Journal of Machine Learning Research 12 (2011) 2831-2855 Submitted 2/11; Revised 7/11; Published 10/11

Neyman-Pearson Classification, Convexity and Stochastic Constraints

Philippe Rigollet RIGOLLET@PRINCETON.EDU
Xin Tong XTONG@PRINCETON.EDU
Department of Operations Research and Financial Engineering
Princeton University
Princeton, NJ 08544, USA

Editor: Gábor Lugosi

Abstract

Motivated by problems of anomaly detection, this paper implements the Neyman-Pearson paradigm to deal with asymmetric errors in binary classification with a convex loss φ. Given a finite collection of classifiers, we combine them and obtain a new classifier that satisfies simultaneously the two following properties with high probability: (i) its φ-type I error is below a pre-specified level and (ii), it has φ-type II error close to the minimum possible. The proposed classifier is obtained by minimizing an empirical convex objective with an empirical convex constraint. The novelty of the method is that the classifier output by this computationally feasible program is shown to satisfy the original constraint on type I error. New techniques to handle such problems are developed and they have consequences on chance constrained programming. We also evaluate the price to pay in terms of type II error for being conservative on type I error.

Keywords: binary classification, Neyman-Pearson paradigm, anomaly detection, empirical constraint, empirical risk minimization, chance constrained optimization

1. Introduction

The Neyman-Pearson (NP) paradigm in statistical learning extends the objective of classical binary classification in that, while the latter focuses on minimizing classification error that is a weighted sum of type I and type II errors, the former minimizes type II error with an upper bound α on type I error. With slight abuse of language, in verbal discussion we do not distinguish type I/II error from probability of type I/II error.

For learning with the NP paradigm, it is essential to avoid one kind of error at the expense of the other. As an illustration, consider the following problem in medical diagnosis: failing to detect a malignant tumor has far more severe consequences than flagging a benign tumor. So it makes sense to put priority on controlling the false negative rate. Other scenarios include spam filtering, machine monitoring, target recognition, etc.

In the learning context, as true errors are inaccessible, we cannot enforce almost surely the desired upper bound for type I error. The best we can hope is that a data dependent classifier has type I error bounded with high probability. Ideally, a good classification rule $\hat{f}$ in NP context should satisfy two properties. The first is that type I error of the classifier $\hat{f}$ is bounded from above by a pre-specified level with pre-specified high probability; the second is that $\hat{f}$ has good performance bounds for excess type II error. As will be illustrated, it is unlikely that these two goals can be fulfilled simultaneously. Following the original spirit of NP paradigm, we put priority on type I error and insist on the pre-specified upper bound α. Our proposed learning procedure meets the

conservative attitude on type I error, and has good performance bound measured by the excess φ-type II error. We also discuss the general consequence of being conservative in NP learning.

The paper is organized as follows. In Section 2, the classical setup for binary classification is reviewed and the main notation is introduced. A parallel between binary classification and statistical hypothesis testing is drawn in Section 3 with emphasis on the NP paradigm in both frameworks. The main propositions and theorems are stated in Section 4 while proofs and technical results are relegated to Appendix A. Finally, Section 5 illustrates an application of our results to chance constrained optimization.

In the rest of the paper, we denote by x_j the j-th coordinate of a vector $x \in \mathbb{R}^d$.

2. Binary Classification

In this section, we review the classical setup of binary classification together with the convexification procedure that we employ throughout the paper. Moreover, we introduce the Neyman-Pearson paradigm in this setup.

2.1 Classification Risk and Classifiers

Let (X,Y) be a random couple where $X \in \mathcal{X} \subset \mathbb{R}^d$ is a vector of covariates and $Y \in \{-1,1\}$ is a label that indicates to which class X belongs. A *classifier* h is a mapping $h : \mathcal{X} \to [-1,1]$ whose sign returns the predicted class given X. An error occurs when $-h(X)Y \geq 0$ and it is therefore natural to define the classification loss by $\mathbb{1}(-h(X)Y \geq 0)$, where $\mathbb{1}(\cdot)$ denotes the indicator function.

The expectation of the classification loss with respect to the joint distribution of (X,Y) is called *(classification) risk* and is defined by

$$R(h) = \mathbb{P}\left(-h(X)Y \geq 0\right).$$

Clearly the indicator function is not convex, and for computational convenience, a common practice is to replace it by a convex surrogate (see, e.g., Bartlett et al., 2006, and references therein).

To this end, we rewrite the risk function as

$$R(h) = \mathbb{E}[\varphi(-h(X)Y)],$$

where $\varphi(z) = \mathbb{1}\left(z \geq 0\right)$. Convex relaxation can be achieved by simply replacing the indicator function by a convex surrogate.

Definition 1 *A function $\varphi : [-1,1] \to \mathbb{R}^+$ is called a* convex surrogate *if it is non-decreasing, continuous and convex and if $\varphi(0) = 1$.*

Commonly used examples of convex surrogates are the hinge loss $\varphi(x) = (1+x)_+$, the logit loss $\varphi(x) = \log_2(1+e^x)$ and the exponential loss $\varphi(x) = e^x$.

For a given choice of φ, define the φ-risk

$$R_\varphi(h) = \mathbb{E}[\varphi(-Yh(X))]\,.$$

Hereafter, we assume that φ is fixed and refer to R_φ as the risk when there is no confusion. In our subsequent analysis, this convex relaxation will also be the ground to analyze a stochastic convex

optimization problem subject to stochastic constraints. A general treatment of such problems can be found in Section 5.

Because of overfitting, it is unreasonable to look for mappings minimizing empirical risk over all classifiers. Indeed, one could have a small empirical risk but a large true risk. Hence, we resort to regularization. There are in general two ways to proceed. The first is to restrict the candidate classifiers to a specific class $\mathcal{H}$, and the second is to change the objective function by, for example, adding a penalty term. The two approaches can be combined, and sometimes they are obviously equivalent.

In this paper, we pursue the first idea by defining the class of candidate classifiers as follows. Let $h_1,\dots,h_M, M \geq 2$ be a given collection of classifiers. In our setup, we allow M to be large. In particular, our results remain asymptotically meaningful as long as $M = o(e^n)$. Such classifiers are usually called base classifiers and can be constructed in a very naive manner. Typical examples include decision stumps or small trees. While the h_j's may have no satisfactory classifying power individually, for over two decades, boosting type of algorithms have successfully exploited the idea that a suitable weighted majority vote among these classifiers may result in low classification risk (Schapire, 1990). Consequently, we restrict our search for classifiers to the set of functions consisting of convex combinations of the h_j's:

$$\mathcal{H}^{\text{conv}} = \{\mathsf{h}_\lambda = \sum_{j=1}^{M} \lambda_j h_j, \lambda \in \Lambda\},$$

where Λ denotes the flat simplex of $\mathbb{R}^M$ and is defined by $\Lambda = \{\lambda \in \mathbb{R}^M : \lambda_j \geq 0, \sum_{j=1}^M \lambda_j = 1\}$. In effect, classification rules given by the sign of $h \in \mathcal{H}^{\text{conv}}$ are exactly the set of rules produced by the weighted majority votes among the base classifiers $h_1,\dots,h_M$.

By restricting our search to classifiers in $\mathcal{H}^{\text{conv}}$, the best attainable φ-risk is called *oracle risk* and is abusively denoted by $R_\varphi(\mathcal{H}^{\text{conv}})$. As a result, we have $R_\varphi(h) \geq R_\varphi(\mathcal{H}^{\text{conv}})$ for any $h \in \mathcal{H}^{\text{conv}}$ and a natural measure of performance for a classifier $h \in \mathcal{H}^{\text{conv}}$ is given by its excess risk defined by $R_\varphi(h) - R_\varphi(\mathcal{H}^{\text{conv}})$.

The excess risk of a data driven classifier h_n is a random quantity and we are interested in bounding it with high probability. Formally, the statistical goal of binary classification is to construct a classifier h_n such that the oracle inequality

$$R_\varphi(h_n) \leq R_\varphi(h_{\mathcal{H}^{\text{conv}}}) + \Delta_n(\mathcal{H}^{\text{conv}}, \delta)$$

holds with probability $1-\delta$, where $\Delta_n(\cdot,\cdot)$ should be as small as possible.

In the scope of this paper, we focus on candidate classifiers in the class $\mathcal{H}^{\text{conv}}$. Some of the following results such as Theorem 3 can be extended to more general classes of classifiers with known complexity such as classes with bounded VC-dimension, as for example in Cannon et al. (2002). However, our main argument for bounding φ-type II error (defined in next subsection) relies on Proposition 4 which, in turn, depends heavily on the convexity of the problem, and it is not clear how it can be extended to more general classes of classifiers.

2.2 The Neyman-Pearson Paradigm

We make the convention that when $h(X) \geq 0$ the predicted class is $+1$, and -1 otherwise. Under this convention, the risk function in classical binary classification can be expressed as a convex combination of type I error $R^-(h) = \mathbb{P}(-Yh(X) \geq 0 | Y = -1)$ and type II error

$R^+(h) = \mathbb{P}(-Yh(X) > 0 | Y = 1)$:

$$R(h) = \mathbb{P}(Y = -1)R^-(h) + \mathbb{P}(Y = 1)R^+(h).$$

While the goal of classical binary classification is $\min_{h\in\mathcal{H}} R(h)$, where $\mathcal{H}$ is the set of candidate classifiers, the NP classification targets on

$$\min_{\substack{h\in\mathcal{H}\\ R^-(h)\le\alpha}} R^+(h)\,.$$

More generally, we can define the φ-type I and φ-type II errors respectively by

$$R^-_\varphi(h) = \mathbb{E}\left[\varphi(-Yh(X))|Y=-1\right] \qquad \text{and} \qquad R^+_\varphi(h) = \mathbb{E}\left[\varphi(-Yh(X))|Y=1\right].$$

Our main theorems concern about $R^-_\varphi(\cdot)$ and $R^+_\varphi(\cdot)$, but we will come back and address how convexification and conservativeness affect $R^-(\cdot)$ and $R^+(\cdot)$.

Following the NP paradigm, for a given class $\mathcal{H}$ of classifiers, we seek to solve the constrained minimization problem:

$$\min_{\substack{h\in\mathcal{H}\\ R^-_\varphi(h)\le\alpha}} R^+_\varphi(h), \tag{1}$$

where $\alpha \in (0,1)$, the significance level, is a constant specified by the user.

NP classification is closely related to the NP approach to statistical hypothesis testing. We now recall a few key concepts about the latter. Many classical works have addressed the theory of statistical hypothesis testing, in particular Lehmann and Romano (2005) provides a thorough treatment of the subject.

Statistical hypothesis testing bears strong resemblance with binary classification if we assume the following model. Let P^- and P^+ be two probability distributions on $\mathcal{X} \subset \mathbb{R}^d$. Let $p \in (0,1)$ and assume that Y is a random variable defined by

$$Y = \begin{cases} 1 & \text{with probability } p\,,\\ -1 & \text{with probability } 1-p\,. \end{cases}$$

Assume further that the conditional distribution of X given Y is given by P^Y. Given such a model, the goal of statistical hypothesis testing is to determine whether X was generated from P^- or P^+. To that end, we construct a test $\phi : \mathcal{X} \to [0,1]$ and the conclusion of the test based on ϕ is that X is generated from P^+ with probability $\phi(X)$ and from P^- with probability $1-\phi(X)$. Note that randomness here comes from an exogenous randomization process such as flipping a biased coin. Two kinds of errors arise: type I error occurs when rejecting P^- when it is true, and type II error occurs when accepting P^- when it is false. The Neyman-Pearson paradigm in hypothesis testing amounts to choosing ϕ that solves the following constrained optimization problem

$$\begin{array}{ll} \text{maximize} & \mathbb{E}[\phi(X)|Y=1]\,,\\ \text{subject to} & \mathbb{E}[\phi(X)|Y=-1] \le \alpha\,, \end{array}$$

where $\alpha \in (0,1)$ is the significance level of the test. In other words, we specify a significance level α on type I error, and minimize type II error. We call a solution to this problem *a most powerful test* of level α. The Neyman-Pearson Lemma gives mild sufficient conditions for the existence of such a test.

Theorem 2 (Neyman-Pearson Lemma) *Let P^- and P^+ be probability distributions possessing densities p^- and p^+ respectively with respect to some measure μ. Let $f_k(x) = \mathbb{1}(L(x) \geq k)$, where the likelihood ratio $L(x) = p^+(x)/p^-(x)$ and k is such that $P^-(L(X) > k) \leq \alpha$ and $P^-(L(X) \geq k) \geq \alpha$. Then,*

- *f_k is a level $\alpha = \mathbb{E}[\varphi_k(X)|Y = -1]$ most powerful test.*
- *For a given level α, the most powerful test of level α is defined by*

$$\phi(X) = \begin{cases} 1 & \text{if} \quad L(X) > k \\ 0 & \text{if} \quad L(X) < k \\ \frac{\alpha - P^-(L(X)>k)}{P^-(L(X)=k)} & \text{if} \quad L(X) = k\,. \end{cases}$$

Notice that in the learning framework, ϕ cannot be computed since it requires the knowledge of the likelihood ratio and of the distributions P^- and P^+. Therefore, it remains merely a theoretical proposition. Nevertheless, the result motivates the NP paradigm pursued here.

3. Neyman-Pearson Classification Via Convex Optimization

Recall that in NP classification with a loss function φ, the goal is to solve the problem (1). This cannot be done directly as conditional distributions P^- and P^+, and hence R^-_φ and R^+_φ, are unknown. In statistical applications, information about these distributions is available through two i.i.d. samples $X^-_1, \ldots, X^-_{n^-}$, $n^- \geq 1$ and $X^+_1, \ldots, X^+_{n^+}$, $n^+ \geq 1$, where $X^-_i \sim P^-, i = 1, \ldots, n^-$ and $X^+_i \sim P^+, i = 1, \ldots, n^+$. We do not assume that the two samples $(X^-_1, \ldots, X^-_{n^-})$ and $(X^+_1, \ldots, X^+_{n^+})$ are mutually independent. Presently the sample sizes n^- and n^+ are assumed to be deterministic and will appear in the subsequent finite sample bounds. A different sampling scheme, where these quantities are random, is investigated in Section 4.3.

3.1 Conservativeness on Type I Error

While the binary classification problem has been extensively studied, theoretical proposition on how to implement the NP paradigm remains scarce. To the best of our knowledge, Cannon et al. (2002) initiated the theoretical treatment of the NP classification paradigm and an early empirical study can be found in Casasent and Chen (2003). The framework of Cannon et al. (2002) is the following. Fix a constant $\varepsilon_0 > 0$ and let $\mathcal{H}$ be a given set of classifiers with finite VC dimension. They study a procedure that consists of solving the following relaxed empirical optimization problem

$$\min_{\substack{h \in \mathcal{H} \\ \hat{R}^-(h) \leq \alpha + \varepsilon_0/2}} \hat{R}^+(h), \tag{2}$$

where

$$\hat{R}^-(h) = \frac{1}{n^-} \sum_{i=1}^{n^-} \mathbb{1}(h(X^-_i) \geq 0)\,, \quad \text{and} \quad \hat{R}^+(h) = \frac{1}{n^+} \sum_{i=1}^{n^+} \mathbb{1}(h(X^-_i) \leq 0)$$

denote the empirical type I and empirical type II errors respectively. Let $\hat{h}$ be a solution to (2). Denote by h^* a solution to the original Neyman-Pearson optimization problem:

$$h^* \in \operatorname*{argmin}_{\substack{h \in \mathcal{H} \\ R^-(h) \leq \alpha}} R^+(h)\,, \tag{3}$$

The main result of Cannon et al. (2002) states that, simultaneously with high probability, the type II error $R^+(\hat{h})$ is bounded from above by $R^+(h^*)+\varepsilon_1$, for some $\varepsilon_1 > 0$ and the type I error of $\hat{h}$ is bounded from above by $\alpha+\varepsilon_0$. In a later paper, Cannon et al. (2003) considers problem (2) for a data-dependent family of classifiers $\mathcal{H}$, and bound estimation errors accordingly. Several results for traditional statistical learning such as PAC bounds or oracle inequalities have been studied in Scott (2005) and Scott and Nowak (2005) in the same framework as the one laid down by Cannon et al. (2002). A noteworthy departure from this setup is Scott (2007) where sensible performance measures for NP classification that go beyond analyzing separately two kinds of errors are introduced. Furthermore, Blanchard et al. (2010) develops a general solution to semi-supervised novelty detection by reducing it to NP classification. Recently, Han et al. (2008) transposed several results of Cannon et al. (2002) and Scott and Nowak (2005) to NP classification with convex loss.

The present work departs from previous literature in our treatment of type I error. In fact, the classifiers in all the papers mentioned above take a compromise on the pre-determined upper bound on type I error, that is, they ensure that $\mathbb{P}(R^-(\hat{h}) > \alpha+\varepsilon_0)$ is small, for some $\varepsilon_0 > 0$. However, it is our primary interest to make sure that $R^-(\hat{h}) \le \alpha$ with high probability, following the original principle of the Neyman-Pearson paradigm that type I error should be controlled by a pre-specified level α. As we follow an empirical risk minimization procedure, to control $\mathbb{P}(R^-(\hat{h}) > \alpha)$, it is necessary to have $\hat{h}$ be a solution to some program with a strengthened constraint on empirical type I error. If our concern is only on type I error, we can just do so. However, we also want to evaluate the excess type II error. Our conservative attitude on type I error faces new technical challenges which we summarize here. In the approach of Cannon et al. (2002) and of Scott and Nowak (2005), the relaxed constraint on the type I error is constructed such that the constraint $\hat{R}^-(h) \le \alpha+\varepsilon_0/2$ on type I error in (2) is satisfied by h^* (defined in (3)) with high probability, and that this classifier accommodates the excess type II error well. As a result, the control of type II error mainly follows as a standard exercise to control suprema of empirical processes. This is not the case here; we have to develop methods to control the optimum value of an optimization problem under a stochastic constraint. Such methods have consequences not only in NP classification but also on chance constraint programming as explained in Section 5.

3.2 Convexified NP Classifier

Concerned about computational feasibility, our proposed classifier is the solution to a convex program, which is an empirical form NP classification problem (1) where the distribution of the observations is unknown. In view of the arguments presented in the previous subsection, we cannot simply replace the unknown risk functions by their empirical counterparts. The treatment of the convex constraint should be done carefully and we proceed as follows.

For any classifier h and a given convex surrogate φ, define $\hat{R}^-_\varphi$ and $\hat{R}^+_\varphi$ to be the empirical counterparts of R^-_φ and R^+_φ respectively by

$$\hat{R}^-_\varphi(h) = \frac{1}{n^-}\sum_{i=1}^{n^-} \varphi(h(X_i^-))\,, \quad \text{and} \quad \hat{R}^+_\varphi(h) = \frac{1}{n^+}\sum_{i=1}^{n^+} \varphi(-h(X_i^+))\,.$$

Moreover, for any $a > 0$, let $\mathcal{H}^{\varphi,a} = \{h \in \mathcal{H}^{\text{conv}} : R^-_\varphi(h) \le a\}$ be the set of classifiers in $\mathcal{H}^{\text{conv}}$ whose convexified type I errors are bounded from above by a, and let $\mathcal{H}^{\varphi,a}_{n^-} = \{h \in \mathcal{H}^{\text{conv}} : \hat{R}^-_\varphi(h) \le a\}$ be the set of classifiers in $\mathcal{H}^{\text{conv}}$ whose empirical convexified type I errors are bounded by a. To make our analysis meaningful, we assume that $\mathcal{H}^{\varphi,\alpha} \neq \emptyset$.

We are now in a position to construct a classifier in $\mathcal{H}^{\text{conv}}$ according to the Neyman-Pearson paradigm. For any $\tau > 0$ such that $\tau \leq \alpha\sqrt{n^-}$, define the convexified NP classifier $\tilde{h}^\tau$ as any classifier that solves the following optimization problem

$$\min_{\substack{h \in \mathcal{H}^{\text{conv}} \\ \hat{R}_\varphi^-(h) \leq \alpha - \tau/\sqrt{n^-}}} \hat{R}_\varphi^+(h)\,. \tag{4}$$

Note that this problem consists of minimizing a convex function subject to a convex constraint and can therefore be solved by standard algorithms (see, e.g., Boyd and Vandenberghe, 2004, and references therein).

In the next section, we present a series of results on type I and type II errors of classifiers that are more general than $\tilde{h}^\tau$.

4. Performance Bounds

In this section, we will first evaluate our proposed classifier $\tilde{h}^\tau$ against φ I/II errors. These benchmarks are necessary because $\tilde{h}^\tau$ is constructed based on them. Moreover, in view of the decision theory framework, such errors are just expected loss with a general loss function φ, which are interesting to investigate. As the true type I and type II errors are usually the main concern in statistical learning, we will also address the effect of convexification in terms of the excess type II error. Interestingly, given that we want to be conservative on type I error, neither working on φ errors nor working on true errors leads to a most desirable type II error. The price to pay for being conservative will be characterized explicitly.

4.1 Control of Type I Error

First, we identify classifiers h such that $R_\varphi^-(h) \leq \alpha$ with high probability. This is done by enforcing its empirical counterpart $\hat{R}_\varphi^-(h)$ be bounded from above by the quantity

$$\alpha_\kappa = \alpha - \kappa/\sqrt{n^-},$$

for a proper choice of positive constant κ.

Theorem 3 *Fix constants $\delta, \alpha \in (0,1), L > 0$ and let $\varphi : [-1,1] \to \mathbb{R}^+$ be a given L-Lipschitz convex surrogate. Define*

$$\kappa = 4\sqrt{2}L\sqrt{\log\left(\frac{2M}{\delta}\right)}\,.$$

Then for any (random) classifier $h \in \mathcal{H}^{\text{conv}}$ that satisfies $\hat{R}_\varphi^-(h) \leq \alpha_\kappa$, we have

$$R^-(h) \leq R_\varphi^-(h) \leq \alpha\,.$$

with probability at least $1-\delta$. Equivalently

$$\mathbb{P}\left[\mathcal{H}_{n^-}^{\varphi,\alpha_\kappa} \subset \mathcal{H}^{\varphi,\alpha}\right] \geq 1-\delta\,. \tag{5}$$

4.2 Simultaneous Control of the Two Errors

Theorem 3 guarantees that any classifier that satisfies the strengthened constraint on the empirical φ-type I error will have φ-type I error and true type I error bounded from above by α. We now check that the constraint is not too strong so that the φ-type II error is overly deteriorated. Indeed, an extremely small α_κ would certainly ensure a good control of type I error but would deteriorate significantly the best achievable φ-type II error. Below, we show not only that this is not the case for our approach but also that the convexified NP classifier $\tilde{h}^\tau$ defined in Section 3.2 with $\tau = \alpha_\kappa$ suffers only a small degradation of its φ-type II error compared to the best achievable. Analogous to classical binary classification, a desirable result is that with high probability,

$$R^+_\varphi(\tilde{h}^\kappa) - \min_{h \in \mathcal{H}^{\varphi,\alpha}} R^+_\varphi(h) \le \tilde{\Delta}_n(\mathcal{F}), \tag{6}$$

where $\tilde{\Delta}_n(\mathcal{F})$ goes to 0 as $n = n^- + n^+ \to \infty$.

The following proposition is pivotal to our argument.

Proposition 4 *Fix constant $\alpha \in (0,1)$ and let $\varphi : [-1,1] \to \mathbb{R}^+$ be a given continuous convex surrogate. Assume further that there exists $\nu_0 > 0$ such that the set of classifiers $\mathcal{H}^{\varphi,\alpha-\nu_0}$ is nonempty. Then, for any $\nu \in (0,\nu_0)$,*

$$\min_{h \in \mathcal{H}^{\varphi,\alpha-\nu}} R^+_\varphi(h) - \min_{h \in \mathcal{H}^{\varphi,\alpha}} R^+_\varphi(h) \le \varphi(1) \frac{\nu}{\nu_0 - \nu}\,.$$

This proposition ensures that if the convex surrogate φ is continuous, strengthening the constraint on type I error (φ-type I error) does not increase too much the best achievable φ-type II error. We should mention that the proof does not use the Lipschitz property of φ, but only that it is uniformly bounded by $\varphi(1)$ on $[-1,1]$. This proposition has direct consequences on chance constrained programming as discussed in Section 5.

The next theorem shows that the NP classifier $\tilde{h}^\kappa$ defined in Section 3.2 is a good candidate to perform classification with the Neyman-Pearson paradigm. It relies on the following assumption which is necessary to verify the condition of Proposition 4.

Assumption 1 *There exists a positive constant $\varepsilon < 1$ such that the set of classifiers $\mathcal{H}^{\varphi,\varepsilon\alpha}$ is nonempty.*

Note that this assumption can be tested using (5) for large enough n^-. Indeed, it follows from this inequality that with probability $1-\delta$,

$$\mathcal{H}^{\varphi,\varepsilon\alpha-\kappa/\sqrt{n^-}}_{n^-} \subset \mathcal{H}^{\varphi,\varepsilon\alpha-\kappa/\sqrt{n^-}+\kappa/\sqrt{n^-}} = \mathcal{H}^{\varphi,\varepsilon\alpha}\,.$$

Thus, it is sufficient to check if $\mathcal{H}^{\varphi,\varepsilon\alpha-\kappa/\sqrt{n^-}}_{n^-}$ is nonempty for some $\varepsilon > 0$. Before stating our main theorem, we need the following definition. Under Assumption 1, let $\bar{\varepsilon}$ denote the smallest ε such that $\mathcal{H}^{\varphi,\varepsilon\alpha} \neq \emptyset$ and let n_0 be the smallest integer such that

$$n_0 \ge \left(\frac{4\kappa}{(1-\bar{\varepsilon})\alpha}\right)^2. \tag{7}$$

Theorem 5 *Let* φ, κ, δ *and* α *be the same as in Theorem 3, and* $\tilde{h}^\kappa$ *denote any solution to* (4). *Moreover, let Assumption 1 hold and assume that* $n^- \geq n_0$ *where* n_0 *is defined in* (7). *Then, the following hold with probability* $1-2\delta$,

$$R^-(\tilde{h}^\kappa) \leq R^-_\varphi(\tilde{h}^\kappa) \leq \alpha \tag{8}$$

and

$$R^+_\varphi(\tilde{h}^\kappa) - \min_{h\in\mathcal{H}^{\varphi,\alpha}} R^+_\varphi(h) \leq \frac{4\varphi(1)\kappa}{(1-\bar{\varepsilon})\alpha\sqrt{n^-}} + \frac{2\kappa}{\sqrt{n^+}}. \tag{9}$$

In particular, there exists a constant $C>0$ *depending on* α, $\varphi(1)$ *and* $\bar{\varepsilon}$, *such that* (9) *yields*

$$R^+_\varphi(\tilde{h}^\kappa) - \min_{h\in\mathcal{H}^{\varphi,\alpha}} R^+_\varphi(h) \leq C\left(\sqrt{\frac{\log(2M/\delta)}{n^-}} + \sqrt{\frac{\log(2M/\delta)}{n^+}}\right).$$

Note here that Theorem 4.2 is not exactly of the type (6). The right hand side of (9) goes to zero if both n^- and n^+ go to infinity. Inequality (9) conveys a message that accuracy of the estimate depends on information from both classes of labeled data. This concern motivates us to consider a different sampling scheme.

4.3 A Different Sampling Scheme

In this subsection (only), we consider a model for observations that is more standard in statistical learning theory (see, e.g., Devroye et al., 1996; Boucheron et al., 2005).

Let $(X_1,Y_1),\ldots,(X_n,Y_n)$ be n independent copies of the random couple $(X,Y)\in\mathcal{X}\times\{-1,1\}$. Denote by P_X the marginal distribution of X and by $\eta(x)=\mathbb{E}[Y|X=x]$ the regression function of Y onto X. Denote by p the probability of positive label and observe that

$$p=\mathbb{P}[Y=1]=\mathbb{E}\left(\mathbb{P}[Y=1|X]\right)=\frac{1+\mathbb{E}[\eta(X)]}{2}.$$

In what follows, we assume that $P_X(\eta(X)=-1)\vee P_X(\eta(X)=1)<1$ so that $p\in(0,1)$.

Let $N^-=\mathrm{card}\{Y_i : Y_i=-1\}$ be the random number of instances labeled -1 and $N^+=n-N^-=\mathrm{card}\{Y_i : Y_i=1\}$. In this setup, the NP classifier is defined as in Section 3.2 where n^- and n^+ are replaced by N^- and N^+ respectively. To distinguish this classifier from $\tilde{h}^\tau$ previously defined, we denote the NP classifier obtained with this sampling scheme by $\tilde{h}^\tau_n$.

Let the event $\mathcal{F}$ be defined by

$$\mathcal{F}=\{R^-_\varphi(\tilde{h}^\kappa_n)\leq\alpha\}\cap\{R^+_\varphi(\tilde{h}^\kappa_n)-\min_{h\in\mathcal{H}^{\varphi,\alpha}} R^+_\varphi(h)\leq\frac{4\varphi(1)\kappa}{(1-\bar{\varepsilon})\alpha\sqrt{N^-}}+\frac{2\kappa}{\sqrt{N^+}}\}.$$

Denote $\mathcal{B}_{n^-}=\{Y_1=\cdots=Y_{n^-}=-1, Y_{n^-+1}=\cdots=Y_n=1\}$. Although the event $\mathcal{B}_{n^-}$ is different from the event $\{N^-=n^-\}$, symmetry leads to the following key observation:

$$\mathbb{P}(\mathcal{F}|N^-=n^-)=\mathbb{P}(\mathcal{F}|\mathcal{B}_{n^-}).$$

Therefore, under the conditions of Theorem 5, we find that for $n^-\geq n_0$ the event $\mathcal{F}$ satisfies

$$\mathbb{P}(\mathcal{F}|N^-=n^-)\geq 1-2\delta. \tag{10}$$

We obtain the following corollary of Theorem 5.

Corollary 6 *Let φ, κ, δ and α be the same as in Theorem 3, and $\tilde{h}_n^\kappa$ be the NP classifier obtained with the current sampling scheme. Then under Assumption 1, if $n > 2n_0/(1-p)$, where n_0 is defined in* (7)*, we have with probability $(1-2\delta)(1-e^{-\frac{n(1-p)^2}{2}})$,*

$$R^-(\tilde{h}_n^\kappa) \le R_\varphi^-(\tilde{h}_n^\kappa) \le \alpha \tag{11}$$

and

$$R_\varphi^+(\tilde{h}_n^\kappa) - \min_{h\in\mathcal{H}^{\varphi,\alpha}} R_\varphi^+(h) \le \frac{4\varphi(1)\kappa}{(1-\bar{\varepsilon})\alpha\sqrt{N^-}} + \frac{2\kappa}{\sqrt{N^+}}\,. \tag{12}$$

Moreover, with probability $1-2\delta-e^{-\frac{n(1-p)^2}{2}}-e^{-\frac{np^2}{2}}$, we have simultaneously (11) *and*

$$R_\varphi^+(\tilde{h}_n^\kappa) - \min_{h\in\mathcal{H}^{\varphi,\alpha}} R_\varphi^+(h) \le \frac{4\sqrt{2}\varphi(1)\kappa}{(1-\bar{\varepsilon})\alpha\sqrt{n(1-p)}} + \frac{2\sqrt{2}\kappa}{\sqrt{np}}\,. \tag{13}$$

4.4 Price to Pay For Being Conservative

We have shown that the the computational feasible classifier $\tilde{h}^\kappa$ satisfies oracle inequalities which take the optimal φ-type II errors as the benchmark. In this subsection, the excess type II error will be measured, and we will characterize the price to pay by being conservative on type I error.

Much like its counterparts in classical binary classification, the next strikingly simple relation addresses the consequence of convexification in the NP paradigm.

Theorem 7 *Let $\tilde{h}$ be any classifier, then*

$$R^+(\tilde{h}) - \min_{R^-(h)\le\alpha} R^+(h) \le R_\varphi^+(\tilde{h}) - \inf_{R^-(h)\le\alpha} R_\varphi^+(h)\,.$$

This theorem applies to any classifier; in particular, it holds for our proposed $\tilde{h}^\kappa$. As the proof of Theorem 7 indicates, $\min_{R^-(h)\le\alpha} R^+(h) = \inf_{R^-(h)\le\alpha} R_\varphi^+(h)$. So the bound in the theorem can be very tight, depending on the nature of $\tilde{h}$.

Now relax the range of base classifiers $h_1, \ldots, h_M$ to be $[-B,B]$. Denote by $\mathcal{H}_B^{\varphi,\alpha}$ the set of convex combinations of the base classifiers that have φ-type I error bounded from above by α.

Therefore, we have the following observation:

$$R^+(\tilde{h}^\kappa) - \min_{R^-(h)\le\alpha} R^+(h) \le T_1 + T_2 + T_3\,,$$

where

$$\begin{aligned}
T_1 &= R_\varphi^+(\tilde{h}^\kappa) - \min_{h\in\mathcal{H}_B^{\varphi,\alpha}} R_\varphi^+(h)\,,\\
T_2 &= \min_{h\in\mathcal{H}_B^{\varphi,\alpha}} R_\varphi^+(h) - \inf_{\substack{R^-(h)\le\alpha\\ -B\le h\le B}} R_\varphi^+(h)\,,\\
T_3 &= \inf_{\substack{R^-(h)\le\alpha\\ -B\le h\le B}} R_\varphi^+(h) - \inf_{R^-(h)\le\alpha} R_\varphi^+(h)\,.
\end{aligned}$$

With the new set of base classifiers taking ranges in $[-B,B]$, Theorem 5 holds if we replace κ by $\kappa_B = 4\sqrt{2}L_B B\sqrt{\log(2M/\delta)}$, where L_B is the Lipschitz constant of φ on $[-B,B]$. Therefore, the

convergence rate of T_1 is explicitly controlled. We can see that with a fixed sample size, choosing a set of base classifiers with smaller range will result in a tighter bound for the excess φ-type II error. However, if one concerns more about the true type II error, choosing a smaller B should not be a better option, because only signs matters for true type I and II errors. This intuition is reflected in the term T_3. When B increases, T_3 decreases. More specifically, it can be shown that

$$T_3 = \big(P^+(\mathcal{X}^+)\varphi(-B) + P^+(\mathcal{X}^-)\varphi(0)\big) - P^+(\mathcal{X}^-) = P^+(\mathcal{X}^+)\varphi(-B)\,,$$

where $\mathcal{X}^+ \subset \mathcal{X}$ is the part of feature space mapped to label $+1$ by the optimal NP classifier that solves $\min_{R^-(h)\le\alpha} R^+(h)$, and $\mathcal{X}^-$ is the part that mapped to label -1; this is what NP Lemma says when there is no need for randomization. Therefore, T_3 diminishes towards 0 as B increases, and the trade-off between T_1 and T_3 is very clear. When $\varphi(x) = (1+x)_+$ is the hinge loss, the best trade-off occurs at $B \in (0,1)$. When $B(\ge 1)$ goes to infinity, $T_3 = 0$ stays the same while the upper bound of T_1 blows up.

Note that $\mathcal{H}_B^{\varphi,\alpha} \subset \{h : R^-(h) \le \alpha, -B \le h \le B\}$, so T_2 reflects the price to pay for being conservative on type I error. It also reflect the bias for choosing a specific candidate pool of classifiers, that is, convex combinations of base classifiers. As long as the base classifiers are rich enough, the latter bias should be small. However in our belief, the price to pay for being conservative is unavoidable. Even if we do not resort to convexification, getting the best insurance on type I error still demands a high premium on type II error.

The same attitude is shared in the seminal paper Cannon et al. (2002), where it was claimed without justification that if we use $\alpha' < \alpha$ for the empirical program, "it seems unlikely that we can control the estimation error $R^+(\hat{h}) - R^+(h^*)$ in a distribution independent way". The following proposition confirms this opinion in a certain sense.

Fix $\alpha \in (0,1), n^- \ge 1, n^+ \ge 1$ and $\alpha' < \alpha$. Let $\hat{h}(\alpha')$ be the classifier defined as any solution of the following optimization problem:

$$\min_{\substack{h\in\mathcal{H} \\ \hat{R}^-(h)\le\alpha'}} \hat{R}^+(h)\,.$$

The following negative result holds not only for this estimator but also for the oracle $h^*(\alpha')$ defined as the solution of

$$\min_{\substack{h\in\mathcal{H} \\ R^-(h)\le\alpha'}} R^+(h)\,.$$

Note that $h^*(\alpha')$ is not a classifier but only a pseudo-classifier since it depends on the unknown distribution of the data.

Proposition 8 *There exist base classifiers h_1, h_2 and a probability distribution for (X,Y) for which, regardless of the sample sizes n^- and n^+, any pseudo-classifier $h_{\tilde{\lambda}} = \tilde{\lambda}h_1 + (1-\tilde{\lambda})h_2$, $0 \le \tilde{\lambda} \le 1$, such that $R^-(h_{\tilde{\lambda}}) < \alpha$, it holds*

$$R^+(h_{\tilde{\lambda}}) - \min_{R^-(h_\lambda)\le\alpha, \lambda\in[0,1]} R^+(h_\lambda) \ge \alpha\,.$$

In particular, the excess type II risk of $h^*(\alpha - \varepsilon_{n^-}), \varepsilon_{n^-} > 0$ does not converge to zero as sample sizes increase even if $\varepsilon_{n^-} \to 0$. Moreover, when $\alpha \le 1/2$ for any (pseudo-)classifier $h_{\tilde{\lambda}}$ ($0 \le \tilde{\lambda} \le 1$) such that $\hat{R}^-(h_{\tilde{\lambda}}) < \alpha$, it holds

$$R^+(h_{\tilde{\lambda}}) - \min_{R^-(h_\lambda)\le\alpha, \lambda\in[0,1]} R^+(h_\lambda) \ge \alpha\,.$$

with probability at least $\alpha \wedge 1/4$. *In other words, if we let* $\mathcal{A} = \{h_\lambda : \hat{R}^-(h_\lambda) < \alpha, \lambda \in [0,1]\}$, *and* $\mathcal{B} = \{h_\lambda : R^+(h_\lambda) - \min_{R^-(h_\lambda) \le \alpha, \lambda \in [0,1]} R^+(h_\lambda) \ge \alpha, \lambda \in [0,1]\}$, *then* $\mathbb{P}(\mathcal{A} \subset \mathcal{B}) \ge \alpha \wedge 1/4$. *In particular, the excess type II risk of* $\hat{h}(\alpha - \varepsilon_{n^-})$, $\varepsilon_{n^-} > 0$ *does not converge to zero with positive probability, as sample sizes increase even if* $\varepsilon_{n^-} \to 0$.

The proof of this result is postponed to Appendix A. The fact that the oracle $h^*(\alpha - \varepsilon_{n^-})$ satisfies the lower bound indicates that the problem comes from using a strengthened constraint. Note that the condition $\alpha \le 1/2$ is purely technical and can be removed. Nevertheless, it is always the case in practice that $\alpha \le 1/2$. When the number of base classifiers is great then two, we believe that similar counterexamples can be still constructed, though the technicality will be move involved.

In view of this negative result and our previous discussion, we have to accept the price to pay for being conservative on type I error, and our classifier $\tilde{h}^\kappa$ is no exception. As such conservativeness follows from the original spirit of the Neyman-Pearson paradigm, we need to pay whatever we have to pay. The positive sides are that our proposed procedure is computationally feasible, and it attains good rates under a different (but still meaningful) criterion.

5. Chance Constrained Optimization

Implementing the Neyman-Pearson paradigm for the convexified binary classification bears strong connections with chance constrained optimization. A recent account of such problems can be found in Ben-Tal et al. (2009, Chapter 2) and we refer to this book for references and applications. A chance constrained optimization problem is of the following form:

$$\min_{\lambda \in \Lambda} f(\lambda) \quad \text{s.t.} \quad \mathbb{P}\{F(\lambda, \xi) \le 0\} \ge 1 - \alpha, \tag{14}$$

where $\xi \in \Xi$ is a random vector, $\Lambda \subset \mathbb{R}^M$ is convex, α is a small positive number and f is a deterministic real valued convex function. Problem (14) can be viewed as a relaxation of robust optimization. Indeed, for the latter, the goal is to solve the problem

$$\min_{\lambda \in \Lambda} f(\lambda) \quad \text{s.t.} \quad \sup_{\xi \in \Xi} F(\lambda, \xi) \le 0\,, \tag{15}$$

and this essentially corresponds to (14) for the case $\alpha = 0$. For simplicity, we take F to be scalar valued but extensions to vector valued functions and conic orders are considered in Ben-Tal et al. (2009, Chapter 10). Moreover, it is standard to assume that $F(\cdot, \xi)$ is convex almost surely.

Problem (14) may not be convex because the chance constraint $\{\lambda \in \Lambda : \mathbb{P}\{F(\lambda, \xi) \le 0\} \ge 1 - \alpha\}$ is not convex in general and thus may not be tractable. To solve this problem, Prékopa (1995) and Lagoa et al. (2005) have derived sufficient conditions on the distribution of ξ for the chance constraint to be convex. On the other hand, Calafiore and Campi (2006) initiated a different treatment of the problem where no assumption on the distribution of ξ is made, in line with the spirit of statistical learning. In that paper, they introduced the so-called *scenario approach* based on a sample $\xi_1, \ldots, \xi_n$ of independent copies of ξ. The scenario approach consists of solving

$$\min_{\lambda \in \Lambda} f(\lambda) \quad \text{s.t.} \quad F(\lambda, \xi_i) \le 0, i = 1, \ldots, n. \tag{16}$$

Calafiore and Campi (2006) showed that under certain conditions, if the sample size n is bigger than some $n(\alpha, \delta)$, then with probability $1 - \delta$, the optimal solution $\hat{\lambda}^{sc}$ of (16) is feasible for (14). The

authors did not address the control of the term $f(\hat{\lambda}^{sc}) - f^*$ where f^* denotes the optimal objective value in (14). However, in view of Proposition 8, it is very unlikely that this term can be controlled well.

In an attempt to overcome this limitation, a new *analytical* approach was introduced by Nemirovski and Shapiro (2006). It amounts to solving the following convex optimization problem

$$\min_{\lambda\in\Lambda, t\in\mathbb{R}^s} f(\lambda) \quad \text{s.t.} \quad G(\lambda,t) \le 0, \tag{17}$$

in which t is some additional instrumental variable and where $G(\cdot,t)$ is convex. The problem (17) provides a conservative convex approximation to (14), in the sense that every x feasible for (17) is also feasible for (14). Nemirovski and Shapiro (2006) considered a particular class of conservative convex approximation where the key step is to replace $\mathbb{P}\{F(\lambda,\xi) \ge 0\}$ by $\mathbb{E}\varphi(F(\lambda,\xi))$ in (14), where φ a nonnegative, nondecreasing, convex function that takes value 1 at 0. Nemirovski and Shapiro (2006) discussed several choices of φ including hinge and exponential losses, with a focus on the latter that they name *Bernstein Approximation*.

The idea of a conservative convex approximation is also what we employ in our paper. Recall that P^- the conditional distribution of X given $Y = -1$. In a parallel form of (14), we cast our target problem as

$$\min_{\lambda\in\Lambda} R^+(h_\lambda) \quad \text{s.t.} \quad P^-\{h_\lambda(X) \le 0\} \ge 1-\alpha, \tag{18}$$

where Λ is the flat simplex of $\mathbb{R}^M$.

Problem (18) differs from (14) in that $R^+(h_\lambda)$ is not a convex function of λ. Replacing $R^+(h_\lambda)$ by $R^+_\varphi(h_\lambda)$ turns (18) into a standard chance constrained optimization problem:

$$\min_{\lambda\in\Lambda} R^+_\varphi(h_\lambda) \quad \text{s.t.} \quad P^-\{h_\lambda(X) \le 0\} \ge 1-\alpha. \tag{19}$$

However, there are two important differences in our setting, so that we cannot use directly Scenario Approach or Bernstein Approximation or other analytical approaches to (14). First, $R^+_\varphi(f_\lambda)$ is an *unknown* function of λ. Second, we assume minimum knowledge about P^-. On the other hand, chance constrained optimization techniques in previous literature assume knowledge about the distribution of the random vector ξ. For example, Nemirovski and Shapiro (2006) require that the moment generating function of the random vector ξ is efficiently computable to study the Bernstein Approximation.

Given a finite sample, it is not feasible to construct a strictly conservative approximation to the constraint in (19). On the other hand, it is possible to ensure that if we learned $\hat{h}$ from the sample, this constraint is satisfied with high probability $1-\delta$, that is, the classifier is approximately feasible for (19). In retrospect, our approach to (19) is an innovative hybrid between the analytical approach based on convex surrogates and the scenario approach.

We do have structural assumptions on the problem. Let $g_j, j \in \{1,\dots,M\}$ be arbitrary functions that take values in $[-1,1]$ and $F(\lambda,\xi) = \sum_{j=1}^N \lambda_j g_j(\xi)$. Consider a convexified version of (14):

$$\min_{\lambda\in\Lambda} f(\lambda) \quad \text{s.t.} \quad \mathbb{E}[\varphi(F(\lambda,\xi))] \le \alpha, \tag{20}$$

where φ is a L-Lipschitz convex surrogate, $L > 0$. Suppose that we observe a sample $(\xi_1,\dots,\xi_n)$ that are independent copies of ξ. We propose to approximately solve the above problem by

$$\min_{\lambda\in\Lambda} f(\lambda) \quad \text{s.t.} \quad \sum_{i=1}^n \varphi(F(\lambda,\xi_i)) \le n\alpha - \kappa\sqrt{n},$$

for some $\kappa >$ to be defined. Denote by $\tilde{\lambda}$ any solution to this problem and by f_φ^* the value of the objective at the optimum in (20). The following theorem summarizes our contribution to chance constrained optimization.

Theorem 9 *Fix constants $\delta, \alpha \in (0,1/2), L > 0$ and let $\varphi : [-1,1] \to \mathbb{R}^+$ be a given L-Lipschitz convex surrogate. Define*

$$\kappa = 4\sqrt{2}L\sqrt{\log\left(\frac{2M}{\delta}\right)}\,.$$

Then, the following hold with probability at least $1-2\delta$

(i) *$\tilde{\lambda}$ is feasible for* (14)*.*

(ii) *If there exists $\varepsilon \in (0,1)$ such that the constraint $\mathbb{E}[\varphi(F(\lambda,\xi))] \le \varepsilon\alpha$ is feasible for some $\lambda \in \Lambda$, then for*

$$n \ge \left(\frac{4\kappa}{(1-\varepsilon)\alpha}\right)^2\,,$$

we have

$$f(\tilde{\lambda}) - f_\varphi^* \le \frac{4\varphi(1)\kappa}{(1-\varepsilon)\alpha\sqrt{n}}\,.$$

In particular, as M and n go to infinity with all other quantities kept fixed, we obtain

$$f(\tilde{\lambda}) - f_\varphi^* = O\left(\sqrt{\frac{\log M}{n}}\right)\,.$$

The proof essentially follows that of Theorem 5 and we omit it. The limitations of Theorem 9 include rigid structural assumptions on the function F and on the set Λ. While the latter can be easily relaxed using more sophisticated empirical process theory, the former is inherent to our analysis.

Acknowledgments

Philippe Rigollet is supported by the National Science Foundation (DMS-0906424 & DMS-1053987).

Appendix A. Proof of the Main Results

We gather in this appendix the proofs of the main results of the paper.

A.1 Proof of Theorem 3

We begin with the following lemma, which is extensively used in the sequel. Its proof relies on standard arguments to bound suprema of empirical processes. Recall that $\{h_1, \dots, h_M\}$ is family of M classifiers such that $h_j : \mathcal{X} \to [-1,1]$ and that for any λ in the simplex $\Lambda \subset R^M$, h_λ denotes the convex combination defined by

$$\mathsf{h}_\lambda = \sum_{j=1}^N \lambda_j h_j\,.$$

The following standard notation in empirical process theory will be used. Let $X_1,\dots,X_n \in \mathcal{X}$ be n i.i.d random variables with marginal distribution P. Then for any measurable function $f : \mathcal{X} \to \mathbb{R}$, we write

$$P_n(f) = \frac{1}{n}\sum_{i=1}^{n} f(X_i) \qquad \text{and} \qquad P(f) = \mathbb{E}f(X) = \int f \mathrm{d}P .$$

Moreover, the Rademacher average of f is defined as

$$R_n(f) = \frac{1}{n}\sum_{i=1}^{n} \varepsilon_i f(X_i) ,$$

where $\varepsilon_1,\dots,\varepsilon_n$ are i.i.d. Rademacher random variables such that $\mathbb{P}(\varepsilon_i = 1) = \mathbb{P}(\varepsilon_i = -1) = 1/2$ for $i = 1,\dots,n$.

Lemma 10 *Fix $L > 0, \delta \in (0,1)$. Let $X_1,\dots,X_n$ be n i.i.d random variables on $\mathcal{X}$ with marginal distribution P. Moreover, let $\varphi : [-1,1] \to \mathbb{R}$ an L-Lipschitz function. Then, with probability at least $1-\delta$, it holds*

$$\sup_{\lambda\in\Lambda} |(P_n - P)(\varphi \circ h_\lambda)| \le \frac{4\sqrt{2}L}{\sqrt{n}} \sqrt{\log\left(\frac{2M}{\delta}\right)} .$$

Proof Define $\bar{\varphi}(\cdot) \doteq \varphi(\cdot) - \varphi(0)$, so that $\bar{\varphi}$ is an L-Lipschitz function that satisfies $\bar{\varphi}(0) = 0$. Moreover, for any $\lambda \in \Lambda$, it holds

$$(P_n - P)(\varphi \circ \mathsf{h}_\lambda) = (P_n - P)(\bar{\varphi} \circ \mathsf{h}_\lambda) .$$

Let $\Phi : \mathbb{R} \to \mathbb{R}_+$ be a given convex increasing function. Applying successively the symmetrization and the contraction inequalities (see, e.g., Koltchinskii, 2011, Chapter 2), we find

$$\mathbb{E}\Phi\left(\sup_{\lambda\in\Lambda} |(P_n - P)(\bar{\varphi} \circ \mathsf{h}_\lambda)|\right) \le \mathbb{E}\Phi\left(2\sup_{\lambda\in\Lambda} |R_n(\bar{\varphi} \circ \mathsf{h}_\lambda)|\right) \le \mathbb{E}\Phi\left(4L\sup_{\lambda\in\Lambda} |R_n(\mathsf{h}_\lambda)|\right) .$$

Observe now that $\lambda \mapsto |R_n(\mathsf{h}_\lambda)|$ is a convex function and Theorem 32.2 in Rockafellar (1997) entails that

$$\sup_{\lambda\in\Lambda} |R_n(\mathsf{h}_\lambda)| = \max_{1\le j\le M} \left|R_n(h_j)\right| .$$

We now use a Chernoff bound to control this quantity. To that end, fix $s,t > 0$, and observe that

$$\begin{aligned}
\mathbb{P}\left(\sup_{\lambda\in\Lambda} |(P_n - P)(\varphi \circ \mathsf{h}_\lambda)| > t\right) &\le \frac{1}{\Phi(st)}\mathbb{E}\Phi\left(s\sup_{\lambda\in\Lambda} |(P_n - P)(\bar{\varphi} \circ \mathsf{h}_\lambda)|\right) \\
&\le \frac{1}{\Phi(st)}\mathbb{E}\Phi\left(4Ls\max_{1\le j\le M} \left|R_n(h_j)\right|\right) . \qquad (21)
\end{aligned}$$

Moreover, since Φ is increasing,

$$\begin{aligned}
\mathbb{E}\Phi\left(4Ls\max_{1\le j\le M} \left|R_n(h_j)\right|\right) &= \mathbb{E}\max_{1\le j\le M} \Phi\left(4Ls\left|R_n(h_j)\right|\right) \\
&\le \sum_{j=1}^{M} \mathbb{E}\left[\Phi\left(4LsR_n(h_j)\right) \vee \Phi\left(-4LsR_n(h_j)\right)\right] \\
&\le 2\sum_{j=1}^{M} \mathbb{E}\Phi\left(4LsR_n(h_j)\right) . \qquad (22)
\end{aligned}$$

Now choose $\Phi(\cdot) = \exp(\cdot)$, then

$$\mathbb{E}\Phi\left(4LsR_n(h_j)\right) = \prod_{i=1}^{n} \mathbb{E}\cosh\left(\frac{4Lsh_j(X_i)}{n}\right) \leq \exp\left(\frac{8L^2s^2}{n}\right),$$

where cosh is the hyperbolic cosine function and where in the inequality, we used the fact that $|h_j(X_i)| \leq 1$ for any i, j and $\cosh(x) \leq \exp(x^2/2)$. Together with (21) and (22), it yields

$$\mathbb{P}\left(\sup_{\lambda\in\Lambda} |(P_n - P)(\varphi \circ \mathsf{h}_\lambda)| > t\right) \leq 2M \inf_{s>0} \exp\left(\frac{8L^2s^2}{n} - st\right) \leq 2M \exp\left(-\frac{nt^2}{32L^2}\right).$$

Choosing

$$t = \frac{4\sqrt{2}L}{\sqrt{n}} \sqrt{\log\left(\frac{2M}{\delta}\right)},$$

completes the proof of the Lemma. ■

We now proceed to the proof of Theorem 3. Note first that from the properties of φ, $R^-(h) \leq R^-_\varphi(h)$. Next, we have for any data-dependent classifier $h \in \mathcal{H}^{\text{conv}}$ such that $\hat{R}^-_\varphi(h) \leq \alpha_\kappa$:

$$R^-_\varphi(h) \leq \hat{R}^-_\varphi(h) + \sup_{h\in\mathcal{H}^{\text{conv}}} \left|\hat{R}^-_\varphi(h) - R^-_\varphi(h)\right| \leq \alpha - \frac{\kappa}{\sqrt{n^-}} + \sup_{h\in\mathcal{H}^{\text{conv}}} \left|\hat{R}^-_\varphi(h) - R^-_\varphi(h)\right|.$$

Lemma 10 implies that, with probability $1-\delta$

$$\sup_{h\in\mathcal{H}^{\text{conv}}} \left|\hat{R}^-_\varphi(h) - R^-_\varphi(h)\right| = \sup_{\lambda\in\Lambda} \left|(P^-_{n^-} - P^-)(\varphi \circ \mathsf{h}_\lambda)\right| \leq \frac{\kappa}{\sqrt{n^-}}.$$

The previous two displays imply that $R^-_\varphi(h) \leq \alpha$ with probability $1-\delta$, which completes the proof of Theorem 3.

A.2 Proof of Proposition 4

The proof of this proposition builds upon the following lemma.

Lemma 11 *Let $\gamma(\alpha) = \inf_{h_\lambda \in \mathcal{H}^{\varphi,\alpha}} R^+_\varphi(h_\lambda)$, then γ is a non-increasing convex function on $[0,1]$.*

Proof First, it is clear that γ is a non-increasing function of α because for $\alpha' > \alpha$, $\{\mathsf{h}_\lambda \in \mathcal{H}^{\text{conv}} : R^-_\varphi(\mathsf{h}_\lambda) \leq \alpha\} \subset \{\mathsf{h}_\lambda \in \mathcal{H}^{\text{conv}} : R^-_\varphi(\mathsf{h}_\lambda) \leq \alpha'\}$.

We now show that γ is convex. To that end, observe first that since φ is continuous on $[-1,1]$, the set $\{\lambda \in \Lambda : \mathsf{h}_\lambda \in \mathcal{H}^{\varphi,\alpha}\}$ is compact. Moreover, the function $\lambda \mapsto R^+_\varphi(\mathsf{h}_\lambda)$ is convex. Therefore, there exists $\lambda^* \in \Lambda$ such that

$$\gamma(\alpha) = \inf_{\mathsf{h}_\lambda \in \mathcal{H}^{\varphi,\alpha}} R^+_\varphi(\mathsf{h}_\lambda) = \min_{\mathsf{h}_\lambda \in \mathcal{H}^{\varphi,\alpha}} R^+_\varphi(\mathsf{h}_\lambda) = R^+_\varphi(\mathsf{h}_{\lambda^*}).$$

Now, fix $\alpha_1, \alpha_2 \in [0,1]$. From the above considerations, there exist $\lambda_1, \lambda_2 \in \Lambda$ such that $\gamma(\alpha_1) = R^+_\varphi(\mathsf{h}_{\lambda_1})$ and $\gamma(\alpha_2) = R^+_\varphi(\mathsf{h}_{\lambda_2})$. For any $\theta \in (0,1)$, define the convex combinations $\bar{\alpha}_\theta = \theta\alpha_1 + (1-\theta)\alpha_2$ and $\bar{\lambda}_\theta = \theta\lambda_1 + (1-\theta)\lambda_2$. Since $\lambda \mapsto R^-_\varphi(\mathsf{h}_\lambda)$ is convex, it holds

$$R^-_\varphi(\mathsf{h}_{\bar{\lambda}_\theta}) \leq \theta R^-_\varphi(\mathsf{h}_{\lambda_1}) + (1-\theta) R^-_\varphi(\mathsf{h}_{\lambda_2}) \leq \theta\alpha_1 + (1-\theta)\alpha_2 = \bar{\alpha}_\theta,$$

so that $h^-_{\lambda_\theta} \in \mathcal{H}^{\varphi,\bar{\alpha}_\theta}$. Hence, $\gamma(\bar{\alpha}_\theta) \le R^+_\varphi(h^-_{\lambda_\theta})$. Together with the convexity of φ, it yields

$$\gamma(\theta\alpha_1 + (1-\theta)\alpha_2) \le R^+_\varphi(h^-_{\lambda_\theta}) \le \theta R^+_\varphi(h_{\lambda_1}) + (1-\theta)R^+_\varphi(h_{\lambda_2}) = \theta\gamma(\alpha_1) + (1-\theta)\gamma(\alpha_2)\,.$$

■

We now complete the proof of Proposition 4. For any $x \in [0,1]$, let $\gamma(x) = \inf_{h\in\mathcal{H}^{\varphi,x}} R^+_\varphi(h)$ and observe that the statement of the proposition is equivalent to

$$\gamma(\alpha-\nu) - \gamma(\alpha) \le \varphi(1)\frac{\nu}{\nu_0-\nu}\,, \quad 0 < \nu < \nu_0\,.$$

Lemma 11 together with the assumption that $\mathcal{H}^{\varphi,\alpha-\nu_0} \neq \emptyset$ imply that γ is a non-increasing convex real-valued function on $[\alpha-\nu_0, 1]$ so that

$$\gamma(\alpha-\nu) - \gamma(\alpha) \le \nu \sup_{g\in\partial\gamma(\alpha-\nu)} |g|\,,$$

where $\partial\gamma(\alpha-\nu)$ denotes the sub-differential of γ at $\alpha-\nu$. Moreover, since γ is a non-increasing convex function on $[\alpha-\nu_0, \alpha-\nu]$, it holds

$$\gamma(\alpha-\nu_0) - \gamma(\alpha-\nu) \ge (\nu-\nu_0) \sup_{g\in\partial\gamma(\alpha-\nu)} |g|\,.$$

The previous two displays yield

$$\gamma(\alpha-\nu) - \gamma(\alpha) \le \nu\frac{\gamma(\alpha-\nu_0)-\gamma(\alpha-\nu)}{\nu-\nu_0} \le \nu\frac{\varphi(1)}{\nu-\nu_0}\,.$$

A.3 Proof of Theorem 5

Define the events $\mathcal{E}^-$ and $\mathcal{E}^+$ by

$$\mathcal{E}^- = \bigcap_{h\in\mathcal{H}^{\mathrm{conv}}} \{|\hat{R}^-_\varphi(h) - R^-_\varphi(h)| \le \frac{\kappa}{\sqrt{n^-}}\}\,,$$
$$\mathcal{E}^+ = \bigcap_{h\in\mathcal{H}^{\mathrm{conv}}} \{|\hat{R}^+_\varphi(h) - R^+_\varphi(h)| \le \frac{\kappa}{\sqrt{n^+}}\}\,.$$

Lemma 10 implies

$$\mathbb{P}(\mathcal{E}^-) \wedge \mathbb{P}(\mathcal{E}^+) \ge 1-\delta\,. \tag{23}$$

Note first that Theorem 3 implies that (8) holds with probability $1-\delta$. Observe now that the l.h.s of (9) can be decomposed as

$$R^+_\varphi(\tilde{h}^\kappa) - \min_{h\in\mathcal{H}^{\varphi,\alpha}} R^+_\varphi(h) = A_1 + A_1 + A_3\,,$$

where

$$A_1 = \left(R^+_\varphi(\tilde{h}^\kappa) - \hat{R}^+_\varphi(\tilde{h}^\kappa)\right) + \left(\hat{R}^+_\varphi(\tilde{h}^\kappa) - \min_{h\in\mathcal{H}^{\varphi,\alpha_\kappa}_{n^-}} R^+_\varphi(h)\right)$$
$$A_2 = \min_{h\in\mathcal{H}^{\varphi,\alpha_\kappa}_{n^-}} R^+_\varphi(h) - \min_{h\in\mathcal{H}^{\varphi,\alpha_{2\kappa}}} R^+_\varphi(h)$$
$$A_3 = \min_{h\in\mathcal{H}^{\varphi,\alpha_{2\kappa}}} R^+_\varphi(h) - \min_{h\in\mathcal{H}^{\varphi,\alpha}} R^+_\varphi(h).$$

To bound A_1 from above, observe that

$$A_1 \le 2 \sup_{h \in \mathcal{H}^{\varphi,\alpha_\kappa}_{n^-}} |\hat{R}^+_\varphi(h) - R^+_\varphi(h)| \le 2 \sup_{h \in \mathcal{H}^{\mathrm{conv}}} |\hat{R}^+_\varphi(h) - R^+_\varphi(h)|.$$

Therefore, on the event $\mathcal{E}^+$ it holds

$$A_1 \le \frac{2\kappa}{\sqrt{n^+}}.$$

We now treat A_2. Note that $A_2 \le 0$ on the event $\mathcal{H}^{\varphi,\alpha_{2\kappa}} \subset \mathcal{H}^{\varphi,\alpha_\kappa}_{n^-}$. But this event contains $\mathcal{E}^-$ so that $A_2 \le 0$ on the event $\mathcal{E}^-$.

Finally, to control A_3, observe that under Assumption 1, Proposition 4 can be applied with $\nu = 2\kappa/\sqrt{n^-}$ and $\nu_0 = (1-\bar{\varepsilon})\alpha$. Indeed, the assumptions of the theorem imply that $\nu \le \nu_0/2$. It yields

$$A_3 \le \frac{4\varphi(1)\kappa}{(1-\bar{\varepsilon})\alpha\sqrt{n^-}}.$$

Combining the bounds on A_1, A_2 and A_3 obtained above, we find that (9) holds on the event $\mathcal{E}^- \cap \mathcal{E}^+$ that has probability at least $1-2\delta$ in view of (23).

The last statement of the theorem follows directly from the definition of κ.

A.4 Proof of Corollary 6

Now prove (12),

$$\begin{aligned}
\mathbb{P}(\mathcal{F}) &= \sum_{n^-=0}^{n} \mathbb{P}(\mathcal{F}|N^- = n^-)\mathbb{P}(N^- = n^-) \\
&\ge \sum_{n^-=n_0}^{n} \mathbb{P}(\mathcal{F}|N^- = n^-)\mathbb{P}(N^- = n^-) \\
&\ge (1-2\delta)\mathbb{P}(N^- \ge n_0),
\end{aligned}$$

where in the last inequality, we used (10). Applying now Lemma 12, we obtain

$$\mathbb{P}(N^- \ge n_0) \ge 1 - e^{-\frac{n(1-p)^2}{2}}.$$

Therefore,

$$\mathbb{P}(\mathcal{F}) \ge (1-2\delta)(1-e^{-\frac{n(1-p)^2}{2}}),$$

which completes the proof of (12).

The proof of (13) follows by observing that

$$\left\{ R^+_\varphi(\tilde{h}^\kappa_n) - \min_{h \in \mathcal{H}^{\varphi,\alpha}} R^+_\varphi(h) > \frac{4\sqrt{2}\varphi(1)\kappa}{(1-\bar{\varepsilon})\alpha\sqrt{n(1-p)}} + \frac{2\sqrt{2}\kappa}{\sqrt{np}} \right\} \subset (\mathcal{A}_1 \cap \mathcal{A}_2^c) \cup \mathcal{A}_2 \cup \mathcal{A}_3,$$

where

$$\begin{aligned}
\mathcal{A}_1 &= \left\{ R^+_\varphi(\tilde{h}^\kappa_n) - \min_{h \in \mathcal{H}^{\varphi,\alpha}} R^+_\varphi(h) > \frac{4\varphi(1)\kappa}{(1-\bar{\varepsilon})\alpha\sqrt{N^-}} + \frac{2\kappa}{\sqrt{N^+}} \right\} \subset \mathcal{F}^c, \\
\mathcal{A}_2 &= \{N^- < n(1-p)/2\}, \\
\mathcal{A}_3 &= \{N^+ < np/2\}.
\end{aligned}$$

Since $\mathcal{A}_2^c \subset \{N^- \geq n_0\}$, we find

$$\mathbb{P}(\mathcal{A}_1 \cap \mathcal{A}_2^c) \leq \sum_{n^- \geq n_0} \mathbb{P}(\mathcal{F}^c | N^- = n^-)\mathbb{P}(N^- = n^-) \leq 2\delta.$$

Next, using Lemma 12, we get

$$\mathbb{P}(\mathcal{A}_2) \leq e^{-\frac{n(1-p)^2}{2}} \qquad \text{and} \qquad \mathbb{P}(\mathcal{A}_3) \leq e^{-\frac{np^2}{2}}.$$

Hence, we find

$$\mathbb{P}\left\{ R_\varphi^+(\tilde{h}_n^\kappa) - \min_{h \in \mathcal{H}^{\varphi,\alpha}} R_\varphi^+(h) > \frac{4\sqrt{2}\varphi(1)\kappa}{(1-\bar{\varepsilon})\alpha\sqrt{n(1-p)}} + \frac{2\sqrt{2}\kappa}{\sqrt{np}} \right\} \leq 2\delta + e^{-\frac{n(1-p)^2}{2}} + e^{-\frac{np^2}{2}},$$

which completes the proof of the corollary.

A.5 Proof of Theorem 7

First observe that for any $\hat{h}$, $R^+(\hat{h}) \leq R_\varphi^+(\hat{h})$. Then the result follows from the claim that

$$\min_{R^-(h) \leq \alpha} R^+(h) = \inf_{R^-(h) \leq \alpha} R_\varphi^+(h).$$

It is clear $\min_{R^-(h)\leq\alpha} R^+(h) \leq \inf_{R^-(h)\leq\alpha} R_\varphi^+(h)$, it remains to prove the other direction. By the Neyman-Pearson Lemma, We can decompose the feature space $\mathcal{X}$ into a disjoint union of $\mathcal{X}^+$ and $\mathcal{X}^-$, and the optimal (pseudo) classifier that solves $\min_{R^-(h)\leq\alpha} R^+(h)$ assigns label $+1$ for any $x \in \mathcal{X}^+$, and -1 for any $x \in \mathcal{X}^-$. Note that if any two classifiers g_1 and g_2 have the same signs, that is, $\operatorname{sgn}(g_1) = \operatorname{sgn}(g_2)$, then $R^-(g_1) = R^-(g_2)$ and $R^+(g_1) = R^+(g_2)$. On the other hand, for φ-type I and II errors, values of classifiers do matter.

Let $\bar{h}_{B,\varepsilon}(x) = B \cdot \mathbb{I}(x \in \mathcal{X}^+) + (-\varepsilon) \cdot \mathbb{I}(x \in \mathcal{X}^-)$. Then clearly for any $B, \varepsilon > 0$, $\bar{h}_{B,\varepsilon}$ solves $\min_{R^-(h)\leq\alpha} R^+(h)$. Also, for any $B, \varepsilon > 0$,

$$\inf_{R^-(h)\leq\alpha} R_\varphi^+(h) \leq R_\varphi^+(\bar{h}_{B,\varepsilon}) = P^+(\mathcal{X}^+)\varphi(-B) + P^+(\mathcal{X}^-)\varphi(\varepsilon).$$

Taking the limit, we have

$$\lim_{B\to\infty,\varepsilon\to 0} R_\varphi^+(\bar{h}_{B,\varepsilon}) = \lim_{B\to\infty,\varepsilon\to 0} P^+(\mathcal{X}^+)\varphi(-B) + P^+(\mathcal{X}^-)\varphi(\varepsilon) = P^+(\mathcal{X}^-) = R^+(\bar{h}_{B,\varepsilon}).$$

Therefore, $\inf_{R^-(h)\leq\alpha} R_\varphi^+(h) \leq \min_{R^-(h)\leq\alpha} R^+(h)$, which completes the proof.

A.6 Proof of Proposition 8

Let the base classifiers be defined as

$$h_1(x) = -1 \quad \text{and} \quad h_2(x) = \mathbb{1}(x \leq \alpha) - \mathbb{1}(x > \alpha), \quad \forall x \in [0,1]$$

For any $\lambda \in [0,1]$, denote the convex combination of h_1 and h_2 by $\mathsf{h}_\lambda = \lambda h_1 + (1-\lambda)h_2$, that is,

$$\mathsf{h}_\lambda(x) = (1-2\lambda)\mathbb{1}(x \leq \alpha) - \mathbb{1}(x > \alpha).$$

Suppose the conditional distributions of X given $Y=1$ or $Y=-1$, denoted respectively by P^+ and P^-, are both uniform on $[0,1]$. Recall that $R^-(\mathsf{h}_\lambda) = P^-(\mathsf{h}_\lambda(X) \geq 0)$ and $R^+(\mathsf{h}_\lambda) = P^+(\mathsf{h}_\lambda(X) < 0)$. Then, we have

$$R^-(\mathsf{h}_\lambda) = P^-(\mathsf{h}_\lambda(X) \geq 0) = \alpha \mathbb{1}(\lambda \leq 1/2). \tag{24}$$

Therefore, for any $\tau \in [0,\alpha]$, we have

$$\{\lambda \in [0,1] : R^-(\mathsf{h}_\lambda) \leq \tau\} = \begin{cases} [0,1] & \text{if } \tau = \alpha, \\ (1/2,1] & \text{if } \tau < \alpha. \end{cases}$$

Observe now that

$$R^+(\mathsf{h}_\lambda) = P^+(\mathsf{h}_\lambda(X) < 0) = (1-\alpha)\mathbb{1}(\lambda < 1/2) + \mathbb{1}(\lambda \geq 1/2). \tag{25}$$

For any $\tau \in [0,\alpha]$, it yields

$$\inf_{\lambda \in [0,1]: R^-(\mathsf{h}_\lambda) \leq \tau} R^+(\mathsf{h}_\lambda) = \begin{cases} 1-\alpha & \text{if } \tau = \alpha, \\ 1 & \text{if } \tau < \alpha. \end{cases}$$

Consider now a classifier $\bar{\mathsf{h}}_\lambda$ such that $R^-(\bar{\mathsf{h}}_\lambda) \leq \tau$ for some $\tau < \alpha$. Then from (24), we see that must have $\lambda > 1/2$. Together with (25), this imples that $R^+(\bar{\mathsf{h}}_\lambda) = 1$. It yields

$$R^+(\bar{\mathsf{h}}_\lambda) - \min_{\lambda : R^-(\mathsf{h}_\lambda) \leq \alpha} R^+(\mathsf{h}_\lambda) = 1 - (1-\alpha) = \alpha.$$

This completes the first part of the proposition. Moreover, in the same manner as (24), it can be easily proved that

$$\hat{R}^-(\mathsf{h}_\lambda) = \frac{1}{n^-} \sum_{i=1}^{n^-} \mathbb{1}(\mathsf{h}_\lambda(X_i^-) \geq 0) = \alpha_{n^-} \mathbb{1}(\lambda \leq 1/2), \tag{26}$$

where

$$\alpha_{n^-} = \frac{1}{n^-} \sum_{i=1}^{n^-} \mathbb{1}(X_i^- \leq \alpha) \tag{27}$$

If a classifier $\hat{\mathsf{h}}_\lambda$ is such that $\hat{R}^-(\hat{\mathsf{h}}_\lambda) < \alpha_{n^-}$, then (26) implies that $\lambda > 1/2$. Using again (25), we find also that $R^+(\hat{\mathsf{h}}_\lambda) = 1$. It yields

$$R^+(\hat{\mathsf{h}}_\lambda) - \min_{\lambda : R^-(\mathsf{h}_\lambda) \leq \alpha} R^+(\mathsf{h}_\lambda) = 1 - (1-\alpha) = \alpha.$$

It remains to show that $\hat{R}^-(\hat{\mathsf{h}}_\lambda) < \alpha_{n^-}$ with positive probability for any classifier such that $\hat{R}^-(\hat{\mathsf{h}}_\lambda) \leq \tau$ for some $\tau < \alpha$. Note that a sufficient condition for a classifier $\hat{\mathsf{h}}_\lambda$ to satisfy this constraint is to have $\alpha \leq \alpha_{n^-}$. It is therefore sufficient to find a lower bound on the probability of the event $\mathcal{A} = \{\alpha_{n^-} \geq \alpha\}$. Such a lower bound is provided by Lemma 13, which guarantees that $\mathbb{P}(\mathcal{A}) \geq \alpha \wedge 1/4$.

Appendix B. Technical Lemmas on Binomial Distributions

The following lemmas are purely technical on the tails of Binomial distributions.

Lemma 12 *Let N be a binomial random variables with parameters $n \geq 1$ and $q \in (0,1)$. Then, for any $t > 0$ such that $t \leq nq/2$, it holds*

$$\mathbb{P}(N \geq t) \geq 1 - e^{-\frac{nq^2}{2}} .$$

Proof Note first that $n - N$ has binomial distribution with parameters $n \geq 1$ and $1-q$. Therefore, we can write $n - N = \sum_{i=1}^{n} Z_i$ where Z_i are i.i.d. Bernoulli random variables with parameter $1-q$. Thus, using Hoeffding's inequality, we find that for any $s \geq 0$,

$$\mathbb{P}(n - N - n(1-q) \geq s) \leq e^{-\frac{2s^2}{n}} .$$

Applying the above inequality with $s = n - n(1-q) - t \geq nq/2 \geq 0$ yields

$$\mathbb{P}(N \geq t) = \mathbb{P}(n - N - n(1-q) \leq n - n(1-q) - t) \geq 1 - e^{-\frac{nq^2}{2}} .$$

■

The next lemma provides a lower bound on the probability that a binomial distribution exceeds its expectation. Our result is uniform in the size of the binomial and it can be easily verified that it is sharp by considering sizes $n = 1$ and $n = 2$ and by looking at Figure 1. In particular, we do not resort to Gaussian approximation which improves upon the lower bounds that can be derived from the inequalities presented in Slud (1977).

Lemma 13 *Let N be a binomial random variable with parameters $n \geq 1$ and $0 < q \leq 1/2$. Then, it holds*

$$\mathbb{P}(N \geq nq) \geq q \wedge (1/4) .$$

Proof We introduce the following local definition, which is limited to the scope of this proof. Fix $n \geq 1$ and for any $q \in (0,1)$, let P_q denote the distribution of a binomial random variable with parameters n and q. Note first that if $n = 1$, the result is trivial since

$$P_q(N \geq q) = \mathbb{P}(Z \geq q) = \mathbb{P}(Z = 1) = q ,$$

where Z is a Bernoulli random variable with parameter q.

Assume that $n \geq 2$. Note that if $q \leq 1/n$, then $P_q(N \geq nq) \geq \mathbb{P}(Z = 1) = q$, where Z is a Bernoulli random variable with parameter q. Moreover, for any any integer k such that $k/n < q \leq (k+1)/n$, we have

$$P_q(N \geq nq) = P_q(N \geq k+1) \geq P_{\frac{k}{n}}(N \geq k+1) . \tag{28}$$

The above inequality can be easily proved by taking the derivative over the interval $(k/n, (k+1)/n]$, of the function

$$q \mapsto \sum_{j=k+1}^{n} \binom{n}{j} q^j (1-q)^{n-j} .$$

We now show that

$$P_{\frac{k}{n}}(N \geq k+1) \geq P_{\frac{k-1}{n}}(N \geq k) , \quad 2 \leq k \leq n/2 . \tag{29}$$

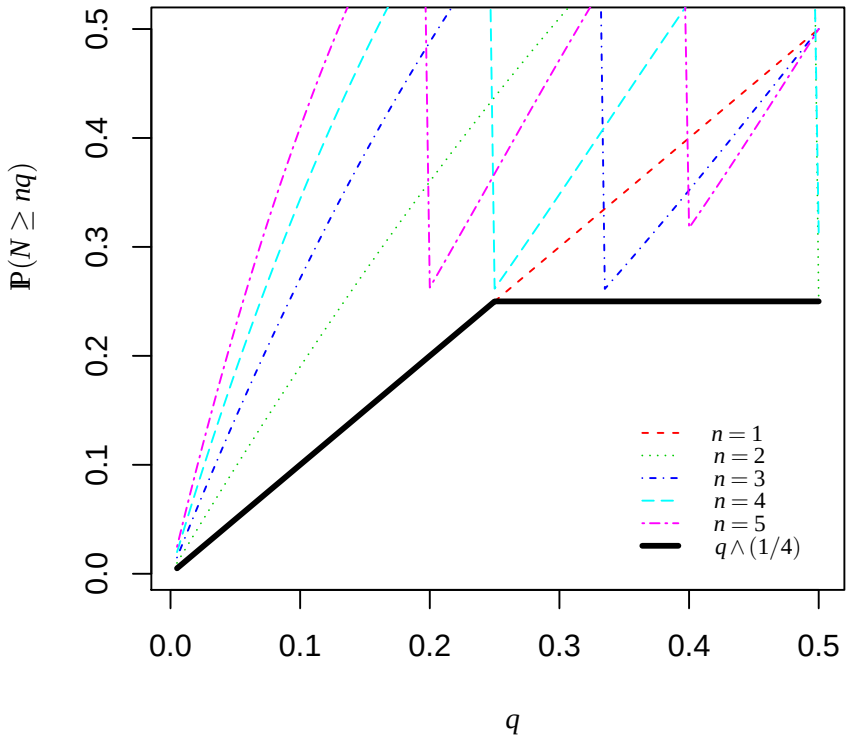

Figure 1: Tail probabilities $\mathbb{P}(N \geq nq)$ where N is a binomial random variable with parameters n and q.

Let $U_1, \ldots, U_n$ be n i.i.d. random variables uniformly distributed on the interval $[0,1]$ and denote by $U_{(k)}$ the corresponding kth order statistic such that $U_{(1)} \leq \ldots \leq U_{(n)}$. Following Feller (1971, Section 7.2), it is not hard to show that

$$P_{\frac{k}{n}}(N \geq k+1) = \mathbb{P}(U_{(k+1)} \leq \frac{k}{n}) = n\binom{n-1}{k}\int_0^{\frac{k}{n}} t^k(1-t)^{n-k-1}\mathrm{d}t\,,$$

and in the same manner,

$$P_{\frac{k-1}{n}}(N \geq k) = \mathbb{P}(U_{(k)} \leq \frac{k-1}{n}) = n\binom{n-1}{k-1}\int_0^{\frac{k-1}{n}} t^{k-1}(1-t)^{n-k}\mathrm{d}t\,.$$

Note that

$$\binom{n-1}{k-1} = \binom{n-1}{k}\frac{k}{n-k}\,,$$

so that (29) follows if we prove

$$k\int_0^{\frac{k-1}{n}} t^{k-1}(1-t)^{n-k}\mathrm{d}t \leq (n-k)\int_0^{\frac{k}{n}} t^k(1-t)^{n-k-1}\mathrm{d}t\,. \tag{30}$$

We can establish the following chain of equivalent inequalities.

$$\begin{aligned}
& k\int_0^{\frac{k-1}{n}} t^{k-1}(1-t)^{n-k}\mathrm{d}t \leq (n-k)\int_0^{\frac{k}{n}} t^k(1-t)^{n-k-1}\mathrm{d}t \\
\Leftrightarrow \quad & \int_0^{\frac{k}{n}} \frac{\mathrm{d}t^k}{\mathrm{d}t}(1-t)^{n-k}\mathrm{d}t \leq -\int_0^{\frac{k}{n}} t^k\frac{\mathrm{d}(1-t)^{n-k}}{\mathrm{d}t}\mathrm{d}t + k\int_{\frac{k-1}{n}}^{\frac{k}{n}} t^{k-1}(1-t)^{n-k}\mathrm{d}t \\
\Leftrightarrow \quad & \int_0^{\frac{k}{n}} \frac{\mathrm{d}}{\mathrm{d}t}\left[t^k(1-t)^{n-k}\right]\mathrm{d}t \leq k\int_{\frac{k-1}{n}}^{\frac{k}{n}} t^{k-1}(1-t)^{n-k}\mathrm{d}t \\
\Leftrightarrow \quad & \left(\frac{k}{n}\right)^k\left(1-\frac{k}{n}\right)^{n-k} \leq k\int_{\frac{k-1}{n}}^{\frac{k}{n}} t^{k-1}(1-t)^{n-k}\mathrm{d}t
\end{aligned}$$

We now study the variations of the function $t \mapsto b(t) = t^{k-1}(1-t)^{n-k}$ on the interval $[(k-1)/n, k/n]$. Taking derivative, it is not hard to see that function b admits a unique local optimum, which is a maximum, at $t_0 = \frac{k-1}{n-1}$ and that $t_0 \in ((k-1)/n, k/n)$ because $k \leq n$. Therefore, the function is increasing on $[(k-1)/n, t_0]$ and decreasing on $[t_0, k/n]$. It implies that

$$\int_{\frac{k-1}{n}}^{\frac{k}{n}} b(t)\mathrm{d}t \geq \frac{1}{n}\min\left[b(\frac{k-1}{n}), b(\frac{k}{n})\right].$$

Hence, the proof of (30) follows from the following two observations:

$$\left(\frac{k}{n}\right)^k\left(1-\frac{k}{n}\right)^{n-k} = \frac{k}{n}\left(\frac{k}{n}\right)^{k-1}\left(1-\frac{k}{n}\right)^{n-k} = \frac{k}{n}b(\frac{k}{n}),$$

and

$$\left(\frac{k}{n}\right)^k\left(1-\frac{k}{n}\right)^{n-k} \leq \frac{k}{n}\left(\frac{k-1}{n}\right)^{k-1}\left(1-\frac{k-1}{n}\right)^{n-k} = \frac{k}{n}b(\frac{k-1}{n}).$$

While the first equality above is obvious, the second inequality can be obtained by an equivalent statement is

$$\begin{aligned}
& \left(\frac{k}{n}\right)^{k-1}\left(\frac{n-k}{n}\right)^{n-k} \leq \left(\frac{k-1}{n}\right)^{k-1}\left(\frac{n-k+1}{n}\right)^{n-k} \\
\Leftrightarrow \quad & \left(\frac{k}{k-1}\right)^{k-1}\left(\frac{n-k}{n-k+1}\right)^{n-k} \leq 1
\end{aligned}$$

Since the function $t \mapsto \left(\frac{t+1}{t}\right)^t$ is increasing on $[0,\infty)$, and $k \leq n-k+1$, the result follows.

To conclude the proof of the Lemma, note that (28) and (29) imply that for any $q > 1/n$,

$$P_q(N \geq nq) \geq P_{\frac{1}{n}}(N \geq 2) = 1 - \left(\frac{n-1}{n}\right)^n - \left(\frac{n-1}{n}\right)^{n-1} \geq 1 - \left(\frac{1}{2}\right)^2 - \frac{1}{2} = \frac{1}{4},$$

where, in the last inequality, we used the fact that the function

$$t \mapsto 1 - \left(\frac{t-1}{t}\right)^t - \left(\frac{t-1}{t}\right)^{t-1}$$

is increasing on $[1, \infty)$. ■

References

P. L. Bartlett, M. I. Jordan, and J. D. McAuliffe. Convexity, classification, and risk bounds. *J. Amer. Statist. Assoc.*, 101(473):138–156, 2006.

A. Ben-Tal, L. El Ghaoui, and A. Nemirovski. *Robust Optimization*. Princeton Series in Applied Mathematics. Princeton University Press, Princeton, NJ, 2009.

G. Blanchard, G. Lee, and C. Scott. Semi-supervised novelty detection. *J. Mach. Learn. Res.*, 11: 2973–3009, 2010.

S. Boucheron, O. Bousquet, and G. Lugosi. Theory of classification: A survey of some recent advances. *ESAIM Probab. Stat.*, 9:323–375, 2005.

S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, Cambridge, 2004.

G. C. Calafiore and M. C. Campi. The scenario approach to robust control design. *IEEE Trans. Automat. Control*, 51(5):742–753, 2006.

A. Cannon, J. Howse, D. Hush, and C. Scovel. Learning with the Neyman-Pearson and min-max criteria. Technical Report LA-UR-02-2951, 2002.

A. Cannon, J. Howse, D. Hush, and C. Scovel. Simple classifiers. Technical Report LA-UR-03-0193, 2003.

D. Casasent and X. Chen. Radial basis function neural networks for nonlinear fisher discrimination and neyman-pearson classification. *Neural Networks*, 16(5-6):529 – 535, 2003.

L. Devroye, L. Györfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition, Volume 31 of Applications of Mathematics (New York)*. Springer-Verlag, New York, 1996.

W. Feller. *An Introduction to Probability Theory and Its Applications. Vol. II.* Second edition. John Wiley & Sons Inc., New York, 1971.

M. Han, D. Chen, and Z. Sun. Analysis to Neyman-Pearson classification with convex loss function. *Anal. Theory Appl.*, 24(1):18–28, 2008.

V. Koltchinskii. *Oracle Inequalities in Empirical Risk Minimization and Sparse Recovery Problems*. École d'Été de Probabilités de Saint-Flour XXXVIII-2008. Lecture Notes in Mathematics 2033. Berlin: Springer. ix, 254 p. EUR 48.10 , 2011.

C. M. Lagoa, X. Li, and M. Sznaier. Probabilistically constrained linear programs and risk-adjusted controller design. *SIAM J. Optim.*, 15(3):938–951 (electronic), 2005.

E. L. Lehmann and J. P. Romano. *Testing Statistical Hypotheses*. Springer Texts in Statistics. Springer, New York, third edition, 2005.

A. Nemirovski and A. Shapiro. Convex approximations of chance constrained programs. *SIAM J. Optim.*, 17(4):969–996, 2006.

A. Prékopa. *Stochastic Programming, volume 324 of Mathematics and its Applications*. Kluwer Academic Publishers Group, Dordrecht, 1995.

R. T. Rockafellar. *Convex Analysis*. Princeton Landmarks in Mathematics. Princeton University Press, Princeton, NJ, 1997. Reprint of the 1970 original, Princeton Paperbacks.

R. Schapire. The strength of weak learnability. *Machine learning*, 5(2):197–227, 1990.

C. Scott. Comparison and design of Neyman-Pearson classifiers. Unpublished, 2005.

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

C. Scott and R. Nowak. A Neyman-Pearson approach to statistical learning. *IEEE Transactions on Information Theory*, 51(11):3806–3819, 2005.

E. V. Slud. Distribution inequalities for the binomial law. *Ann. Probability*, 5(3):404–412, 1977.