefore, we adopt the penalties for finite classes because they lead to bounds having smaller constants and lacking an additional log term.

By applying NP-SRM to DDTs[10] with parameters $L(n)$, $K(n)$, and $\delta_j(n)$ chosen appropriately, we obtain the following result. Note that the condition on $\delta_j(n)$ in the following theorem holds whenever $\delta_j(n) \asymp n^{-\beta}$, $\beta > 1/2$. The proof is found in Appendix III.

Theorem 7: Let $\widehat{h}_n$ be the classifier given by (5), with $\epsilon_0(n_0, \delta_0, k)$ defined by (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 $d \geq 2$ then

$$\sup_{\mathbf{A0},\mathbf{A1}} \left[\mathbb{E}^n R_0(\widehat{h}_n) - \alpha\right] \preccurlyeq n^{-1/(d+1)}$$

and

$$\sup_{\mathbf{A0},\mathbf{A1}} \left[\mathbb{E}^n R_1(\widehat{h}_n) - R_1^*\right] \preccurlyeq n^{-1/(d+1)}$$

where the sup is over all distributions belonging to the box-counting class.

We note in particular that the constants c_0 and c_1 in the definition of the box-counting class need not be known.

D. Optimal Rates for the Box-Counting Class

The rates given in the previous theorem do not match the lower bound of $n^{-1/d}$ mentioned in Section VI-A. At this point, one may as two questions: 1) Is the lower bound or the upper bound loose? and 2) If it is the upper bound, is suboptimality due to the use of DDTs or is it inherent in the NP-SRM learning procedure? It turns out that the lower bound is tight, and suboptimality stems from the NP-SRM learning procedure. It is possible to obtain optimal rates (to within a logarithmic factor) using DDTs and an alternate penalization scheme.[11]

A similar phenomenon appears in the context of standard classification. In [29], we show that DDTs and standard SRM yield suboptimal rates like those in Theorem 7 for standard classification. Subsequently, we were able to obtain the optimal rate with DDTs using a spatially adaptive penalty (which favors unbalanced trees) and a penalized empirical risk procedure [30], [23]. A similar modification works here. That same spatially adaptive penalty may be used to obtain optimal rates for NP classification. Thus, NP-SRM is suboptimal because it does not promote the learning of an unbalanced tree, which is the kind of tree we would expect to accurately approximate a member of the box-counting class. For further discussion of the importance of spatial adaptivity in classification see [28].

E. Implementing Dyadic Decision Trees

The importance of DDTs stems not only from their theoretical properties but also from the fact that NP-SRM may be implemented exactly in polynomial time. In Appendix V, we provide an explicit algorithm to accomplish this task. The algorithm is inspired by the work of Blanchard, Schäfer, and Rozenholc [34] who extend an algorithm of Donoho [35] to perform standard penalized ERM for DDTs.

VII. Conclusion

We have extended several results for learning classifiers from training data to the NP setting. Familiar concepts such as empirical and SRM have counterparts with analogous performance guarantees. Under mild assumptions on the hierarchy of classes, NP-SRM allows one to deduce strong universal consistency and rates of convergence. We have examined rates for DDTs, and presented algorithms for implementing NP-SRM with both histograms and DDTs.

This work should be viewed as an initial step in translating the ever growing field of supervised learning for classification to the NP setting. An important next step is to evaluate the potential impact of NP classification in practical settings where different

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

[10] Since $L(n)$ changes with n, the classes $\mathcal{H}^k$ are not independent of n as they are in the development of Section III. However, a quick inspection of the proofs of Theorems 5 and 6 reveals that those theorems also hold in this slightly more general setting.

[11] We do not present this refined analysis here because it is somewhat specialized to DDTs and would require a substantial amount of additional space, detracting from the focus of the paper.

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 theoretical properties of such NP algorithms should resemble those of their more familiar counterparts.

APPENDIX I
PROOF OF THEOREM 5

Define the sets

$$\Theta_0 = \left\{ Z^n : R_0(\widehat{h}_n) - \alpha > \epsilon_0(n_0, \delta_0, K(n)) \right\}$$

$$\Theta_1 = \Bigg\{ Z^n : R_1(\widehat{h}_n) - R_1^* > \inf_{1 \le k \le K(n)} \left(\epsilon_1(n_1, \delta_1, k) + \inf_{h \in \mathcal{H}_0^k} R_1(h) - R_1^* \right) \Bigg\}$$

$$\Omega_0^k = \left\{ Z^n : \sup_{h \in \mathcal{H}^k} |R_0(h) - \widehat{R}_0(h)| > \frac{1}{2}\epsilon_0(n_0, \delta_0, k) \right\}$$

$$\Omega_1^k = \left\{ Z^n : \sup_{h \in \mathcal{H}^k} |R_1(h) - \widehat{R}_1(h)| > \frac{1}{2}\epsilon_1(n_1, \delta_1, k) \right\}$$

$$\Theta_0^k = \left\{ Z^n : R_0\left(\widehat{h}_n^k\right) - \alpha > \epsilon_0(n_0, \delta_0, k) \right\}$$

$$\Theta_1^k = \left\{ Z^n : R_1\left(\widehat{h}_n^k\right) - \inf_{h \in \mathcal{H}_0^k} R_1(h) > \epsilon_1(n_1, \delta_1, k) \right\}.$$

Our goal is to show

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

Lemma 2:

$$\Theta_0 \cup \Theta_1 \subset \bigcup_{k=1}^{\infty} \left(\Omega_0^k \cup \Omega_1^k \right).$$

Proof: We show the contrapositive

$$\cap_{k=1}^{\infty} \left(\overline{\Omega_0^k} \cap \overline{\Omega_1^k} \right) \subset \overline{\Theta_0} \cap \overline{\Theta_1}.$$

So suppose $Z^n \in \cap_{k=1}^{\infty} \left(\overline{\Omega_0^k} \cap \overline{\Omega_1^k} \right)$. By Lemma 1

$$Z^n \in \cap_{k=1}^{\infty} \left(\overline{\Theta_0^k} \cap \overline{\Theta_1^k} \right).$$

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

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

where

$$\widehat{\mathcal{H}}_0^k = \left\{ h \in \mathcal{H}^k : \widehat{R}_0(h) \le \alpha + \frac{1}{2}\epsilon_0(n_0, \delta_0, k) \right\}$$

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

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

where the second inequality comes from Remark 1 in Section III and a repetition of the argument in the proof of Theorem 3.

APPENDIX II
PROOF OF THEOREM 6

We prove the theorem in the case of VC classes, the case of finite classes being entirely analogous. Our approach is to apply the Borel–Cantelli lemma [36, p. 40] to show

$$\lim_{n \to \infty} R_0(\widehat{h}_n) \le \alpha \quad \text{and} \quad \lim_{n \to \infty} R_1(\widehat{h}_n) \le R_1^*.$$

It then follows that the second inequality must hold with equality, for otherwise there would be a classifier that strictly outperforms the optimal classifier given by the NP lemma (or equivalently, there would be an operating point above the receiver operating characteristic), a contradiction.

First consider the convergence of $R_0(\widehat{h}_n)$ to α. By the Borel–Cantelli lemma, 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

$$\Phi_0^n = \left\{ Z^n : R_0(\widehat{h}_n) - \alpha > \epsilon \right\}$$

$$\Psi_0^n = \left\{ Z^n : n_0 \le \frac{1}{2}\pi_0 n \right\}$$

$$\Theta_0^n = \left\{ Z^n : R_0(\widehat{h}_n) - \alpha > \epsilon_0(n_0, \delta_0(n), K(n)) \right\}.$$

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

$$\sum_{n=1}^{\infty} \mathbb{P}^n\left(\Phi_0^n\right) \le \sum_{n=1}^{\infty} \mathbb{P}^n\left(\Phi_0^n \cap \overline{\Psi_0^n}\right) + \sum_{n=1}^{\infty} \mathbb{P}^n\left(\Psi_0^n\right). \qquad (9)$$

To bound the second term we use the following lemma.

Lemma 3:

$$\mathbb{P}^n\left(\Psi_0^n\right) \le e^{-n\pi_0/8}.$$

Proof: The relative Chernoff bound [37] states that if $U \sim$ Binomial(n,p), then for all $\gamma > 0$, $\mathbb{P}(U/n \le (1-\gamma)p) \le e^{-np\gamma^2/2}$. 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

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

To bound the first term in (9) we use the following lemma.

Lemma 4: 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 $\epsilon'(n) \le \epsilon$. 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

$$\epsilon_0(n_0, \delta_0(n), K(n)) \le \epsilon'(n) \le \epsilon$$

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

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

where in the last line we use Theorem 5.

Now consider the convergence of $R_1(\widehat{h}_n)$ to R_1^*. As before, it suffices to show that for each $\epsilon > 0$

$$\sum_{n=1}^{\infty} \mathbb{P}^n\left(R_1(\widehat{h}_n) - R_1^* > \epsilon\right) < \infty.$$

Let $\epsilon > 0$ and define the sets

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

Arguing as before, it suffices to show

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

and

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

The first expression is bounded using an analogue of Lemma 3. To bound the second expression we employ the following analog of Lemma 4.

Lemma 5: 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 $k^* \le K(N)$ satisfying i) $\epsilon'(n, k^*) \le \epsilon/2$ and ii) $\inf_{h\in\mathcal{H}_0^{k^*}} R_1(h) - R_1^* \le \epsilon/2$. Suppose $n > N$ and $Z^n \in \Phi_1^n \cap \overline{\Psi_1^n}$. Since $Z^n \in \overline{\Psi_1^n}$ we have

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

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

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

APPENDIX III
PROOF OF THEOREM 7

Define the sets

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

Observe

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

where the next to last step follows from Theorem 5 and Lemma 3, 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

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

Since $Z^n \in \overline{\Psi_0}$, we know $\frac{1}{n_0} \le \frac{2}{\pi_0 n}$. By assumption, $\log(1/\delta_0(n)) = O(\log n)$. Furthermore, from the discussion prior to the statement of Theorem 7, we know

$$\log|\mathcal{H}^{K(n)}| \le K(n)(\log 8 + \log d)$$

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

$$R_0(\widehat{h}_n) - \alpha \preccurlyeq n^{-1/(d+1)} \qquad (10)$$

as desired.

For the miss probability, it can similarly be shown that

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

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 $\tilde{k} \le K(n)$ 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 $4dc_1 m^{d-1} \le K(n)$ and $m \le 2^L$ 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 side length $1/m$. 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**), $|\mathcal{B}_m| \le c_1 m^{d-1}$ 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

$$s(A') \equiv s(A) + 1 (\mathrm{mod}\ d).$$

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 $R_0(\tilde{h}) \le \alpha$. Note that $\tilde{h}$ has depth $J = d \log_2 m$.

By the following lemma we know $\tilde{h} \in \mathcal{H}_0^{\tilde{k}}$ for some $\tilde{k} \le K(n)$.

Lemma 6: Let $\tilde{k}$ denote the number of leaf nodes of $\tilde{h}$. Then $\tilde{k} \le K(n)$.

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

$$\sum_{j=0}^{J} c_1 2^{\lceil j/d \rceil (d-1)} \le \sum_{\ell=0}^{J/d} dc_1 2^{\ell(d-1)} \le 2dc_1 2^{(J/d)(d-1)} = 2dc_1 m^{d-1}$$

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 $\tilde{k} \le 4dc_1 m^{d-1} \le K(n)$. □

```
Input: Training sample $Z^n$
Initialize: $\mathcal{B}(\ell) = \emptyset$ for $0 \le \ell \le \ell_0$.
    $\mathcal{C}(0) = 1$ and $\mathcal{C}(\ell) = \infty$ for $1 \le \ell \le \ell_0$.
For $m = 1$ to $k^d$
    $p = n_0^m$
    For $q = 0, 1, \ldots, \ell_0 - p$ such that $\mathcal{C}(q) < \infty$
        $\ell = p + q$
        If $\mathcal{C}(\ell) > \mathcal{C}(q) - n_1^m/n_1$
            $\mathcal{C}(\ell) = \mathcal{C}(q) - n_1^m/n_1$
            $\mathcal{B}(\ell) = \mathcal{B}(q) \cup A_m$
        End
    End
End
Output: $\mathcal{B}(\ell_0)$
```

Fig. 3. Algorithm for NP-ERM with histograms.

Applying the lemma we have

$$\epsilon_1(n_1, \delta_1(n), \tilde{k}) \le \epsilon_1(n_1, \delta_1(n), K(n)) \preccurlyeq n^{-1/(d+1)}$$

where the last step follows by the same argument that produced (10).

To bound the approximation error, observe

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

where the second inequality follows from **A0** and the third from **A1**. This completes the proof.

Appendix IV
An Algorithm for NP-SRM With Histograms

NP-SRM for histograms can be implemented by solving NP-ERM for each $\mathcal{H}^k$. This yields classifiers $\widehat{h}_n^k, k = 1, \ldots, K$. The final NP-SRM is then determined by simply selecting $\widehat{h}_n^k$ that minimizes $\widehat{R}_1(\widehat{h}_n^k) + \frac{1}{2}\epsilon_1(n_1, \delta_1, k)$. The challenge is in implementing NP-ERM. Thus, let k be fixed.

An algorithm for NP-ERM over $\mathcal{H}^k$ is given in Fig. 3. The following notation is used. Let $A_m, m = 1, \ldots, k^d$, denote the individual hypercubes of side length $1/k$ that comprise histograms in $\mathcal{H}^k$. Let $n_j^m = |\{i : X_i \in A_m, Y_i = j\}|$ be the number of class j training samples in cell m. Let ℓ_0 be the largest integer such that $\ell_0/n_0 \le \alpha + \frac{1}{2}\epsilon_0(n_0, \delta_0, k)$. For $\ell = 0, 1, \ldots, \ell_0$ let $\mathcal{B}(\ell)$ denote the histogram classifier having minimum empirical miss probability $(\widehat{R}_1)$ among all histograms having $\widehat{R}_0 = \ell/n$. Let $\mathcal{C}(\ell)$ denote the empirical miss probability of $\mathcal{B}(\ell)$. Assume that histogram classifiers are represented by the cells labeled one, so that each $\mathcal{B}(\ell)$ is a collection of cells.

NP-ERM amounts to determining $\mathcal{B}(\ell_0)$. The algorithm builds up this optimal classifier in a recursive fashion. The proof that the algorithm attains the NP-ERM solutions is a straightforward inductive argument, and is omitted. The computational complexity of the algorithm is $O\left(k^d \ell_0\right)$. Assuming K is chosen so that NP-SRM is consistent, we have $k^d \le K^d = o(n)$, and hence, the complexity of NP-ERM here is $O(n\ell_0) = O(nn_0)$.

APPENDIX V
AN ALGORITHM FOR NP-SRM WITH 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 then is selected by the user (usually the maximum L such that the algorithm runs efficiently; see below) and take $K = 2^{dL}$, the largest possible meaningful value. We replace the symbol h for a generic classifier by the notation T for trees. Let $|T|$ denote the number of leaf nodes of T. We seek an algorithm implementing

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

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. Given $T \in \mathcal{T}_L$, define $\mathcal{A}_L(T)$ to be the set of all $A \in \mathcal{A}_L$ that are nodes of T. If $A \in \mathcal{A}_L(T)$, let T_A denote the subtree of T rooted at A. Given $A \in \mathcal{A}_L$, define

$$\mathcal{T}_L(A) = \{T_A : T \in \mathcal{T}_L, A \in \mathcal{A}_L(T)\}.$$

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 each $(k, \ell) \in \mathcal{I}_A$ define

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

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

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

The algorithm is stated formally in Fig. 4.

```
Input: Minimum empirical risk trees $T^{k,\ell}$ and $\mathcal{I}$
Initialize: $C^{\text{min}} = \infty$
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 $C^{\text{temp}} < C^{\text{min}}$
            $C^{\text{min}} = C^{\text{temp}}$
            $\widehat{T} = T^{k,\ell}$
        End
    End
End
Output: $\widehat{T}$
```

Fig. 4. Algorithm for NP-SRM with DDTs.

The MERTs may be computed as follows. Recall that for a hyperrectangle A we define $A^{s,1}$ and $A^{s,2}$ to be the hyperrectangles formed by splitting A at its midpoint along coordinate s. 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. First, the base of the recursion. Define $n_0^A = |\{(X_i, Y_i) \mid X_i \in A, Y_i = 0\}|$, the number of class 0 samples in cell A. When A is a cell at maximum depth $J = dL$, $T_A^{1,0} = \{A\}$ (labeled with class 0) and $T_A^{1,n_0^A} = \{A\}$ (labeled with class 1). Furthermore, $\mathcal{I}_A = \{(1,0), (1, n_0^A)\}$.

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 $j < J$

$$T_A^{k,\ell} = \arg\min\left\{\widehat{R}_1(T_A) : T_A = \text{merge}\left(A, T_{A^{s,1}}^{k',\ell'}, T_{A^{s,2}}^{k'',\ell''}\right), \\ k' + k'' = k, \ell' + \ell'' = \ell, s = 1, \ldots, d\right\}.$$

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.* [34] to derive an algorithm for penalized ERM 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 Fig. 5.

Proposition 5: The algorithm in Fig. 5 requires $O\left(n^2 n_0 d^2 L^{d+1} \log(nL^d)\right)$ operations.

Proof: The proof is a minor variation on an argument given by Blanchard *et al.* [34]. For each training point X_i there are exactly $(L+1)^d$ cells in $\mathcal{A}_L$ containing the point (see [34]). 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 $O\left(n_0 n d^2 L\right)$ 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. Computational and memory constraints limit the algorithm to problems for which $d < 15$[38]. However, if one desires a computationally efficient algorithm that achieves the rates in Theorem 7 for all d, there is an alternative. As shown in the proof of Theorem 7, it suffices to consider *cyclic* DDT's (defined in the proof). For NP-SRM with cyclic DDT's the algorithm of Fig. 5 can be can be simplified so that it requires $O\left(n_0 n^2 d^2 L^2\right)$ 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 $\mathcal{I}_A = \{(1,0),(1,n_0^A)\}$
For $j = J$ down to 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 $\mathcal{I}_A = \{(1,0),(1,n_0^A)\}$
                Set $T_A^{1,0} = \{A\}$ (label = 0)
                Set $T_A^{1,n_0^A} = \{A\}$ (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 $(k,\ell)$ to $\mathcal{I}_A$
                    $T_A^{k,\ell} = \text{MERGE}(A, T_{A'}^{k',\ell'}, T_{A''}^{k'',\ell''})$
                Else If $\widehat{R}_1(T_{A'}^{k',\ell'}) + \widehat{R}_1(T_{A''}^{k'',\ell''}) < \widehat{R}_1(T_A^{k,\ell})$
                    $T_A^{k,\ell} \leftarrow \text{MERGE}(A, T_{A'}^{k',\ell'}, T_{A''}^{k'',\ell''})$
                End
            End
        End
    End
End
Output: $T^{k,\ell}$ and $\mathcal{I}$
```

Fig. 5. Algorithm for computing minimum empirical risk trees for DDTs.

Acknowledgment

The authors wish to thank Rebecca Willett for helpful comments and suggestions.

References

[1] E. Lehmann, *Testing Statistical Hypotheses*. New York: Wiley, 1986.

[2] H. V. Poor, *An Introduction to Signal Detection and Estimation*. New York: Springer-Verlag, 1988.

[3] H. V. Trees, *Detection, Estimation, and Modulation Theory: Part I*. New York: Wiley, 2001.

[4] E. L. Lehmann, "The Fisher, Neyman-Pearson theories of testing hypotheses: One theory or two?," *J. Amer. Statist. Assoc.*, vol. 88, no. 424, pp. 1242–1249, Dec. 1993.

[5] P. Sebastiani, E. Gussoni, I. S. Kohane, and M. Ramoni, "Statistical challenges in functional genomics," *Statist. Sci.*, vol. 18, no. 1, pp. 33–60, 2003.

[6] S. Dudoit, J. Fridlyand, and T. P. Speed, "Comparison of discrimination methods for the classification of tumors using gene expression data," *J. Amer. Statist. Assoc.*, vol. 97, no. 457, pp. 77–87, Mar. 2002.

[7] B. Zadrozny, J. Langford, and N. Abe, "Cost sensitive learning by cost-proportionate example weighting," in *Proc. 3rd Int. Conf. Data Mining*, Melbourne, FL, 2003.

[8] P. Domingos, "Metacost: A general method for making classifiers cost sensitive," in *Proc. 5th Int. Conf. Knowledge Discovery and Data Mining*, San Diego, CA, 1999, pp. 155–164.

[9] C. Elkan, "The foundations of cost-sensitive learning," in *Proc. 17th Int. Joint Conf. Artificial Intelligence*, Seattle, WA, 2001, pp. 973–978.

[10] D. Margineantu, "Class probability estimation and cost-sensitive classification decisions," in *Proc. 13th European Conf. Machine Learning*, Helsinki, Finland, 2002, pp. 270–281.

[11] D. Casasent and X.-W. Chen, "Radial basis function neural networks for nonlinear Fisher discrimination and Neyman-Pearson classification," *Neural Netw.*, vol. 16, pp. 529–535, 2003.

[12] A. Cannon, J. Howse, D. Hush, and C. Scovel. (2002) Learning With the Neyman-Pearson and min-max Criteria. Los Alamos National Laboratory. [Online]. Available: http://www.c3.lanl.gov/kelly/ml/pubs/2002minmax/paper.pdf

[13] ——, (2003) Simple Classifiers. Los Alamos National Laboratory. [Online]. Available: http://www.c3.lanl.gov/ml/pubs/2003sclassifiers/abstract.shtml

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

[15] V. Vapnik and C. Chervonenkis, "On the uniform convergence of relative frequencies of events to their probabilities," *Theory Probab. Its Applic.*, vol. 16, no. 2, pp. 264–280 , 1971.

[16] ——, *Theory of Pattern Recognition* (in Russian). Moscow, U.S.S.R.: Nauka, 1971.

[17] V. Koltchinskii, "Rademacher penalties and structural risk minimization," *IEEE Trans. Inf. Theory*, vol. 47, no. 5, pp. 1902–1914, Jul. 2001.

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

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

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

[21] S. Boucheron, O. Bousquet, and G. Lugosi. (2004) Theory of Classification: A Survey of Recent Advances. [Online]. Available: http://www.econ.upf.es/lugosi/

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

[23] C. Scott and R. Nowak, "On the adaptive properties of decision trees," in *Advances in Neural Information Processing Systems 17*, L. K. Saul, Y. Weiss, and L. Bottou, Eds. Cambridge, MA: MIT Press, 2005, pp. 1225–1232.

[24] I. Steinwart and C. Scovel, "Fast rates to bayes for kernel machines," in *Advances in Neural Information Processing Systems 17*, L. K. Saul, Y. Weiss, and L. Bottou, Eds. Cambridge, MA: MIT Press, 2005, pp. 1337–1344.

[25] G. Blanchard, G. Lugosi, and N. Vayatis, "On the rate of convergence of regularized boosting classifiers," *J. Mach. Learn. Res.*, vol. 4, pp. 861–894, 2003.

[26] A. B. Tsybakov and S. A. van de Geer. (2004) Square Root Penalty: Adaptation to the Margin in Classification and in Edge Estimation. [Online]. Available: http://www.proba.jussieu.fr/pageperso/tsybakov/tsybakov.html

[27] L. Devroye, "Any discrimination rule can have an arbitrarily bad probability of error for finite sample size," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. PAMI-4, pp. 154–157, Mar. 1982.

[28] C. Scott and R. Nowak. Minimax Optimal Classification With Dyadic Decision Trees. Rice Univ., Houston, TX. [Online]. Available: http://www.stat.rice.edu/~cscott

[29] ——, "Dyadic classification trees via structural risk minimization," in *Advances in Neural Information Processing Systems 15*, S. Becker, S. Thrun, and K. Obermayer, Eds. Cambridge, MA, 2003.

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

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

[32] A. Cohen, W. Dahmen, I. Daubechies, and R. A. DeVore, "Tree approximation and optimal encoding," *Appl. Comput. Harmonic Anal.*, vol. 11, no. 2, pp. 192–226, 2001.

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

[34] G. Blanchard, C. Schäfer, and Y. Rozenholc, "Oracle bounds and exact algorithm for dyadic classification trees," in *Learning Theory: Proc. 17th Annu. Conf. Learning Theory, COLT 2004*, J. Shawe-Taylor and Y. Singer, Eds. Heidelberg, Germany: Sspringer-Verlag, 2004, pp. 378–392.

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

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

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

[38] G. Blanchard, C. Schäfer, Y. Rozenholc, and K.-R. Müller, "Optimal Dyadic Decision Trees," preprint, 2005.