Electronic Journal of Statistics
Vol. 3 (2009) 651–677
ISSN: 1935-7524
DOI: 10.1214/09-EJS363

The false discovery rate for statistical pattern recognition

Clayton Scott[*] **and Gowtham Bellala**

Electrical Engineering and Computer Science
University of Michigan, Ann Arbor
e-mail: clayscot@umich.edu; gowtham@umich.edu

Rebecca Willett[†]

Electrical and Computer Engineering
Duke University
e-mail: willett@duke.edu

Abstract: The false discovery rate (FDR) and false nondiscovery rate (FNDR) have received considerable attention in the literature on multiple testing. These performance measures are also appropriate for classification, and in this work we develop generalization error analyses for FDR and FNDR when learning a classifier from labeled training data. Unlike more conventional classification performance measures, the empirical FDR and FNDR are not binomial random variables but rather a ratio of binomials, which introduces challenges not present in conventional formulations of the classification problem. We develop distribution-free uniform deviation bounds and apply these to obtain finite sample bounds and strong universal consistency. We also present a simulation study demonstrating the merits of variance-based bounds, which we also develop. In the context of multiple testing with FDR/FNDR, our framework may be viewed as a way to leverage training data to achieve distribution free, asymptotically optimal inference under the random effects model.

AMS 2000 subject classifications: Primary 62H30; secondary 68T05.
Keywords and phrases: Statistical learning theory, generalization error, false discovery rate.

Contents

[*] C. Scott was partially supported by NSF award no. 0830480.

[†] R. Willett was partially supported by DARPA grant no. HR0011-06-1-0026 and NSF CAREER award no. NSF-CCF-06-43947.

1. Introduction

Consider the problem of learning a classifier from labeled training data. Traditional learning algorithms are designed to optimize the probability of error, or some combination of false positive and negative rates. These criteria are typically viewed as measures of performance on a *single* future test point. However, it is often the case that we desire to classify *multiple* future test points, in which case these traditional performance measures may be inappropriate. This situation is similar to the multiple testing problem in hypothesis testing, where the goal is also to assign labels to several measurements simultaneously. The basic approach adopted there is to consider alternative measures of size and power that are better suited to multiple inference, and to design decision rules based on these new performance measures. In this paper we extend this idea to classification.

In particular, we investigate measuring classification performance in terms of the false discovery rate (FDR) [32] and its companion quantity the false nondiscovery rate (FNDR). FDR has emerged as the method of choice for quantifying error rates meaningfully in many multiple testing situations, with applications ranging from wavelet denoising [17] to neuroimaging [22] to the analysis of DNA microarrays [19]. Control of the FDR, i.e., the fraction of declared positives (discoveries) that are in fact negative, ensures that follow-up investigations into declared positives must return a certain yield of actual positives. Such control is vital in applications where follow-up studies are time or resource consuming. Several researchers, spurred by the seminal work of [5], have studied FDR control in the context of multiple hypothesis testing.

The problem we consider is different from the usual multiple testing problem in the following respects. Multiple testing is concerned with making discoveries among an observed, unlabeled dataset. It is typically assumed that p-values can be calculated or estimated. Then, to control FDR, p-values are adjusted through one of a variety of single step, step-up or step-down procedures, and thresholded. We are concerned with building a classifier based on labeled training data, and making predictions on future test points. In our setup, we assume the unlabeled test data will be observed after the classifier is learned. Thus, we are in the *inductive* setting; in the *transductive* setting the test points would be observed before learning. We assume measurements are iid, which coincides with the "random effects model" in the multiple testing literature.

Our framework may be viewed as a way to leverage training data, when available, to overcome what are often weakness of multiple testing based on adjusted p-values. First, in many applications, the null distribution cannot not be easily modeled or simulated, so that p-values are difficult to estimate. Second, thresholding adjusting p-values is almost never optimal in terms of minimizing FNDR or some other measure of Type II error [35]. In this work, we show how training data may be used to *adapt* to both the null and alternative distributions, while making no assumptions whatsoever on either. Adopting the perspective of statistical learning theory, we develop distribution free results on the generalization error analysis of FDR and FNDR, including uniform deviation bounds, finite sample performance guarantees, strong universal consistency, and variance-based bounds.

1.1. Motivation

We offer two motivating examples. First, consider a network operator who seeks to detect anomalous activities in a network such as malware propagation. For example, suppose that every ten minutes, a feature vector X is observed, where X consists of different measurable quantities such as traffic volumes on different links. The process of determining the presence of an anomaly is laborious, so the operator would like a predictor based on X to tell him whether to invest the time and energy to determine if an anomaly occurred. Now suppose that some historical data $(X_i, Y_i)_{i=1}^n$ are available, such as measurements from the past month gathered under similar conditions (e.g., time of day), where every label Y_i was carefully determined. A classifier based on this historical data should take into consideration the fact that it will be used multiple times, and that the network operator is willing to investigate many events if they turn out to be positive, but does not want to chase down too many false leads. In this setting, the most natural approach may be to control the false discovery rate (FDR) of the classifier.

Alternatively, consider a neuroimaging (e.g., functional MRI) experiment conducted on multiple subjects performing some task. For each subject, we would like to identify which voxels in the brain are active during different phases of the experiment. In particular, suppose each voxel is associated with a vector X consisting of features derived from the voxel time series. Activity detection is then commonly accomplished by assuming that inactive voxels generate Gaussian-distributed measurements, calculating p-values, and adjusting these p-values to ensure a small FDR. This is a useful alternative to earlier approaches which determined thresholds based on a target false positive rate, since a 5% threshold generated many false positives and suggested significant scattered brain activity which was not reasonable from a biological perspective [22].

Despite the relative success of FDR for p-value adjustment in neuroimaging, this approach has some limitations. For instance, as mentioned above, accurate modeling of inactive voxel time-series can be a challenge. Furthermore, robust theoretical analyses of this approach focus on voxel-wise thresholding decisions, even though there are biological reasons to expect active voxels to be clustered together. In contrast, if training data are available (provided by, say, a neuroimaging expert), our approach can easily handle features that include spatial coordinates, and the shape of possible activation patterns can be influenced by specifying a prior on the space of possible classifiers.

1.2. Notation

More formally, in this paper we consider the following scenario: Let $\mathcal{X}$ be a set and $Z = (X, Y)$ be a random variable taking values in $\mathcal{Z} = \mathcal{X} \times \{0, 1\}$. The variable X corresponds to a pattern or feature vector and Y to a class label associated with X; $Y = 0$ corresponds to the null hypothesis (e.g., that no target is present) and $Y = 1$ corresponds to the alternative hypothesis (e.g., that a target is present). The distribution on Z is unknown and is denoted by $\mathbf{P}$. Assume we make n independent and identically distributed training observations of Z, denoted $Z^n = (X_i, Y_i)_{i=1}^n$. This corresponds to the *random effects model* common in the multiple testing literature [19].

A classifier is a function $h : \mathcal{X} \longrightarrow \{0, 1\}$ mapping feature vectors to class labels. Let $\mathcal{H}$ denote a collection of different classifiers. A false discovery occurs when $h(X) = 1$ but the true label is $Y = 0$. Similarly, a false nondiscovery occurs when $h(X) = 0$ but $Y = 1$. We define the *false discovery rate* (FDR)

$$\mathcal{R}_D(h) := \begin{cases} \mathbf{P}(Y = 0 \,|\, h(X) = 1), & \text{if } \mathbf{P}(h(X) = 1) > 0, \\ \infty, & \text{else,} \end{cases}$$

and the *false nondiscovery rate* (FNDR)

$$\mathcal{R}_{ND}(h) := \begin{cases} \mathbf{P}(Y = 1 \,|\, h(X) = 0), & \text{if } \mathbf{P}(h(X) = 0) > 0, \\ \infty, & \text{else.} \end{cases}$$

1.3. Related Concepts

These definitions, which are natural in the classification setting, coincide with the so-called *positive* FDR/FNDR of Storey [33, 34], so named because it can be seen to equal the expected fraction of false discoveries/nondiscoveries, conditioned on a positive number of discoveries/nondiscoveries having been made. Storey makes some decision-theoretic connections to classification [34], but does not consider learning from data.

Storey's definition does not cover the case where the conditioning event has probability zero. Our definition in these cases is motivated primarily by technical considerations. In particular, it allows us to treat extreme cases where true positives (or negatives) are never observed, and thereby establish a universally consistent rule. Our definition has the effect of assigning high costs to classifiers that fail to make at least some discoveries (and nondiscoveries). This is consistent with the multiple testing perspective, where often the goal is to generate discoveries for further examination, just not too many false ones. Further comments on our definitions of FDR and FNDR are given after the proof of Theorem 2.

In certain communities, different terms embody the idea behind FDR. In the medical diagnostic testing literature, the *positive predictive value* (PPV) is defined as the "proportion of patients with positive test results who are correctly diagnosed" [2]. In database information retrieval problems, the *precision* is defined as the ratio of the number of relevant documents retrieved by a search to the total number of documents retrieved by a search [38]. Both PPV and precision are equal to 1 - FDR. Precision is discussed further is Section 7.2.

Finally, several researchers have recently investigated connections between multiple testing and statistical learning theory. McAllester's PAC-Bayesian learning theory may be viewed as an extension of multiple testing procedures to (possibly uncountably) infinite collections of hypotheses [26]. Blanchard and Fleuret present an extension of the Occam's razor principle for generalization error analysis in classification, and apply it to derive p-value adjustment procedures for controlling FDR [8]. Arlot et al. develop concentration inequalities that apply to multiple testing with correlated observations [3]. Scott and Blanchard present a learning theoretic analysis of semi-supervised novelty detection, which includes multiple testing under the random effects model as a special case [28]. None of these works consider FDR/FNDR as performance criteria for classification.

1.4. Connections to Cost-Sensitive Learning and Related Problems

In Sections 3 and Section 4 we consider the performance measure $\mathcal{E}_\lambda(h) := \mathcal{R}_{ND}(h) + \lambda \mathcal{R}_D(h)$. This criterion is evocative of the cost-sensitive Bayes risk,

$$\mathbf{P}(h(X) = 0, Y = 1) + \gamma \mathbf{P}(h(X) = 1, Y = 0),$$

$\gamma > 0$. The global minimizer of this risk has the form

$$h(x) = \mathbf{1}_{\{\eta(x) \geq c\}}, \tag{1}$$

where $\eta(x) := \mathbf{P}(Y = 1 | X = x)$ and $c = 1/(1 + \gamma)$. Proof of this fact is a direct generalization of the case of the probability of error, when $\gamma = 1$ [15]. Furthermore, many statistical learning algorithms have been developed to perform cost-sensitive classification [4, 21], and their theoretical properties are typically similar to those of standard classification algorithms [7].

Unfortunately, algorithms for cost-sensitive classification cannot be easily applied to our problem. In particular, we may write

$$\begin{aligned} \mathcal{E}_\lambda(h) &= \mathbf{P}(Y = 1 | h(X) = 0) + \lambda \mathbf{P}(Y = 0 | h(X) = 1) \\ &\propto \mathbf{P}(h(X) = 0, Y = 1) + \left[\lambda \frac{\mathbf{P}(h(X) = 0)}{\mathbf{P}(h(X) = 1)} \right] \mathbf{P}(h(X) = 1, Y = 0). \end{aligned}$$

The term in brackets is not constant, but depends on the classifier h as well as the unknown distribution.

Despite this lack of any simple reduction, it is still true that the family of classifiers in (1) are the global minimizers of $\mathcal{E}_\lambda$, with λ and the corresponding c obeying a monotone relationship. Storey [34] gives a proof of this fact for the case where the two class-conditional distributions are continuous. A precise statement of the result requires that this family be extended by a standard randomization argument if its receiver operating characteristic (ROC) is not concave.

This suggests that it may still be possible to minimize $\mathcal{E}_\lambda$ by performing cost-sensitive classification with a certain cost γ. The *critical issue* is that γ is an implicit function of λ, and cannot be determined a priori without knowledge of the underlying distribution. Thus, when only data are given, applying existing cost-sensitive classification methods to our problem would require estimating γ. In practice, this would most likely entail learning cost-sensitive classifiers $\widehat{h}_{\gamma_i}$ for some grid of values $\{\gamma_i\}$ that grows increasingly dense as $n \to \infty$. Then, the best of these candidates would be selected by minimizing an estimate of $\mathcal{E}_\lambda(h)$. Such a procedure would likely be expensive computationally. From an analytical standpoint, it seems plausible that generalization error analyses for cost-sensitive classification could be useful; however, the need to search for a γ that approximately minimizes our criterion would certainly complicate the analysis. The objective of our work is to develop a much more direct approach, which does not require repeated cost-sensitive classification.

Similarly, in Section 7, we consider minimizing $\mathcal{R}_{ND}(h)$ subject to $\mathcal{R}_D(h) < \alpha$. This has clear ties to Neyman-Pearson classification [10, 29, 27]. Once again, however, there is no direct reduction from this problem to ours. Furthermore, we present a general result for learning with a Type I error constraint whose proof is much simpler than those given in [10, 30].

Connections to other learning problems may also be drawn. Liu et al. [24] apply VC theory to study the problem of optimizing precision subject to a constraint on recall. Their precision bound is loose and evidently does not lead to a consistent procedure. Recently, Clémencon and Vayatis have studied learning theoretic aspects of learning the entire precision-recall curve [13]. More generally, several authors have analyzed learning the entire ROC for bipartite ranking [1, 37, 12, 14]. Our problem amounts to estimating a specific point on this curve, where the point of interest (for us) on this curve depends not only on the value of λ, but also on the unknown distribution. Therefore, adapting these methods to our setting is likely to be no easier than it would be to adapt methods for cost-sensitive classification.

1.5. Overview

In the next section we present and prove uniform deviation bounds for FDR and FNDR. In Section 3, we discuss performance measures based on FDR and

FNDR, and in Section 4 we establish the strong universal consistency of a learning rule with respect to the measure $\mathcal{E}_\lambda$. A variance-based bounding technique is employed in Section 5 to a faster rate of convergence for the zero-error case, and in Section 6 we present an experimental evaluation and comparison of our bounds. Section 7 treats performance measures which constrain FDR, and the final section offers a concluding discussion.

Unlike traditional performance measures, whose empirical versions are related to binomial random variables, empirical versions of FDR and FNDR are related to ratios of binomial variables. Thus, our analytical methods are combinations of both existing and new techniques.

2. Uniform Deviation Bounds

Define empirical analogues to the FDR and FNDR according to

$$\begin{aligned}
\widehat{\mathcal{R}}_D(h) \quad &:= \quad \begin{cases} \frac{1}{n_D(h,Z^n)} \sum_{i=1}^n \mathbf{1}_{\{Y_i=0,h(X_i)=1\}}, & n_D(h,Z^n)>0, \\ \infty, & n_D(h,Z^n)=0, \end{cases} \\
\widehat{\mathcal{R}}_{ND}(h) \quad &:= \quad \begin{cases} \frac{1}{n_{ND}(h,Z^n)} \sum_{i=1}^n \mathbf{1}_{\{Y_i=1,h(X_i)=0\}}, & n_{ND}(h,Z^n)>0, \\ \infty, & n_{ND}(h,Z^n)=0, \end{cases}
\end{aligned}$$

where $n_D(h,Z^n) = \sum_{i=1}^n \mathbf{1}_{\{h(X_i)=1\}}$ and $n_{ND}(h,Z^n) = \sum_{i=1}^n \mathbf{1}_{\{h(X_i)=0\}}$ are binomial random variables. When $n_D(h,Z^n)$ and $n_{ND}(h,Z^n)$ are greater than zero, $\widehat{\mathcal{R}}_D(h)$ and $\widehat{\mathcal{R}}_{ND}(h)$ are known as the *false discovery proportion* and *false nondiscovery proportion*, respectively. Storey showed that $E[\widehat{\mathcal{R}}_D(h)|n_D(h,Z^n)>0] = \mathcal{R}_D(h)$, and similarly for FNDR [34]. This section describes a uniform bound on the amount by which the empirical estimate of FDR/FNDR can deviate from the true value. Note that unlike the usual empirical estimates for the probability of error/false positive rate/false negative rate, here both numerator and denominator are random, and both depend on h.

Assume $\mathcal{H}$ is countable, and let $[\![h]\!]$ be a real valued functional on $\mathcal{H}$ such that $\sum_{h\in\mathcal{H}} 2^{-[\![h]\!]} \le 1$. If $[\![h]\!]$ is integer valued, such a functional can be identified with a prefix code for $\mathcal{H}$, in which case $[\![h]\!]$ is the codelength associated to h. If $\sum_{h\in\mathcal{H}} 2^{-[\![h]\!]} = 1$, then $2^{-[\![h]\!]}$ may be viewed as a prior distribution on $\mathcal{H}$.

For $\delta > 0$, we introduce the *penalty* terms

$$\begin{aligned}
\phi_D(h,\delta) \quad &:= \quad \begin{cases} \sqrt{\frac{[\![h]\!]\ln 2 + \ln(2/\delta)}{2n_D(h,Z^n)}}, & n_D(h,Z^n)>0, \\ \infty, & n_D(h,Z^n)=0, \end{cases} \\
\phi_{ND}(h,\delta) \quad &:= \quad \begin{cases} \sqrt{\frac{[\![h]\!]\ln 2 + \ln(2/\delta)}{2n_{ND}(h,Z^n)}}, & n_{ND}(h,Z^n)>0, \\ \infty, & n_{ND}(h,Z^n)=0. \end{cases}
\end{aligned}$$

The interpretation of these expressions as penalties comes from the learning algorithms studied below, where we minimize the empirical error plus a penalty to avoid overfitting. Note that the penalties are data dependent.

Theorem 1. *With probability at least* $1-\delta$ *with respect to the draw of the training data,*

$$|\mathcal{R}_D(h) - \widehat{\mathcal{R}}_D(h)| \leq \phi_D(h,\delta) \tag{2}$$

for all $h \in \mathcal{H}$*. Similarly, with probability at least* $1-\delta$ *with respect to the draw of the training data,*

$$|\mathcal{R}_{ND}(h) - \widehat{\mathcal{R}}_{ND}(h)| \leq \phi_{ND}(h,\delta) \tag{3}$$

for all $h \in \mathcal{H}$*. The results are independent of the underlying probability distribution.*

Because of the form of the penalty terms, the bound is larger for classifiers h that are more complex, as represented through the codelength $[\![h]\!]$, and smaller when more discoveries/nondiscoveries are made. This result leads to finite sample bounds and strong universal consistency for certain learning rules based on minimization of the penalized empirical error, as developed in the sequel.

Proof. We prove the first statement, the second being similar. For added clarity, write the penalty as $\phi_D(h,\delta,n_D(h,Z^n))$, where

$$\phi_D(h,\delta,k) := \begin{cases} \sqrt{\frac{[\![h]\!]\ln 2 + \ln(2/\delta)}{2k}}, & k > 0, \\ \infty, & k = 0. \end{cases}$$

Consider a fixed $h \in \mathcal{H}$. The fundamental concentration inequality underlying our bounds is Hoeffding's [23], which, in one form, states that if S_k is the sum of $k > 0$ independent random variables bounded between zero and one, and $\mu = \mathbf{E}[S_k]$, then

$$\mathbf{P}(|\mu - S_k| > k\epsilon) \leq 2e^{-2k\epsilon^2}.$$

To apply Hoeffding's inequality, we need the following conditioning argument. Let $V = (V_1,\ldots,V_n) \in \{0,1\}^n$ be a binary indicator vector, with $V_i = h(X_i)$. Let $\mathcal{V}_k$ denote the set of all $v = (v_1,\ldots,v_n) \in \{0,1\}^n$ such that $\sum_{i=1}^n v_i = k$. We may then write

$$\begin{aligned} &\mathbf{P}(|\mathcal{R}_D(h) - \widehat{\mathcal{R}}_D(h)| > \phi_D(h,\delta,n_D(h,Z^n))) \\ &\quad = \sum_{k=0}^{n} \sum_{v\in\mathcal{V}_k} \mathbf{P}(|\mathcal{R}_D(h) - \widehat{\mathcal{R}}_D(h)| > \phi_D(h,\delta,k)|V=v)\mathbf{P}(V=v) \\ &\quad = \sum_{k=0}^{n} \sum_{v\in\mathcal{V}_k} \mathbf{P}(|k\mathcal{R}_D(h) - k\widehat{\mathcal{R}}_D(h)| > k\phi_D(h,\delta,k)|V=v)\mathbf{P}(V=v), \end{aligned}$$

First note that $|\mathcal{R}_D(h) - \widehat{\mathcal{R}}_D(h)| \leq \phi_D(h,\delta)$ with probability one when $n_D(h,Z^n) = 0$. We now apply Hoeffding's inequality for each $k \geq 1$ and $v \in \mathcal{V}_k$, conditioning on $V = v$. Setting $S_k = k\widehat{\mathcal{R}}_D(h)$, we have

$$\begin{aligned}\mu &= \mathbf{E}[S_k | V = v] \\ &= k\mathbf{E}[\widehat{\mathcal{R}}_D(h) | V = v] \\ &= \mathbf{E}[\sum_{i=1}^{n} \mathbf{1}_{\{Y_i=0,h(X_i)=1\}} | V = v] \\ &= \mathbf{E}[\sum_{i:v_i=1} \mathbf{1}_{\{Y_i=0\}} | V = v] \\ &= k\mathbf{P}(Y = 0 | h(X) = 1) \\ &= k\mathcal{R}_D(h),\end{aligned}$$

where in the next to last step we use independence of the realizations. Therefore, applying Hoeffding's inequality conditioned on $V = v \in \mathcal{V}_k$ yields

$$\begin{aligned}&\mathbf{P}(|\mathcal{R}_D(h) - \widehat{\mathcal{R}}_D(h)| > \phi_D(h, \delta, n_D(h, Z^n))) \\ \leq\ & \sum_{k=1}^{n} \sum_{v\in\mathcal{V}_k} 2e^{-2k\phi_D^2(h,\delta,k)} \mathbf{P}(V = v) \\ \leq\ & \sum_{k=1}^{n} \sum_{v\in\mathcal{V}_k} \delta 2^{-[\![h]\!]} \mathbf{P}(V = v) \\ =\ & \delta 2^{-[\![h]\!]}(1 - \mathbf{P}(\textstyle\sum V_i = 0)) \leq \delta 2^{-[\![h]\!]}.\end{aligned}$$

The result now follows by applying the union bound over all $h \in \mathcal{H}$. □

The technique of conditioning on the random denominator of a ratio of binomials has also been applied in others settings [25, 29, 11]. Unlike those works, however, here the binomial denominator depends on the classifier h. This presents difficulties for extending the above techniques to uncountable classes $\mathcal{H}$. See the final section for further discussion of this issue.

3. Measuring Performance

We would like to be able to make FDR/FNDR related guarantees about how a data-based classifier $\widehat{h}$ performs. For this, we need to specify a performance measure or optimality criterion that incorporates both FDR and FNDR quantities simultaneously. One possibility is to specify a number $0 < \alpha < 1$ and seek the classifier such that $\mathcal{R}_{ND}(h)$ is minimal while $\mathcal{R}_D(h) \leq \alpha$. We consider this setting in Section 7. Another is to specify a constant $\lambda > 0$ reflecting the relative cost of FDR to FNDR, and minimize

$$\mathcal{E}_\lambda(h) := \mathcal{R}_{ND}(h) + \lambda\mathcal{R}_D(h).$$

This measure was introduce by Storey [34], but was not studied in a learning context. The uniform deviation bounds of the previous section immediately imply the following computable bound on a classifier's performance with respect to this measure.

Corollary 1. *For any $\delta > 0$ and $n \geq 1$, with probability at least $1 - 2\delta$ with respect to the draw of the training data,*

$$\mathcal{E}_\lambda(h) \leq \widehat{\mathcal{R}}_{ND}(h) + \phi_{ND}(h, \delta) + \lambda[\widehat{\mathcal{R}}_D(h) + \phi_D(h, \delta)]$$

for all $h \in \mathcal{H}$.

In the next section, we analyze a learning rule based on minimizing the bound of Corollary 1, and establish its strong universal consistency.

4. Strong Universal Consistency

Denote the globally optimal value of the performance measure by

$$\mathcal{E}^*_\lambda := \inf_h \mathcal{E}_\lambda(h),$$

where the inf is over all measurable $h : \mathcal{X} \to \{0, 1\}$. We seek a learning rule $\widehat{h}_{\lambda,n}$ such that $\mathcal{E}_\lambda(\widehat{h}_{\lambda,n}) \to \mathcal{E}^*_\lambda$ almost surely, regardless of the underlying probability distribution. Thus let $\{\mathcal{H}_k\}_{k \geq 1}$ be a family of finite sets of classifiers with universal approximation capability. That is, assume that $\lim_{k\to\infty} \inf_{h\in\mathcal{H}_k} \mathcal{E}_\lambda(h) = \mathcal{E}^*_\lambda$ for all distributions on (X, Y). Furthermore, assume this family to be *nested*, meaning $\mathcal{H}_1 \subseteq \mathcal{H}_2 \subseteq \mathcal{H}_3 \cdots$. For example, if $\mathcal{X} = [0, 1]^d$, we may take $\mathcal{H}_k$ to be the collection of histogram classifiers based on a binwidth of 2^{-k}. Recall that we can set $[\![h]\!] = \log_2 |\mathcal{H}_k|$ for $h \in \mathcal{H}_k$, where $|\mathcal{H}_k|$ is the cardinality of $\mathcal{H}_k$. For histograms, we have $|\mathcal{H}_k| = 2^{2^{kd}}$ and hence $[\![h]\!] = 2^{kd} \ln 2$.

The bound of Corollary 1 suggests bound minimization as a strategy for selecting a classifier empirically. However, rather than minimizing over all possible classifiers in some $\mathcal{H}_k$, we first discard those classifiers whose empirical numbers of discoveries or nondiscoveries are too small. In these cases, the penalties are possibly quite large, and we are unable to obtain tight concentrations of empirical FDR/FNDR measures around their true values. However, as n increases, we are able to admit classifiers with increasingly small proportions of (non)discoveries, so that in the limit, we can still approximate arbitrary distributions. This aspect is another unique feature of FDR/FNDR compared to traditional performance measures.

Formally, define

$$\widehat{\mathcal{H}}_n := \{h \in \mathcal{H}_{k_n} : \frac{n_{ND}(h, Z^n)}{n} \geq p_n, \frac{n_D(h, Z^n)}{n} \geq p_n\},$$

where $p_n := (\ln n)^{-1}$. Here k_n is such that $k_n \to \infty$ as $n \to \infty$ and $\ln |\mathcal{H}_{k_n}| = o(n/ \ln n)$. For the histogram example, $\ln |\mathcal{H}_{k_n}| = 2^{k_n d} \ln 2$, and thus the assumed conditions on the growth of k_n are essentially the same (up to a logarithmic factor) as for consistency of histograms in other problems. For example, in standard classification, $2^{k_n d} = o(n)$ is required [15].

For concreteness, set $\delta_n = 1/n^2$. Denote the bound of Corollary 1 by

$$\tilde{\mathcal{E}}_\lambda(h) := \widehat{\mathcal{R}}_{ND}(h) + \phi_{ND}(h, \delta_n) + \lambda[\widehat{\mathcal{R}}_D(h) + \phi_D(h, \delta_n)],$$

and define the classification rule

$$\widehat{h}_{\lambda,n} \quad := \quad \underset{h \in \widehat{\mathcal{H}}_n}{\arg\min} \ \tilde{\mathcal{E}}_\lambda(h).$$

If $\widehat{\mathcal{H}}_n = \emptyset$, then $\widehat{h}_{\lambda,n}$ may be defined arbitrarily.

Theorem 2. *For any distribution on (X, Y), and any $\lambda > 0$,*

$$\mathcal{E}_\lambda(\widehat{h}_{\lambda,n}) \to \mathcal{E}^*_\lambda$$

almost surely. That is, $\widehat{h}_{\lambda,n}$ is strongly universally consistent.

Proof. First consider the case where there is no measurable $h : \mathcal{X} \to \{0, 1\}$ such that both $\mathbf{P}(h(X) = 0) > 0$ and $\mathbf{P}(h(X) = 1) > 0$. This occurs when X is deterministic. Then $\mathcal{E}^*_\lambda = \infty$, and trivially $\widehat{h}_{\lambda,n}$ achieves optimal performance. So assume this is not the case.

By the Borel-Cantelli lemma [18,15], it suffices to show that for each $\epsilon > 0$

$$\sum_{n=1}^{\infty} \mathbf{P}(\Omega^n) < \infty,$$

where

$$\Omega^n := \{Z^n : \mathcal{E}_\lambda(\widehat{h}_{\lambda,n}) - \mathcal{E}^*_\lambda \geq \epsilon\}.$$

Introduce the event

$$\Theta^n = \{Z^n : \widehat{\mathcal{H}}_n \neq \emptyset\}.$$

Since

$$\mathbf{P}(\Omega^n) \leq \mathbf{P}(\Omega^n \cap \Theta^n) + \mathbf{P}(\overline{\Theta^n}),$$

we have

$$\sum_{n=1}^{\infty} \mathbf{P}(\Omega^n) \leq \sum_{n=1}^{\infty} \mathbf{P}(\Omega^n \cap \Theta^n) + \sum_{n=1}^{\infty} \mathbf{P}(\overline{\Theta^n}). \tag{4}$$

We will bound these two terms separately.

Consider the second term.

Lemma 1. *Let $\nu > 0$ and assume $\mathcal{E}^*_\lambda < \infty$. There exist h' and N_1 such that $\mathcal{E}_\lambda(h') \leq \mathcal{E}^*_\lambda + \nu$ and, for all $n > N_1$, $\mathbf{P}(h' \in \widehat{\mathcal{H}}_n) \geq 1 - 1/n^2$.*

Proof. By the universal approximation assumption, there exist m and $h' \in \mathcal{H}_{k_m}$ such that $\mathcal{E}_\lambda(h') \leq \mathcal{E}^*_\lambda + \nu$. Since $\mathcal{E}^*_\lambda < \infty$, this h' necessarily has both $\mathbf{P}(h'(X) = 0) > 0$ and $\mathbf{P}(h'(X) = 1) > 0$. Denote $q := \min\{\mathbf{P}(h'(X) = 1), \mathbf{P}(h'(X) = 0)\} > 0$. Introduce

$$\tau_n := \sqrt{\frac{\ln(2/\delta_n)}{2n}}.$$

By Hoeffding's inequality, with probability at least $1-\delta_n$, $|\mathbf{P}(h'(X)=1) - n_D(h', Z^n)/n| = |\mathbf{P}(h'(X)=0) - n_{ND}(h', Z^n)/n| \leq \tau_n$. Since $\delta_n = 1/n^2$, we have that $\tau_n = o(p_n)$. Now choose N_1 such that $\tau_{N_1} \leq p_{N_1}$ and $2p_{N_1} \leq q$. Then, for a sample of size $n = N_1$, $\min\{n_D(h', Z^n)/N_1, n_{ND}(h', Z^n)/N_1\} \geq q - \epsilon_{N_1} \geq 2p_{N_1} - \tau_{N_1} \geq p_{N_1}$ with probability at least $1 - \delta_{N_1} = 1 - 1/{N_1}^2$. Since p_n is decreasing and $\{\mathcal{H}_k\}$ is nested, the same is true for all $n > N_1$. □

By this lemma we have $\mathbf{P}(\overline{\Theta^n}) \leq \delta_n = 1/n^2$ for all $n > N_1$ (Here we only the need the second part of the conclusion of the lemma; later we use the lemma in its full generality). Thus

$$\sum_{n=1}^{\infty} \mathbf{P}(\overline{\Theta^n}) \leq N_1 + \sum_{n>N_1} \frac{1}{n^2} < \infty.$$

Now consider the first term on the right-hand side of (4). Define the events

$$\begin{aligned} \Omega_1^n &:= \{Z^n : \mathcal{E}_\lambda(\widehat{h}_{\lambda,n}) - \inf_{h\in\widehat{\mathcal{H}}_n} \mathcal{E}_\lambda(h) \geq \frac{\epsilon}{2}\} \\ \Omega_2^n &:= \{Z^n : \inf_{h\in\widehat{\mathcal{H}}_n} \mathcal{E}_\lambda(h) - \mathcal{E}_\lambda^* \geq \frac{\epsilon}{2}\} \end{aligned}$$

Since $\Omega^n \subset \Omega_1^n \bigcup \Omega_2^n$, we have

$$\sum_{n=1}^{\infty} \mathbf{P}(\Omega^n \cap \Theta^n) \leq \sum_{n=1}^{\infty} \mathbf{P}(\Omega_1^n \cap \Theta^n) + \sum_{n=1}^{\infty} \mathbf{P}(\Omega_2^n \cap \Theta^n). \tag{5}$$

We consider the two terms individually and show that each of them is finite.

To bound the first term on the right-hand side of (5) we use the following lemma.

Lemma 2. *If $\widehat{\mathcal{H}}_n \neq \emptyset$, then*

$$\mathcal{E}_\lambda(\widehat{h}_{\lambda,n}) - \inf_{h\in\widehat{\mathcal{H}}_n} \mathcal{E}_\lambda(h) \leq 2 \sup_{h\in\widehat{\mathcal{H}}_n} |\mathcal{E}_\lambda(h) - \tilde{\mathcal{E}}_\lambda(h)|.$$

Proof. Let $h' \in \widehat{\mathcal{H}}_n$ be arbitrary. By the definition of $\widehat{h}_{\lambda,n}$, $\tilde{\mathcal{E}}_\lambda(\widehat{h}_{\lambda,n}) \leq \tilde{\mathcal{E}}_\lambda(h')$. Hence

$$\begin{aligned} \mathcal{E}_\lambda(\widehat{h}_{\lambda,n}) &= \mathcal{E}_\lambda(\widehat{h}_{\lambda,n}) - \tilde{\mathcal{E}}_\lambda(\widehat{h}_{\lambda,n}) + \tilde{\mathcal{E}}_\lambda(\widehat{h}_{\lambda,n}) - \mathcal{E}_\lambda(h') + \mathcal{E}_\lambda(h') \\ &\leq \mathcal{E}_\lambda(\widehat{h}_{\lambda,n}) - \tilde{\mathcal{E}}_\lambda(\widehat{h}_{\lambda,n}) + \tilde{\mathcal{E}}_\lambda(h') - \mathcal{E}_\lambda(h') + \mathcal{E}_\lambda(h') \\ &\leq 2 \sup_{h\in\widehat{\mathcal{H}}_n} |\mathcal{E}_\lambda(h) - \tilde{\mathcal{E}}_\lambda(h)| + \mathcal{E}_\lambda(h'). \end{aligned}$$

Since h' was arbitrary, the result now follows. □

Define the events

$$\begin{aligned}
\Omega_{11}^n &:= \{Z^n : \sup_{h\in\widehat{\mathcal{H}}_n} |\mathcal{R}_{ND}(h) - \widehat{\mathcal{R}}_{ND}(h)| \geq \frac{\epsilon}{16}\} \\
\Omega_{12}^n &:= \{Z^n : \sup_{h\in\widehat{\mathcal{H}}_n} |\mathcal{R}_{D}(h) - \widehat{\mathcal{R}}_{D}(h)| \geq \frac{\epsilon}{16\lambda}\} \\
\Omega_{13}^n &:= \{Z^n : \sup_{h\in\widehat{\mathcal{H}}_n} |\phi_{ND}(h,\delta_n)| \geq \frac{\epsilon}{16}\} \\
\Omega_{14}^n &:= \{Z^n : \sup_{h\in\widehat{\mathcal{H}}_n} |\phi_{D}(h,\delta_n)| \geq \frac{\epsilon}{16\lambda}\}
\end{aligned}$$

From Lemma 2 it follows that

$$\Omega_1^n \subset \bigcup_{i=1}^4 \Omega_{1i}^n$$

and hence it suffices to show

$$\sum_{n=1}^\infty \mathbf{P}(\Omega_{1i}^n \cap \Theta^n)$$

is finite for each $i = 1, 2, 3, 4$. We shall consider Ω_{11} and Ω_{13}, the other two cases following similarly.

For $h \in \widehat{\mathcal{H}}_n$ we have $n_{ND}(h, Z^n)/n \geq p_n$ and therefore

$$\begin{aligned}
\phi_{ND}(h,\delta_n) &= \sqrt{\frac{\ln|\mathcal{H}_{k_n}| + \ln(2n^2)}{2n_{ND}(h,Z^n)}} \\
&\leq \sqrt{(\ln|\mathcal{H}_{k_n}| + \ln(2n^2))\frac{\ln n}{2n}} < \frac{\epsilon}{16}
\end{aligned}$$

for $n \geq N_2$, for some N_2 sufficiently large. Here we use $\delta_n = 1/n^2$ and $\ln|\mathcal{H}_{k_n}| = o(n/\ln n)$. Then

$$\sum_{n=1}^\infty \mathbf{P}(\Omega_{13}^n \cap \Theta^n) \leq N_2.$$

Furthermore, by the uniform deviation bound,

$$\sum_{n=1}^\infty \mathbf{P}(\Omega_{11}^n \cap \Theta^n) \leq N_2 + \sum_{n>N_2} \frac{1}{n^2} < \infty.$$

Now consider the event Ω_2^n. Applying Lemma 1 with $\nu = \epsilon/2$, we have that

$$\sum_{n=1}^\infty \mathbf{P}(\Omega_2^n \cap \Theta_n) \leq N_1 + \sum_{n>N_1} \frac{1}{n^2} < \infty.$$

□

In the definitions of $\mathcal{R}_D(h)$ and $\mathcal{R}_{ND}(h)$, we define these quantities to be infinity when the conditioning event has probability zero (see Introduction). This forces the globally optimal classifier to have both $\mathbf{P}(h(X)=1)>0$ and $\mathbf{P}(h(X)=0)>0$ whenever possible. The same property would hold provided $\mathcal{R}_D(h)>(1+\lambda)/\lambda$ when $\mathbf{P}(h(X)=1)=0$ and $\mathcal{R}_{ND}(h)>1+\lambda$ when $\mathbf{P}(h(X)=0)=0$. Were we to define FDR or FNDR to be smaller, our consistency argument would not apply universally. In particular, it might fail for distributions where the global minimizer of $\mathcal{E}_\lambda$ has either $\mathbf{P}(h(X)=0)=0$ or $\mathbf{P}(h(X)=0)=1$, such as when X is deterministic. In a preliminary version of this work, we defined $\mathcal{R}_D(h)$ and $\mathcal{R}_{ND}(h)$ to be zero when the conditioning event is a null event, and were able to prove consistency under a very mild condition on the underlying distribution [31].

5. Variance-Based Bounds and the Zero-Error Case

The previous bounds are based on Hoeffding's inequality. However, a variety of other inequalities exist for sums of bounded random variables. In this section we explore bounds based on Bernstein's inequality, which often allows for tighter bounds through the incorporation of variance information.

For simplicity let us assume that $\mathcal{H}$ is finite with a uniform prior. The results below extend easily to countable $\mathcal{H}$, where $\ln\frac{|\mathcal{H}|}{\delta}$ is replaced by $[\![h]\!]\ln 2+\ln\frac{1}{\delta}$. We denote

$$\widehat{\mathcal{E}}_\lambda(h)=\widehat{\mathcal{R}}_{ND}(h)+\lambda\widehat{\mathcal{R}}_D(h).$$

Theorem 3. *With probability at least $1-\delta$ with respect to the draw of the training data,*

$$\mathcal{E}_\lambda(h)\le\widehat{\mathcal{E}}_\lambda(h)+\psi(h,\delta,n_D(h,Z^n)) \tag{6}$$

for all $h\in\mathcal{H}$, where

$$\psi(h,\delta,k):=\begin{cases}\sqrt{2\mathcal{E}_\lambda(h)\ln\frac{|\mathcal{H}|}{\delta}\left(\frac{1}{n-k}+\frac{\lambda}{k}\right)}+\frac{2\ln\frac{|\mathcal{H}|}{\delta}}{3}\left(\frac{1}{n-k}+\frac{\lambda}{k}\right), & 0<k<n,\\ \infty, & k=0,n.\end{cases}$$

Proof. We recall Bernstein's inequality [6]:

Lemma 3. *Let $Z_1,\dots,Z_n$ be independent zero-mean random variables bounded in absolute value by c. Denote $S_n=\sum_{i=1}^n Z_i$ and $\sigma^2=\frac{1}{n}\sum_{i=1}^n Var\{Z_i\}$. For any $t>0$,*

$$\mathbf{P}(S_n>t)\le\exp\left(-\frac{t^2}{2n\sigma^2+2ct/3}\right).$$

By the quadratic formula, we may invert the relationship between the bound t and the confidence δ. Thus, for $0<\delta<1$, we have that

$$S_n\le\sqrt{2n\sigma^2\ln\frac{1}{\delta}}+\frac{2c\ln\frac{1}{\delta}}{3}$$

with probability at least $1-\delta$.

Consider a fixed $h \in \mathcal{H}$. Recall the following notation from the proof of Theorem 1. Let $V = (V_1, \ldots, V_n) \in \{0,1\}^n$ be a binary indicator vector, with $V_i = h(X_i)$. Let $\mathcal{V}_k$ denote the set of all $v = (v_1, \ldots, v_n) \in \{0,1\}^n$ such that $\sum_{i=1}^n v_i = k$. Since the bound holds trivially for $k = 0, n$, we have

$$\begin{aligned}
&\mathbf{P}(\mathcal{E}_\lambda(h) - \widehat{\mathcal{E}}_\lambda(h) > \psi(h, \delta, n_D(h, Z^n))) \\
&\quad \leq \sum_{k=1}^{n-1} \sum_{v \in \mathcal{V}_k} \mathbf{P}(\mathcal{E}_\lambda(h) - \widehat{\mathcal{E}}_\lambda(h) > \psi(h, \delta, k) | V = v) \mathbf{P}(V = v).
\end{aligned}$$

Introduce

$$\begin{aligned}
Z_i &= \frac{1}{n-k} \left[\mathcal{R}_{ND}(h) \mathbf{1}_{\{h(X_i)=0\}} - \mathbf{1}_{\{Y_i=1, h(X_i)=0\}} \right] \\
&\quad + \frac{\lambda}{k} \left[\mathcal{R}_D(h) \mathbf{1}_{\{h(X_i)=1\}} - \mathbf{1}_{\{Y_i=0, h(X_i)=1\}} \right].
\end{aligned}$$

Note that, if we condition on $V = v$, then

$$\mathcal{E}_\lambda(h) - \widehat{\mathcal{E}}_\lambda(h) = \sum_{i=1}^n Z_i$$

and for $v \in \mathcal{V}_k$, $E\{Z_i | V = v\} = 0$ and

$$|Z_i| \leq \frac{1}{n-k} + \lambda \frac{1}{k}.$$

In addition, the conditional variance term is

$$\begin{aligned}
\sigma_v^2 &= \frac{1}{n} \sum_{i=1}^n \mathrm{Var}\{Z_i | V = v\} \\
&= \frac{1}{n} \left\{ \sum_{i : v_i = 0} \frac{1}{(n-k)^2} E\left[(\mathcal{R}_{ND}(h) - \mathbf{1}_{\{Y_i=1\}})^2 \Big| h(X_i) = 0 \right] \right. \\
&\qquad \left. + \sum_{i : v_i = 1} \frac{\lambda^2}{k^2} E\left[(\mathcal{R}_D(h) - \mathbf{1}_{\{Y_i=0\}})^2 \Big| h(X_i) = 1 \right] \right\} \\
&\leq \frac{1}{n} \left[\frac{1}{n-k} \mathcal{R}_{ND}(h) + \frac{\lambda^2}{k} \mathcal{R}_D(h) \right] \\
&\leq \frac{1}{n} \mathcal{E}_\lambda(h) \left(\frac{1}{n-k} + \frac{\lambda}{k} \right).
\end{aligned}$$

By Bernstein's inequality,

$$\begin{aligned}
&\sum_{k=1}^{n-1} \sum_{v \in \mathcal{V}_k} \mathbf{P}(\mathcal{E}_\lambda(h) - \widehat{\mathcal{E}}_\lambda(h) > \psi(h, \delta, k) | V = v) \mathbf{P}(V = v) \\
&\qquad \leq \sum_{k=1}^{n-1} \frac{\delta}{|\mathcal{H}|} \mathbf{P}(V = v) \leq \frac{\delta}{|\mathcal{H}|}.
\end{aligned}$$

The result now follows by the union bound over all $h \in \mathcal{H}$. □

The bound can be tighter than the one based on Hoeffding's inequality in Theorem 1. However, it cannot be computed in general because it depends on $\mathcal{E}_\lambda(h)$. In the zero-error case, however, it can be used to derive faster rates, as we now describe.

Define the empirical error minimizer

$$\widehat{h} := \arg\min_{h \in \mathcal{H}} \widehat{\mathcal{E}}_\lambda(h).$$

If there are multiple $h \in \mathcal{H}$ achieving the minimum, we assume $\widehat{h}_\lambda$ has both $n_D(\widehat{h}_\lambda) > 0$ and $n_{ND}(\widehat{h}_\lambda) > 0$ if possible.

Corollary 2. *Suppose there exists h^* in $\mathcal{H}$ such that $\mathcal{E}_\lambda(h^*) = 0$. With probability at least $1 - \delta$, if $0 < \sum_{i=1}^n Y_i < n$, then*

$$\mathcal{E}_\lambda(\widehat{h}) \leq 4 \ln \frac{|\mathcal{H}|}{\delta} \left(\frac{1}{n_{ND}(\widehat{h}, Z^n)} + \frac{\lambda}{n_D(\widehat{h}, Z^n)} \right).$$

Proof. Since h^* has $\mathcal{E}_\lambda(h^*) = 0$, it classifies every point correctly. This implies $\widehat{\mathcal{E}}_\lambda(h^*) = 0$ and therefore $\widehat{\mathcal{E}}_\lambda(\widehat{h}) = 0$. Since $0 < \sum_{i=1}^n Y_i < n$, we deduce $n_{ND}(\widehat{h}, Z^n) > 0$ and $n_D(\widehat{h}, Z^n) > 0$. By Theorem 3,

$$\mathcal{E}_\lambda(\widehat{h}) \leq \sqrt{2\mathcal{E}_\lambda(\widehat{h}) \cdot B} + \frac{2}{3} B,$$

where $B = \ln \frac{|\mathcal{H}|}{\delta} \left(\frac{1}{n_{ND}(\widehat{h}, Z^n)} + \frac{\lambda}{n_D(\widehat{h}, Z^n)} \right)$. We would like to show

$$\mathcal{E}_\lambda(\widehat{h}) \leq cB$$

for a universal constant c. For now, let c be fixed but arbitrary. Suppose $\mathcal{E}_\lambda(\widehat{h}) > cB$. Then

$$\sqrt{\mathcal{E}_\lambda(\widehat{h})} \leq \sqrt{2B} + \frac{2}{3} \frac{B}{\sqrt{\mathcal{E}_\lambda(\widehat{h})}}.$$

Squaring both sides we have

$$\begin{aligned} \mathcal{E}_\lambda(\widehat{h}) &\leq B \left(2 + \frac{4\sqrt{2}}{3} \sqrt{\frac{B}{\mathcal{E}_\lambda(\widehat{h})}} + \frac{4}{9} \frac{B}{\mathcal{E}_\lambda(\widehat{h})} \right) \\ &\leq B \left(2 + \frac{4\sqrt{2}}{3} \frac{1}{\sqrt{c}} + \frac{4}{9} \frac{1}{c} \right) \\ &= c' B. \end{aligned}$$

When $c = 4$, $c' < 4$, contradicting the assumption that $\mathcal{E}_\lambda(\widehat{h}) > cB$. □

There has been recent interest in the learning theory literature in generalizing such fast rates to settings beyond the zero-error case. The general strategy is to apply variance-based concentration inequalities to so-called "relative loss" classes. Then, conditions such as Tsybakov's noise condition can be used to derive faster rates [36,9]. This approach can in principle be applied here, but the development of appropriate "noise conditions" is likely to be more challenging. To extend our arguments to a relative loss class, we would need to condition on both $V = (h(X_1), \ldots, h(X_n))$ and $V^* = (h^*(X_1), \ldots, h^*(X_n))$. This would result in quantities such as $P(Y_i = 1 | h(X_i) = 0, h^*(X_i) = 1)$ appearing in the variance term. The bound would also involve the quantities $n_{ND}(h^*, Z^n)$ and $n_D(h^*, Z^n)$, which are unknown in general.

6. Experimental Comparison of Bounds

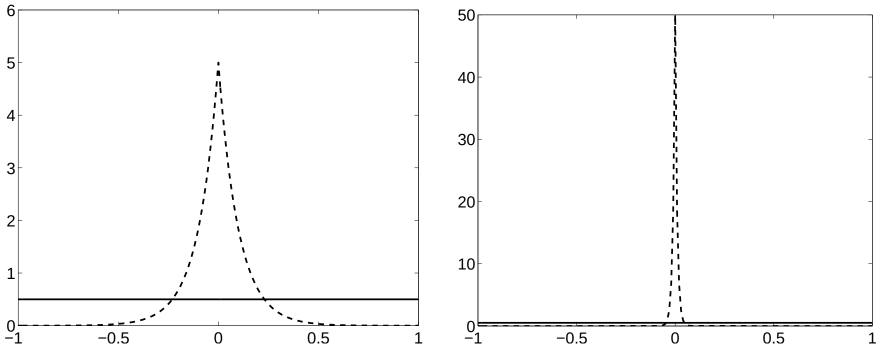
In this section we present a synthetic data experiment to illustrate and compare the two bounding techniques considered in this work, i.e., those based on Hoeffding's and Bernstein's inequalities. These experiments shed light on the tightness of these bounds in terms of the data and various properties of the underlying distributions.

We consider the following joint distribution on (X, Y), where $X \in [-1, 1]$. The a priori probability that $Y = 1$ is denoted p. The conditional distribution of X given $Y = 0$ is uniform on $[-1, 1]$. The conditional distribution of X given $Y = 1$ is a truncated Laplacian distribution, with density $f_1(x) \propto e^{-|x|/\beta}$, $-1 \leq x \leq 1$, where β determines the variance. This model is similar to those used to analyze differential expression in microarray studies [19], although our choices here are guided primarily by tractability considerations. $\mathcal{H}$ is the set of classifiers $h_t = \mathbf{1}_{[-t,t]}$, where $t \in \{\frac{1}{M+1}, \frac{2}{M+1}, \ldots, \frac{M}{M+1}\}$. The endpoints $t = 0$ and $t = 1$ are not included because these classifiers make no discoveries or nondiscoveries, respectively, leading to uninteresting bounds. Thus $|\mathcal{H}| = M$. With this model, $\mathcal{E}_\lambda$ can be calculated analytically, and the difficulty of the problem (i.e., the minimum value of $\mathcal{E}_\lambda$) can be controlled by the parameter β.

To compare the bounds we generated a random sample of size $n = 1000$. We set $\lambda = 2$ and $\delta = 0.1$. M was fixed at 100. We considered 6 different experimental settings, corresponding to $p \in \{0.2, 0.5, 0.8\}$ and $\beta \in \{0.1, 0.01\}$. We generated a single realization of the data for each setting, and computed the bounds for each $h \in \mathcal{H}$. To be fair, we compared the Bernstein-based bound in (6) to a one-sided Hoeffding-based bound derived expressly for $\mathcal{E}_\lambda$. This is slightly tighter than the bound of Corollary 1, which follows from summing bounds on $\mathcal{R}_{ND}$ and $\mathcal{R}_D$. Following the same conditioning argument as in Theorems 1 and 3, and employing a general form of Hoeffding's inequality as stated, e.g., in [16], this bound states that with probability at least $1 - \delta$,

$$\mathcal{E}_\lambda(h) \leq \widehat{\mathcal{E}}_\lambda(h) + \sqrt{\frac{1}{2} \ln \frac{|\mathcal{H}|}{\delta} \left(\frac{1}{n_{ND}(h, Z^n)} + \frac{\lambda^2}{n_D(h, Z^n)} \right)}$$

for all $h \in \mathcal{H}$ such that $0 < n_D(h, Z^n) < n$. Note that applying the inequality $\sqrt{a + b} \leq \sqrt{a} + \sqrt{b}$ recovers the bound of Corollary 1.



(a) $\beta = 0.1$

(b) $\beta = 0.01$

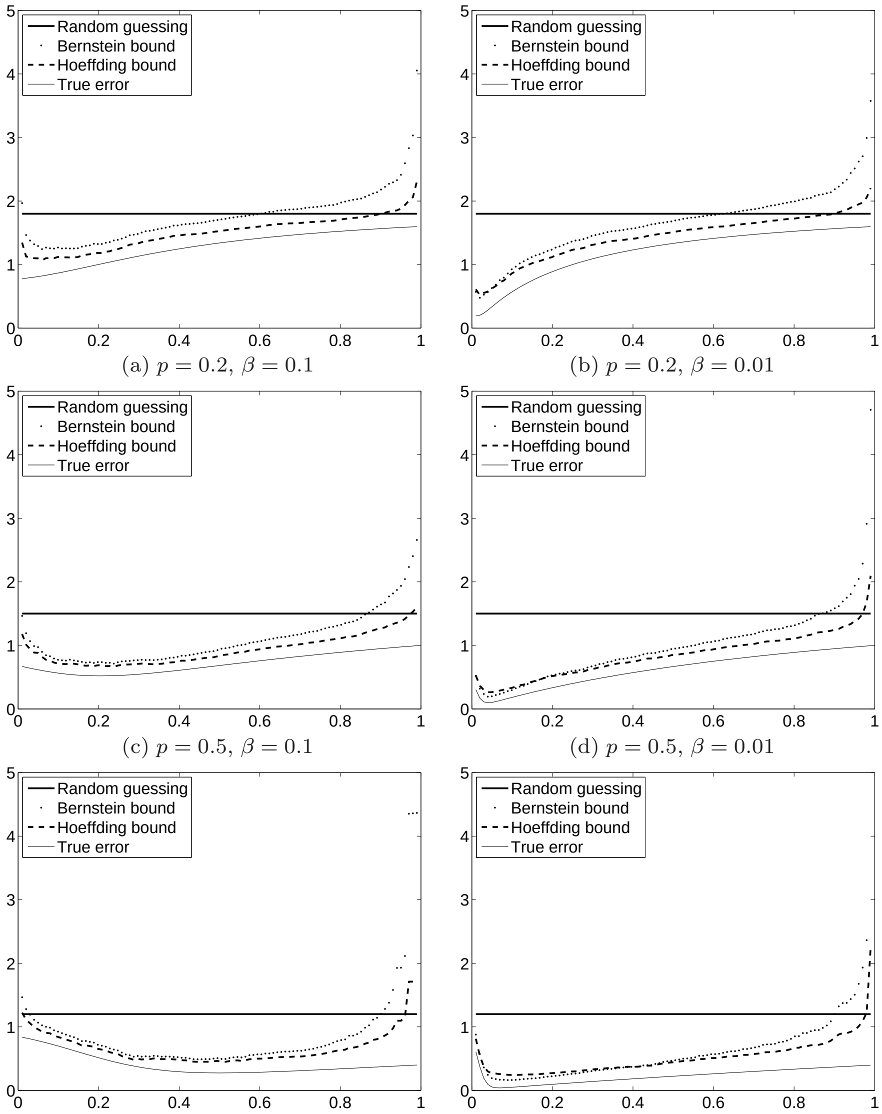
Fig 1. *The class 0 (uniform) and class 1 (truncated Laplacian) distributions used in the simulation study. The variance parameter β of the truncated Laplacian was taken to be $\beta = 0.1$ (a) and $\beta = 0.01$ (b) in the experiments.*

The results are shown in Figure 2. In each plot, the horizontal axis is the t in h_t, and the vertical axis represents the error $\mathcal{E}_\lambda$. For comparison, we also show the performance of the classifier that randomly guesses the class label. The value of $\mathcal{E}_\lambda$ in this case is $p + \lambda(1 - p)$, which is analogous to the value $\frac{1}{2}$ in standard classification. Although our bounds are data-dependent, the general shapes and relative magnitudes of the bounds are fairly stable across realizations.

We first make the following qualitative observations. Both bounds are looser near the endpoints, reflecting the small number of false discoveries or nondiscoveries, which appear as denominators in the bounds. When p is small (meaning fewer class 1 examples in the training data), the bounds are looser at the left endpoint. Similarly, when p is large, the bounds are looser at the right endpoint. As p increases, the optimal “rejection region” grows, as expected.

Regarding performance, for each parameter combination, the bound minimum is always below the random guessing line, meaning the bound is non-trivial for the bound-minimizing learning rule. To assess the dependence of the property on n, we also reran the experiments for $n = 100$, and found that when $\beta = 0.01$, this property still held, but when $\beta = 0.1$, it only held for the Hoeffding-based bounds.

By way of comparison, the bound minimizers tend to be near the true minimizer of $\mathcal{E}_\lambda$. When $\beta = 0.1$, the Hoeffding bound is at least as tight as the Bernstein bound, and typically tighter. When $\beta = 0.01$, $\mathcal{E}_\lambda$ is small enough that the Bernstein bound has a smaller minimum. This again conforms to our expectations. Finally, we simply recall that the Hoeffding-based bound is computable in practice, but the Bernstein-based bound is not, since it depends on $\mathcal{E}_\lambda$. A computable version of the Bernstein bound would necessarily be larger.



(a) $p = 0.2$, $\beta = 0.1$

(b) $p = 0.2$, $\beta = 0.01$

(c) $p = 0.5$, $\beta = 0.1$

(d) $p = 0.5$, $\beta = 0.01$

(e) $p = 0.8$, $\beta = 0.1$

(f) $p = 0.8$, $\beta = 0.01$

Fig 2. *Experimental comparison of Hoeffding- and Bernstein-based bounds for the data described in Section 6. The horizontal axis is the value t corresponding to the classifier $h_t = \mathbf{1}_{[-t,t]}$, and the vertical axis is $\mathcal{E}_\lambda$. Here p is the prior probability of class 1, and β determines the variance of class 1. Smaller β means less overlap between classes.*

7. Constraining FDR

In this section we apply Theorem 1 to analyze a rule that seeks to minimize the FNDR subject to the constraint that FDR $\leq \alpha$, where α is a user-defined significance level. In fact, we first present a more general result, and then deduce results for this and other constrained learning problems as corollaries. Special cases of this bound have been presented previously in different settings [10, 29, 30], but the proof presented here is much simpler.

Thus, let $\mathcal{H}$ be a collection of classifiers as before, but not necessarily finite. Let $\mathcal{R}_0$ and $\mathcal{R}_1$ be measures of Type I and Type II error. For example, these may be FDR and FNDR, false positive rate and false negative rate, or some combination thereof. Assume that for $i = 0, 1$, there exist a data-based estimate $\widehat{\mathcal{R}}_i$ of $\mathcal{R}_i$, and a penalty $\phi_i(h, \delta)$, which define a symmetric confidence interval for $\mathcal{R}_i$. That is, suppose that for any $0 < \delta < 1$,

$$\mathbf{P}_{Z^n}(\sup_{h \in \mathcal{H}} [|\mathcal{R}_i(h) - \widehat{\mathcal{R}}_i(h)| - \phi_i(h, \delta)] > 0) \leq \delta.$$

For $0 < \alpha < 1$ define

$$\begin{aligned} h^*_{\mathcal{H},\alpha} &= \arg\min_{h \in \mathcal{H}} \ \mathcal{R}_1(h) \\ &\quad \text{s. t. } \mathcal{R}_0(h) \leq \alpha. \end{aligned}$$

Consider the learning rule

$$\begin{aligned} \widehat{h}_{\mathcal{H},\alpha} &= \arg\min_{h \in \mathcal{H}} \ \widehat{\mathcal{R}}_1(h) + \phi_1(h, \delta) \\ &\quad \text{s. t. } \widehat{\mathcal{R}}_0(h) \leq \alpha + \phi_0(h, \delta). \end{aligned} \tag{7}$$

Theorem 4. *The learning rule defined in Eqn. (7) is such that, for any $\delta > 0$ and any $n \geq 1$, with probability at least $1 - 2\delta$ with respect to the draw of the training data,*

$$\mathcal{R}_1(\widehat{h}_{\mathcal{H},\alpha}) \leq \mathcal{R}_1(h^*_{\mathcal{H},\alpha}) + 2\phi_1(h^*_{\mathcal{H},\alpha}, \delta)$$

and

$$\mathcal{R}_0(\widehat{h}_{\mathcal{H},\alpha}) \leq \alpha + 2\phi_0(\widehat{h}_{\mathcal{H},\alpha}, \delta)$$

for all $0 < \alpha < 1$ simultaneously.

The result holds regardless of the data-generating distribution. Note that the bounds for the two errors are different in that one depends on a theoretical classifier, while the other involves an empirical classifier. Nonetheless, the bounds are still quite powerful, and lead to strongly consistent learning procedures in several different settings. For example, when applied to Neyman-Pearson classification [29] (see below) and learning minimum volume sets [30], strong consistency has been demonstrated. In these settings, the classifiers $h^*_{\mathcal{H},\alpha}$ and $\widehat{h}_{\mathcal{H},\alpha}$ both belong to the same VC class, in which case ϕ_0 takes the same value for all $h \in \mathcal{H}$, and likewise for ϕ_1. Below we mention how the same general bound gives rise to a strongly consistent rule for FDR/FNDR.

The reader may also note that from the bound, the Type I error may exceed the desired constraint. In special cases, we have previously developed extensions of this result that allow some or all of the slack in the Type I error bound to be transferred to the Type II error bound [30, 27]. Such analysis should also be possible in the general setting presented here.

Proof. Assume that both

$$|\mathcal{R}_0(h) - \widehat{\mathcal{R}}_0(h)| \leq \phi_0(h, \delta) \quad \text{for all } h \in \mathcal{H} \tag{8}$$

and

$$|\mathcal{R}_1(h) - \widehat{\mathcal{R}}_1(h)| \leq \phi_1(h, \delta) \quad \text{for all } h \in \mathcal{H}, \tag{9}$$

which, by assumption, occurs with probability at least $1-2\delta$. By (8), we deduce the second half of the theorem from

$$\mathcal{R}_0(\widehat{h}_{\mathcal{H},\alpha}) \leq \widehat{\mathcal{R}}_0(\widehat{h}_{\mathcal{H},\alpha}) + \phi_0(\widehat{h}_{\mathcal{H},\alpha}, \delta) \leq \alpha + 2\phi_0(\widehat{h}_{\mathcal{H},\alpha}, \delta),$$

where the second inequality follows from $\widehat{\mathcal{R}}_0(\widehat{h}_{\mathcal{H},\alpha}) \leq \alpha + \phi_0(\widehat{h}_{\mathcal{H},\alpha}, \delta)$, which follows from the definition of $\widehat{h}_{\mathcal{H},\alpha}$. To get the first half of the theorem, observe that $\widehat{\mathcal{R}}_0(h^*_{\mathcal{H},\alpha}) \leq \mathcal{R}_0(h^*_{\mathcal{H},\alpha}) + \phi_0(h^*_{\mathcal{H},\alpha}, \delta) \leq \alpha + \phi_0(h^*_{\mathcal{H},\alpha}, \delta)$. Therefore, $h^*_{\mathcal{H},\alpha}$ is among the candidates in the minimization defining $\widehat{h}_{\mathcal{H},\alpha}$. Then

$$\begin{aligned} \mathcal{R}_1(\widehat{h}_{\mathcal{H},\alpha}) &\leq \widehat{\mathcal{R}}_1(\widehat{h}_{\mathcal{H},\alpha}) + \phi_1(\widehat{h}_{\mathcal{H},\alpha}, \delta) \\ &\leq \widehat{\mathcal{R}}_1(h^*_{\mathcal{H},\alpha}) + \phi_1(h^*_{\mathcal{H},\alpha}, \delta) \\ &\leq \mathcal{R}_1(h^*_{\mathcal{H},\alpha}) + 2\phi_1(h^*_{\mathcal{H},\alpha}, \delta). \end{aligned}$$

□

Theorem 4 can immediately be combined with Theorem 1 to give performance guarantees for the case $\mathcal{R}_0(h) = \mathcal{R}_D(h)$ and $\mathcal{R}_1(h) = \mathcal{R}_{ND}(h)$, for a countable class $\mathcal{H}$. In particular, define the rule

$$\begin{aligned} \widehat{h}_{\mathcal{H},\alpha} &= \arg\min_{h \in \mathcal{H}} \widehat{\mathcal{R}}_{ND}(h) + \phi_{ND}(h, \delta) \\ &\quad \text{s. t. } \widehat{\mathcal{R}}_D(h) \leq \alpha + \phi_D(h, \delta). \end{aligned} \tag{10}$$

We have the following.

Corollary 3. *Assume $\mathcal{H}$ is countable. For any $\delta > 0$ and any $n \geq 1$, with probability at least $1 - 2\delta$ with respect to the draw of the training data, the learning rule in (10) satisfies*

$$\mathcal{R}_{ND}(\widehat{h}_{\mathcal{H},\alpha}) \leq \mathcal{R}_{ND}(h^*_{\mathcal{H},\alpha}) + 2\phi_{ND}(h^*_{\mathcal{H},\alpha}, \delta)$$

and

$$\mathcal{R}_D(\widehat{h}_{\mathcal{H},\alpha}) \leq \alpha + 2\phi_D(\widehat{h}_{\mathcal{H},\alpha}, \delta)$$

for all $0 < \alpha < 1$ simultaneously.

To extend such a result to a universally consistent estimator, based on the discussion of Theorem 2, it would be necessary to take $\mathcal{H}$ growing with the sample size n, and to exclude classifiers making too few discoveries or nondiscoveries. The details are similar to those of Section 4, and a formal development is omitted.

7.1. Neyman-Pearson Classification

If we take $\mathcal{R}_0$ and $\mathcal{R}_1$ to be the false positive rate and false negative rate, respectively, we may apply Theorem 4 to recover and generalize known results for Neyman-Pearson classification [10, 29]. Specifically, set

$$\begin{aligned} \mathcal{R}_{FP}(h) &:= \mathbf{P}(h(X)=1 \,|\, Y=0) \\ \mathcal{R}_{FN}(h) &:= \mathbf{P}(h(X)=0 \,|\, Y=1). \end{aligned}$$

There are several possible penalties that provide uniform bounds on the deviation between these quantities and their natural empirical estimates,

$$\begin{aligned} \widehat{\mathcal{R}}_{FP}(h) &:= \begin{cases} \frac{1}{n_0}\sum_{i=1}^n \mathbf{1}_{\{Y_i=0, h(X_i)=1\}}, & n_0>0, \\ 0, & n_0=0, \end{cases} \\ \widehat{\mathcal{R}}_{FN}(h) &:= \begin{cases} \frac{1}{n_1}\sum_{i=1}^n \mathbf{1}_{\{Y_i=1, h(X_i)=0\}}, & n_1>0, \\ 0, & n_1=0, \end{cases} \end{aligned}$$

where $n_j := \sum_{i=1}^n \mathbf{1}_{\{Y_i=j\}}$. Examples of such penalties (e.g., VC and Rademacher penalties) are given in [27]. As a concrete example, we state a result here for the case of countable $\mathcal{H}$. Thus define the penalties

$$\begin{aligned} \phi_{FP}(h,\delta) &= \begin{cases} \sqrt{\frac{[\![h]\!]\ln 2+\ln(2/\delta)}{2n_0}}, & n_0>0, \\ 1, & n_0=0, \end{cases} \\ \phi_{FN}(h,\delta) &= \begin{cases} \sqrt{\frac{[\![h]\!]\ln 2+\ln(2/\delta)}{2n_1}}, & n_1>0, \\ 1, & n_1=0. \end{cases} \end{aligned}$$

Define the rule

$$\begin{aligned} \widehat{h}_{\mathcal{H},\alpha} &= \underset{h\in\mathcal{H}}{\arg\min}\ \widehat{\mathcal{R}}_{FN}(h)+\phi_{FN}(h,\delta) \\ &\quad\ \text{s. t. } \widehat{\mathcal{R}}_{FP}(h)\le \alpha+\phi_{FP}(h,\delta). \end{aligned} \tag{11}$$

We have the following.

Corollary 4. *Assume $\mathcal{H}$ is countable. For any $\delta>0$ and any $n\ge 1$, with probability at least $1-2\delta$ with respect to the draw of the training data, the learning rule in (11) satisfies*

$$\mathcal{R}_{FN}(\widehat{h}_{\mathcal{H},\alpha}) \le \mathcal{R}_{FN}(h^*_{\mathcal{H},\alpha}) + 2\phi_{FN}(h^*_{\mathcal{H},\alpha},\delta)$$

and

$$\mathcal{R}_{FP}(\widehat{h}_{\mathcal{H},\alpha}) \le \alpha + 2\phi_{FP}(\widehat{h}_{\mathcal{H},\alpha},\delta)$$

for all $0<\alpha<1$ simultaneously.

We note that Theorem 4, applied in the context of Neyman-Pearson classification, is a stronger result than those in [10, 29], which do not explicitly allow penalties that depend on the classifier h.

7.2. Precision and Recall

As a final application of Theorem 4, we analyze the precision and recall performance measures, common in database information retrieval problems (see Introduction). Precision and recall can both be defined in terms of quantities already discussed. Denote the precision

$$\mathcal{Q}_{PR}(h) := \mathbf{P}(Y=1 \,|\, h(X)=1) := 1 - \mathcal{R}_D(h)$$

and the recall

$$\mathcal{Q}_{RE}(h) := \mathbf{P}(h(X)=1 \,|\, Y=1) = 1 - \mathcal{R}_{FN}(h),$$

and let $\widehat{\mathcal{Q}}_{PR}(h) := 1 - \widehat{\mathcal{R}}_D(h)$ and $\widehat{\mathcal{Q}}_{RE}(h) := 1 - \widehat{\mathcal{R}}_{FN}(h)$ be the empirical estimates. In this setting the goal is to find the classifier with the largest precision, while maintaining a recall of at least β, where β is a user-specified level. Thus the optimal classifier in a given class $\mathcal{H}$ is

$$\begin{aligned} h^*_{\mathcal{H},\beta} \quad = \quad & \arg\max_{h\in\mathcal{H}} \ \mathcal{Q}_{PR}(h) \\ & \text{s. t. } \mathcal{Q}_{RE}(h) \geq \beta. \end{aligned}$$

Define the rule

$$\begin{aligned} \widehat{h}_{\mathcal{H},\beta} \quad = \quad & \arg\max_{h\in\mathcal{H}} \ \widehat{\mathcal{Q}}_{PR}(h) - \phi_D(h,\delta) \\ & \text{s. t. } \widehat{\mathcal{Q}}_{RE}(h) \geq \beta - \phi_{FN}(h,\delta). \end{aligned} \tag{12}$$

We have the following.

Corollary 5. *Assume $\mathcal{H}$ is countable. For any $\delta > 0$ and any $n \geq 1$, with probability at least $1 - 2\delta$ with respect to the draw of the training data, the learning rule in (12) satisfies*

$$\mathcal{Q}_{PR}(\widehat{h}_{\mathcal{H},\beta}) \geq \mathcal{Q}_{PR}(h^*_{\mathcal{H},\beta}) - 2\phi_D(h^*_{\mathcal{H},\beta}, \delta)$$

and

$$\mathcal{Q}_{RE}(\widehat{h}_{\mathcal{H},\beta}) \geq \beta - 2\phi_{FN}(\widehat{h}_{\mathcal{H},\beta}, \delta)$$

for all $0 < \beta < 1$ simultaneously

Proof. To apply Theorem 4, note that maximizing $\mathcal{Q}_{PR}(h)$ is equivalent to minimizing $\mathcal{R}_D(h)$, that the constraint $\mathcal{Q}_{RE}(h) \geq \beta$ is equivalent to $\mathcal{R}_{FN}(h) \leq \alpha := 1 - \beta$, and similarly for the empirical objective and constraint. Furthermore, since $|\mathcal{Q}_{PR}(h) - \widehat{\mathcal{Q}}_{PR}(h)| = |\mathcal{R}_D(h) - \widehat{\mathcal{R}}_D(h)|$, and $|\mathcal{Q}_{RE}(h) - \widehat{\mathcal{Q}}_{RE}(h)| = |\mathcal{R}_{FN}(h) - \widehat{\mathcal{R}}_{FN}(h)|$, we have that the assumptions of Theorem 4 are satisfied with the stated penalties. □

8. Conclusion

This paper demonstrates that FDR and FNDR control is possible in the context of statistical learning theory, where the distribution of (X, Y) is unknown except through training data. We develop empirical estimates of these quantities and derive uniform deviation bounds which assess the closeness of these empirical estimates to the true FDR and FNDR. Unlike most other performance measures in statistical learning theory, which are related to binomial random variables, the FDR and FNDR measures are related to ratios of binomial random variables, which requires the development of novel bounding techniques. A bound based on Hoeffding's inequality is shown to lead to a universally consistent rule, while a bound based on Bernstein's inequality is shown to lead to a fast rate of convergence in the zero-error case. We also present a general result for learning subject to a Type I error constraint, and apply this to FDR as well as other criteria.

Extending our results to uncountable classes $\mathcal{H}$ is an interesting open question, and may require the development of new techniques. The standard proofs of common generalization error bounds for uncountable classes, such as Rademacher and VC penalties, rely on the introduction of an artificial "ghost" sample [15]. That technique would require every $h \in \mathcal{H}$ to have the same empirical number of discoveries (or nondiscoveries) on both the original and ghost samples, which is generally not the case. Recently El-Yaniv and Pechyony [20] have extended the ghost sample technique to cases where the training and ghost samples have different sizes (their results are stated in the context of transductive learning), and some of their arguments may be useful in this regard.

Another interesting open question is how to achieve rates of convergence faster that $n^{-1/2}$ beyond the zero-error case. Our simulations suggest that a variance-based bound (in our case, Bernstein's inequality) can lead to tighter bounds in "low noise" settings, but formalizing this intuition will evidently entail new techniques and conditions in addition to those currently used to derive fast rates.

A third avenue for future work is to consider the transductive setting, where the test points are available at the time of learning, and the goal is to predict their labels.

Acknowledgements

The authors thank the Associate Editor and anonymous referees for their many constructive comments.

References

[1] Agarwal, S., Graepel, T., Herbrich, R., Har-Peled, S., and Roth, D. (2005). Generalization bounds for the area under the ROC curve. *J. Machine Learning Research 6*, 393–425. MR2249826

[2] Altman, D. G. and Bland, J. M. (1994). Diagnostic tests 2: predictive values. *Brit. Med. J. 309*, 102.

[3] Arlot, S., Blanchard, G., and Roquain, E. (2007). Resampling-based confidence regions and multiple tests for a correlated random vector. In *Learning Theory: 20th Annual Conference on Learning Theory, COLT 2007*, N. H. B. C. Gentile, Ed. Springer-Verlag, Heidelberg, 127–141. MR2397583

[4] Bach, F. R., Heckerman, D., and Horvitz, E. (2006). Considering cost asymmetry in learning classifiers. *J. Machine Learning Research*, 1713–1741. MR2274422

[5] Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Statist. Soc B* **57**, 1, 289–300. MR1325392

[6] Bernstein, S. (1946). *The Theory of Probabilities*. Gastehizdat Publishing House, Moscow.

[7] Beygelzimer, A., Langford, J., and Ravikumar, P. (2009). Error-correcting tournaments. preprint.

[8] Blanchard, G. and Fleuret, F. (2007). Occam's hammer. In *Learning Theory: 20th Annual Conference on Learning Theory, COLT 2007*, N. H. B. C. Gentile, Ed. Springer-Verlag, Heidelberg, 112–126. MR2397582

[9] Bousquet, O., Boucheron, S., and Lugosi, G. (2004). Introduction to statistical learning theory. In *Advanced Lectures in Machine Learning*, O. Bousquet, U. Luxburg, and G. Rätsch, Eds. Springer, 169–207.

[10] Cannon, A., Howse, J., Hush, D., and Scovel, C. (2002). Learning with the Neyman-Pearson and min-max criteria. Tech. Rep. LA-UR 02-2951, Los Alamos National Laboratory.

[11] Chi, Z. and Tan., Z. (2008). Positive false discovery proportions: intrinsic bounds and adaptive control. *Statistica Sinica* **18**, 3, 837–860. MR2440397

[12] Clémenccon, S., Lugosi, G., and Vayatis, N. (2008). Ranking and empirical risk minimization of U-statistics. *Ann. Stat.* **36**, 2, 844–874. MR2396817

[13] Clémencon, S. and Vayatis, N. (2009a). Nonparametric estimation of the precision-recall curve. In to appear*, Proc. 26th International Machine Learning Conference (ICML).*

[14] Clémencon, S. and Vayatis, N. (2009b). Overlaying classifiers: a practical approach for optimal ranking. In *Advances in Neural Information Processing Systems 21*, D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, Eds. 313–320.

[15] Devroye, L., Györfi, L., and Lugosi, G. (1996). *A Probabilistic Theory of Pattern Recognition*. Springer, New York. MR1383093

[16] Devroye, L. and Lugosi, G. (2001). *Combinatorial Methods in Density Estimation*. Springer, New York. MR1843146

[17] Donoho, D. and Jin, J. (2004). Higher criticism for detecting sparse heterogeneous mixtures. *Ann. Stat.* **32**, 3, 962–994. MR2065195

[18] Durrett, R. (1991). *Probability: Theory and Examples.* Wadsworth & Brooks/Cole, Pacific Grove, CA. MR1068527

[19] Efron, B., Tibshirani, R., Storey, J., and Tusher, V. (2001). Empirical Bayes analysis of a microarray experiment. *Journal of the American Statistical Association 96*, 1151–1160. MR1946571

[20] El-Yaniv, R. and Pechyony, D. (2007). Transductive Rademacher complexity and its applications. In *Proc. 20th Annual Conference on Learning Theory, COLT 2007*, N. Bshouty and C. Gentile, Eds. Springer-Verlag, Heidelberg, 157–171. MR2397585

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

[22] Genovese, C. R., Lazar, N. A., and Nichols, T. E. (2002). Thresholding of statistical maps in functional neuroimaging using the false discovery rate. *NeuroImage 15*, 870–878.

[23] Hoeffding, W. (1963). Probability inequalities for sums of bounded random variables. *J. Amer. Statist. Assoc. 58*, 13–30. MR0144363

[24] Liu, B., Lee, W. S., Yu, P. S., and Li, X. (2002). Partially supervised classification of text documents. In *Proc. 19th Int. Conf. Machine Learning (ICML).* Sydney, Australia, 387–394.

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

[26] McAllester, D. (1999). Some PAC-Bayesian theorems. *Machine Learning* **37**, 3, 355–363.

[27] Scott, C. (2007). Performance measures for Neyman-Pearson classification. *IEEE Trans. Inform. Theory* **53**, 8, 2852–2863. MR2400500

[28] Scott, C. and Blanchard, G. (2009). Novelty detection: Unlabeled data definitely help. In *Proceedings of the Twelfth International Conference on Artificial Intelligence and Statistics (AISTATS) 2009* (April 16-18), D. van Dyk and M. Welling, Eds. JMLR: W&CP 5, Clearwater Beach, Florida, 464–471.

[29] Scott, C. and Nowak, R. (2005). A Neyman-Pearson approach to statistical learning. *IEEE Trans. Inform. Theory* **51**, 8, 3806–3819. MR2239000

[30] Scott, C. and Nowak, R. (2006). Learning minimum volume sets. *J. Machine Learning Res. 7*, 665–704. MR2274383

[31] Scott, C. D., Bellala, G., and Willett, R. (2007). Generalization error analysis for FDR controlled classification. In *Proc. IEEE Workshop on Statistical Signal Processing.* Madison, WI.

[32] Soric, B. (1989). Statistical discoveries and effect-size estimation. *J. Amer. Statist. Assoc. 84*, 608–610.

[33] Storey, J. (2002). A direct approach to false discovery rates. *Journal of the Royal Statistical Society, Series B 64*, 479–498. MR1924302

[34] Storey, J. (2003). The positive false discovery rate: A Bayesian interpretation of the q-value. *Annals of Statistics 31:6*, 2013–2035. MR2036398

[35] Storey, J. (2007). The optimal discovery procedure: A new approach to simultaneous significance testing. *Journal of the Royal Statistical Society, Series B 69*, 347–368. MR2323757

[36] Tsybakov, A. B. (2004). Optimal aggregation of classifiers in statistical learning. *Ann. Stat.* **32**, 1, 135–166. MR2051002

[37] Usunier, N., Amini, M., and Gallinari, P. (2005). A data-dependent generalisation error bound for the AUC. In *Proc. ICML 2005 Workshop on ROC Analysis in Machine Learning*.

[38] van Rijsbergen, C. J. (1979). *Information Retrieval*, 2nd ed. Butterworths, London.