Neyman-Pearson classification under a strict constraint

Xin Tong
Operations Research
& Financial Engineering
Princeton University
Princeton, NJ 08540 US
xtong@princeton.edu

Philippe Rigollet
Operations Research
& Financial Engineering
Princeton University
Princeton, NJ 08540 US
rigollet@princeton.edu

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. 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 probability of type I error is below a pre-specified level and (ii), it has probability of type II error close to the minimum possible. The proposed classifier is obtained by minimizing an empirical objective subject to an empirical constraint. The novelty of the method is that the classifier output by this problem is shown to satisfy the original constraint on type I error. This strict enforcement of the constraint has interesting consequences on the control of the type II error and we develop new techniques to handle this situation. Finally, connections with chance constrained optimization are evident and are investigated.

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, where the weighting is proportional to the class priors , 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. Motivations for the NP approach come from many practical problems, where the importance of type I error differs from that of type II error. Typical examples include medical diagnosis or anomaly detection.

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. Henceforth, there are two goals in this project. The first is to design a learning procedure so that type I error of the learned classifier $\hat{f}$ is upper bounded by a pre-specified level with pre-specified high probability; the second is to show that $\hat{f}$ has good performance bounds for excess type II error.

This 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 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. Finally different extensions of the main results to a different sampling scheme and to chance constrained optimization are presented in Section 5. The proofs of the main results are gathered in Section 6.

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

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 computation, 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

$$R(h) = \mathbb{E}[\varphi(-h(X)Y)], \tag{2.1}$$

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

Definition 2.1 *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. 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 subsection 5.2.

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 are obviously equivalent.

In this paper, we pursue the first idea by defining the class of candidate classifiers as follows. Let $h_1, \ldots, h_M, M \geq 2$ 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 $h_1, \ldots, h_M$.

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

$$R_\varphi(h_n) \leq R_\varphi(h_{\mathcal{H}^{\text{conv}}}) + \Delta_n(\mathcal{H}^{\text{conv}}, \delta) \tag{2.2}$$

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 4.1 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 type II error relies on Proposition 4.1 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

In classical binary classification, the risk function can be expressed as a convex combination of type I error $R^-(h) = \mathbb{P}(-Yh(X) \geq 0 | Y = -1)$ and of type II error $R^+(h) = \mathbb{P}(-Yh(X) \geq 0 | Y = 1)$:

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

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

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

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

$$\min_{\substack{h \in \mathcal{H} \\ R^-_\varphi(h) \leq \alpha}} R^+_\varphi(h), \tag{2.3}$$

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 $Y \in \{-1, 1\}$ takes value 1 with probability p and value -1 with probability $1 - p$. 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

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

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 2.1 (Neyman-Pearson Lemma) *Let P^- and P^+ be probability distributions possessing densities p^- and p^+ respectively with respect to some measure μ. Let $\varphi_k(x) = \mathbb{1}(L(x) \geq k)$, 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,*

- *φ_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 propositions. Nevertheless, the result motivates the NP paradigm pursued here.

3 Neyman-Pearson classification via convex optimization

Recall that in NP classification with a convex surrogate φ, the goal is to solve the following optimization problem

$$\min_{\substack{h \in \mathcal{H} \\ R^-_\varphi(h) \leq \alpha}} R^+_\varphi(h). \tag{3.1}$$

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 subsection 5.1.

3.1 Previous results and new input

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

$$\min_{\substack{h \in \mathcal{H} \\ \hat{R}^-(h) \leq \alpha + \varepsilon_0/2}} \hat{R}^+(h), \tag{3.2}$$

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 (3.2). Denote by h^* a solution to the original Neyman-Pearson optimization problem:

$$h^* \in \operatorname*{argmin}_{\substack{h \in \mathcal{H} \\ R^-(h) \leq \alpha}} R^+(h)\,, \tag{3.3}$$

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 $\alpha + \epsilon_0$. In a later paper, Cannon et al. (2003) consider problem (3.2) 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. Moreover, Corollary 1 in Scott (2007) provides an oracle inequality for the type II error of a classifier that satisfies an strict constraint on the type I error. However, this result is not directly comparable to the present paper since the rate at which the type II error decreases is not explicitly controlled. This drawback is inherent to methods based on empirical risk minization as opposed to convexified methods as discussed below. Finally, a related work is that of Blanchard et al. (2010) who develop 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. As a matter of fact, the classifiers in all the papers mentioned above can only ensure that $\mathbb{P}(R^-(\hat{h}) > \alpha + \varepsilon_0)$ is small, for some $\epsilon_0 > 0$. However, it is our primary interest to make sure that $R^-(\hat{h}) \leq \alpha$ 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 will be illustrated, 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 control excess type II error simultaneously.

The difficulty was foreseen in the seminal paper Cannon et al. (2002), where it is 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". We have analytically confirmed this opinion, but due to limited space we refer the interested reader to the full version of this paper (Rigollet and Tong, 2011).

To overcome this dilemma, we resort to a continuous convex surrogate as our loss function. In particular, we design a modified version of empirical risk minimization method such that the data-driven classifier $\hat{h}$ has type I error bounded by α with high probability. Moreover, we consider here a class $\mathcal{H}$ that allows a different treatment of the empirical processes involved.

This new approach comes with 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 $\hat{R}^-(h) \leq \alpha + \varepsilon_0/2$ on type I error in (3.2) is satisfied by h^* with high probability, and that this classifier accommodates 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 a convex optimization problem under a stochastic constraint. Such methods have consequences not only in NP classification but also on chance constrained optimization as explained in subsection 5.2.

3.2 Convexified NP classifier

To solve the problem of NP classification (2.3) where the distribution of the observations is unknown, we resort to empirical risk minimization. In view of the arguments presented in the previous subsection, we cannot simply replace the unknown true 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 $\mathcal{H}^{\varphi,a} = \{h \in \mathcal{H}^{\text{conv}} : R^-_\varphi(h) \leq a\}$ be the set of classifiers in $\mathcal{H}^{\text{conv}}$ whose convexified type I errors are bounded from above by a, and let $\mathcal{H}^{\varphi,a}_{n^-} = \{h \in \mathcal{H}^{\text{conv}} : \hat{R}^-_\varphi(h) \leq a\}$ 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

$$\min_{\substack{h \in \mathcal{H}^{\text{conv}} \\ \hat{R}^-_\varphi(h) \leq \alpha - \tau/\sqrt{n^-}}} \hat{R}^+_\varphi(h)\,. \tag{3.4}$$

Note that this problem consists of minimizing a convex function subject to a convex constraint and can therefore be solved by standard algorithms such as (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 include $\tilde{h}^\tau$.

4 Performance Bounds

4.1 Control of type I error

The first challenge is to 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

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

for a proper choice of positive constant τ.

Theorem 4.1 *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*

$$\tau = 4\sqrt{2}L\sqrt{\log\left(\frac{2M}{\delta}\right)}\,. \tag{4.1}$$

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

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

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

$$\mathbb{P}\left[\mathcal{H}^{\varphi,\alpha_\tau}_{n^-} \subset \mathcal{H}^{\varphi,\alpha}\right] \geq 1-\delta\,. \tag{4.2}$$

4.2 Simultaneous control of the two errors

Theorem 4.1 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 type II error is overly deteriorated. Indeed, an extremely small α_τ would certainly ensure a good control of type I error but would deteriorate significantly the best achievable 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 subsection 3.2 with τ defined in (4.1) suffers only a small degradation of its type II error compared to the best achievable. Analogues to classical binary classification, a desirable result is that with high probability,

$$R_\varphi^+(\tilde{h}^{\alpha_\tau}) - \min_{h \in \mathcal{H}^{\varphi,\alpha}} R_\varphi^+(h) \leq \tilde{\Delta}_n(\mathcal{F}), \tag{4.3}$$

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

The following proposition is pivotal to our argument.

Proposition 4.1 *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 $\nu \in (0,\nu_0)$,*

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

This proposition ensures that if the convex surrogate φ is continuous, strengthening the constraint on type I error does not deteriorate too much the optimal 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 optimization as discussed in subsection 5.2.

The next theorem shows that the NP classifier $\tilde{h}^\tau$ defined in subsection 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 4.1.

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 (4.2) for large enough n^-. Indeed, it follows from this inequality that with probability $1-\delta$,

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

Thus, it is sufficient to check if $\mathcal{H}_{n^-}^{\varphi,\varepsilon\alpha-\tau/\sqrt{n^-}}$ 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

$$n_0 \geq \left(\frac{4\tau}{(1-\bar{\varepsilon})\alpha}\right)^2. \tag{4.4}$$

Theorem 4.2 *Let φ, τ, δ and α be the same as in Theorem 4.1, and $\tilde{h}^\tau$ denote any solution to (3.4). Moreover, let Assumption 1 hold and assume that $n^- \geq n_0$ where n_0 is defined in (4.4). Then, the following hold with probability $1-2\delta$,*

$$R^-(\tilde{h}^\tau) \leq R_\varphi^-(\tilde{h}^\tau) \leq \alpha \tag{4.5}$$

and

$$R_\varphi^+(\tilde{h}^\tau) - \min_{h \in \mathcal{H}^{\varphi,\alpha}} R_\varphi^+(h) \leq \frac{4\varphi(1)\tau}{(1-\bar{\varepsilon})\alpha\sqrt{n^-}} + \frac{2\tau}{\sqrt{n^+}}\,. \tag{4.6}$$

In particular, there exits a constant $C > 0$ depending on α, $\varphi(1)$ and $\bar{\varepsilon}$ such that (4.6) yields

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

Note here that Theorem 4.2 is not exactly of the type (4.3). The right hand side of (4.6) goes to zero if both n^- and n^+ go to infinity. Moreover, inequality (4.6) 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,under which parallel results to Theorem 4.5 and Theorem 4.6 are developed, and relegated to Section 6.

5 Extensions

5.1 A Different Sampling Scheme

We now consider a model for observations that is more standard in statistical learning theory (Devroye et al., 1996, Boucheron et al., 2005, see, e.g.,).

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 subsection 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

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

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 4.2, we find that for $n^- \ge n_0$ the event $\mathcal{F}$ satisfies

$$\mathbb{P}(\mathcal{F}|N^- = n^-) \ge 1 - 2\delta\,. \tag{5.1}$$

We obtain the following corollary of Theorem 4.2.

Corollary 5.1 *Let φ, τ, δ and α be the same as in Theorem 4.1, and $\tilde{h}^\tau_n$ 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* (4.4)*, we have with probability $(1-2\delta)(1-e^{-\frac{n(1-p)^2}{2}})$,*

$$R^-(\tilde{h}^\tau_n) \le R^-_\varphi(\tilde{h}^\tau_n) \le \alpha \tag{5.2}$$

and

$$R^+_\varphi(\tilde{h}^\tau_n) - \min_{h \in \mathcal{H}^{\varphi,\alpha}} R^+_\varphi(h) \le \frac{4\varphi(1)\tau}{(1-\bar{\varepsilon})\alpha\sqrt{N^-}} + \frac{2\tau}{\sqrt{N^+}}\,. \tag{5.3}$$

Moreover, with probability $1 - 2\delta - e^{-\frac{n(1-p)^2}{2}} - e^{-\frac{np^2}{2}}$, we have simultaneously (5.2) *and*

$$R^+_\varphi(\tilde{h}^\tau_n) - \min_{h \in \mathcal{H}^{\varphi,\alpha}} R^+_\varphi(h) \le \frac{4\sqrt{2}\varphi(1)\tau}{(1-\bar{\varepsilon})\alpha\sqrt{n(1-p)}} + \frac{2\sqrt{2}\tau}{\sqrt{np}}\,. \tag{5.4}$$

5.2 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:

$$\min_{\lambda \in \Lambda} f(\lambda) \quad \text{s.t.} \quad \mathbb{P}\{F(\lambda, \xi) \le 0\} \ge 1 - \alpha, \tag{5.1}$$

where ξ 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. For simplicity, we take F to be scalar valued but extensions to vector valued functions and conic orders are considered in (see, e.g., Ben-Tal et al., 2009, Chapter 10). Moreover, it is standard to assume that $F(\cdot, \xi)$ is convex almost surely.

Problem (5.1) 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

$$\min_{\lambda \in \Lambda} f(\lambda) \quad \text{s.t.} \quad F(\lambda, \xi_i) \leq 0, i = 1, \ldots, n. \tag{5.2}$$

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 (5.2) is feasible for (5.1). The authors did not address the control of the term $f(\hat{\lambda}^{sc}) - f^*$ where f^* denotes the optimal objective value in (5.1).

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

$$\min_{\lambda \in \Lambda, t \in \mathbb{R}^s} f(\lambda) \quad \text{s.t.} \quad G(\lambda, t) \leq 0, \tag{5.3}$$

in which t is some additional instrumental variable and where $G(\cdot, t)$ is convex. The problem (5.3) provides a conservative convex approximation to (5.1), in the sense that every λ feasible for (5.3) is also feasible for (5.1). Nemirovski and Shapiro (2006) considered a particular class of conservative convex approximation where the key step is to replace $P_\xi\{F(\lambda, \xi) \geq 0\}$ by $\mathbb{E}\varphi(F(\lambda, \xi))$ in (5.1), where φ a nonnegative, nondecreasing, convex function that takes 1 at 0. Nemirovski and Shapiro (2006) discuss several choices of φ including hinge loss and exponential loss, 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. Denote by P^- the conditional distribution of X given $Y = -1$. In a parallel form of (5.1), we cast our target problem as

$$\min_{\lambda \in \Lambda} R^+(\mathsf{h}_\lambda) \quad \text{s.t.} \quad P^-\{\mathsf{h}_\lambda(X) \leq 0\} \geq 1 - \alpha, \tag{5.4}$$

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

The problem (5.4) differs from (5.1) in that $R^+(\mathsf{h}_\lambda)$ is not a convex function of λ. Replacing $R^+(\mathsf{h}_\lambda)$ by $R^+_\varphi(\mathsf{h}_\lambda)$ turns (5.4) into a standard chance constrained optimization problem:

$$\min_{\lambda \in \Lambda} R^+_\varphi(\mathsf{h}_\lambda) \quad \text{s.t.} \quad P^-\{\mathsf{h}_\lambda(X) \leq 0\} \geq 1 - \alpha. \tag{5.5}$$

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 (5.1). 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 (5.5). Instead, what possible is to ensure that if we learned $\hat{f}_\lambda$ from the sample, this constraint is satisfied with high probability $1 - \delta$, i.e., the classifier is approximately feasible for (5.5). In retrospect, our approach to (5.5) is an innovative hybrid between the analytical approach based on convex surrogates and the scenario approach.

We do have structural assumptions on the scope of the problem. Let $g_j, j \in \{1, \ldots, M\}$ 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 (5.1):

$$\min_{\lambda \in \Lambda} f(\lambda) \quad \text{s.t.} \quad \mathbb{E}[\varphi(F(\lambda, \xi))] \leq \alpha, \tag{5.6}$$

where φ is a L-Lipschitz convex surrogate, $L > 0$. Suppose that we observe a sample $(\xi_1, \ldots, \xi_n)$ that are independent copies of ξ. Denote by f^*_φ the value of the objective at the optimum in (5.6). We propose to approximately solve the above problem by

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

for some $\tau > 0$ to be defined. Denote by $\tilde{\lambda}$ any solution to this problem. The following theorem summarizes our contribution to chance constrained optimization.

Theorem 5.1 *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*

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

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

(i) $\tilde{\lambda}$ is feasible for (5.1).

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

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

we have

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

The proof essentially follows that of Theorem 4.2 and we omit it. The limitations of Theorem 5.1 include rigid structural assumptions on the function F and on the set Λ. Also, we did not address the effect of replacing the indicator function by a convex surrogate; this investigation is beyond the scope of this paper.

6 Proofs

6.1 Proof of Theorem 4.1

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 $\{h_1, \ldots, h_M\}$ 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 $X_1, \ldots, X_n \in \mathcal{X}$ 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 $\varepsilon_1, \ldots, \varepsilon_n$ are i.i.d. Rademacher random variables such that $\mathbb{P}(\varepsilon_i = 1) = \mathbb{P}(\varepsilon_i = -1) = 1/2$ for $i = 1, \ldots, n$.

Lemma 6.1 *Fix $L > 0, \delta \in (0,1)$. Let $X_1, \ldots, X_n$ 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*

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

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, 2008, Section 2), we find

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

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

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

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

$$\begin{aligned}
\mathbb{P}\left(\sup_{\lambda\in\Lambda} |(P_n - P)(\varphi\circ\mathsf{h}_\lambda)| > t\right) &\le \frac{1}{\Phi(st)}\mathbb{E}\Phi\left(s\sup_{\lambda\in\Lambda} |(P_n - P)(\bar\varphi\circ\mathsf{h}_\lambda)|\right) \\
&\le \frac{1}{\Phi(st)}\mathbb{E}\Phi\left(4Ls\max_{1\le j\le M} |R_n(h_j)|\right) . \qquad (6.7)
\end{aligned}$$

Moreover, since Φ is increasing,

$$\begin{aligned}
\mathbb{E}\Phi\left(4Ls\max_{1\le j\le M} |R_n(h_j)|\right) &= \mathbb{E}\max_{1\le j\le M} \Phi\left(4Ls\,|R_n(h_j)|\right) \\
&\le \sum_{j=1}^M \mathbb{E}\left[\Phi\left(4LsR_n(h_j)\right) \vee \Phi\left(-4LsR_n(h_j)\right)\right] \\
&\le 2\sum_{j=1}^M \mathbb{E}\Phi\left(4LsR_n(h_j)\right) . \qquad (6.8)
\end{aligned}$$

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

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

where cosh is the hyperbolic cosine function and where in the inequality, we used the fact that $|h_j(X_i)| \le 1$ for any i, j and $\cosh(x) \le \exp(x^2/2)$. Together with (6.7) and (6.8), it yields

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

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 4.1. Note first that from the properties of φ, $R^-(h) \le R^-_\varphi(h)$. Next, we have for any data-dependent classifier $h \in \mathcal{H}^{\text{conv}}$ such that $\hat R^-_\varphi(h) \le \alpha_\tau$:

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

Lemma 6.1 implies that, with probability $1-\delta$

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

The previous two displays imply that $R^-_\varphi(h) \le \alpha$ with probability $1-\delta$, which completes the proof of Theorem 4.1.

6.2 Proof of Proposition 4.1

The proof of this proposition builds upon the following lemma.

Lemma 6.2 *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 $\alpha' > \alpha$, $\{\mathsf{h}_\lambda \in \mathcal{H}^{\text{conv}} : R^-_\varphi(\mathsf{h}_\lambda) \le \alpha\} \subset \{\mathsf{h}_\lambda \in \mathcal{H}^{\text{conv}} : R^-_\varphi(\mathsf{h}_\lambda) \le \alpha'\}$.

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 exits $\lambda^* \in \Lambda$ 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 exits $\lambda_1, \lambda_2 \in \Lambda$ 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

$$R^-_\varphi(\mathsf{h}_{\bar\lambda_\theta}) \le \theta R^-_\varphi(\mathsf{h}_{\lambda_1}) + (1-\theta) R^-_\varphi(\mathsf{h}_{\lambda_2}) \le \theta\alpha_1 + (1-\theta)\alpha_2 = \bar\alpha_\theta\,,$$

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

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

□

We now complete the proof of Proposition 4.1. 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

$$\gamma(\alpha-\nu) - \gamma(\alpha) \le \varphi(1)\frac{\nu}{\nu_0 - \nu} \quad 0 < \nu < \nu_0\,. \tag{6.9}$$

Lemma 6.2 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

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

6.3 Proof of Theorem 4.2

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

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

Lemma 6.1 implies

$$\mathbb{P}(\mathcal{E}^-) \wedge \mathbb{P}(\mathcal{E}^+) \ge 1-\delta\,. \tag{6.10}$$

Note first Theorem 4.1 implies that (4.5) holds with probability $1-\delta$. Observe now that the l.h.s of (4.6) can be decomposed as

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

where

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

To bound A_1 from above, observe that

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

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

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

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

Finally, to control A_3, observe that under Assumption 1, Proposition 4.1 can be applied with $\nu = 2\tau/\sqrt{n^-}$ and $\nu_0 = (1-\bar{\varepsilon})\alpha$. Indeed, the assumptions of the theorem imply that $\nu \leq \nu_0/2$. It yields

$$A_3 \leq \frac{4\varphi(1)\tau}{(1-\bar{\varepsilon})\alpha\sqrt{n^-}}\,.$$

Combining the bounds on A_1, A_2 and A_3 obtained above, we find that (4.6) holds on the event $\mathcal{E}^- \cap \mathcal{E}^+$ that has probability at least $1-2\delta$ in view of (6.10).

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

6.4 Proof of Corollary 5.1

We will use the following Lemma to bound the left tail of a binomial distribution, whose proof we omit for this short version.

Lemma 6.3 *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*

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

Now prove (5.3),

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

where in the last inequality, we used (5.1). Applying now Lemma 6.3, we obtain

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

Therefore,

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

which completes the proof of (5.3).

The proof of (5.4) follows by observing that

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

where

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

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 6.3, 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

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

which completes the proof of the corollary.

References

P. Bartlett, M. Jordan, and J. Mcauliffe. Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 2006.

Aharon Ben-Tal, Laurent El Ghaoui, and Arkadi Nemirovski. *Robust optimization*. Princeton Series in Applied Mathematics. Princeton University Press, Princeton, NJ, 2009. ISBN 978-0-691-14368-2.

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

Stéphane Boucheron, Olivier Bousquet, and Gábor Lugosi. Theory of classification: a survey of some recent advances. *ESAIM Probab. Stat.*, 9:323–375, 2005. ISSN 1292-8100. doi: 10.1051/ps:2005018.

Stephen Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge University Press, Cambridge, 2004. ISBN 0-521-83378-7.

Giuseppe C. Calafiore and Marco C. Campi. The scenario approach to robust control design. *IEEE Trans. Automat. Control*, 51(5):742–753, 2006. ISSN 0018-9286. doi: 10.1109/TAC.2006.875041.

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.

Luc Devroye, László Györfi, and Gábor Lugosi. *A probabilistic theory of pattern recognition*, volume 31 of *Applications of Mathematics (New York)*. Springer-Verlag, New York, 1996. ISBN 0-387-94618-7.

M. Han, D. Chen, and Z. Sun. Analysis to neyman-pearson classification with convex loss function. *Analysis in Theory and Applications*, 24(1):18–28, 2008.

V. Koltchinskii. *Saint Flour Lectures Oracle Inequalities in Empirical Risk Minimization and Sparse Recovery Problems*. 2008.

Constantino M. Lagoa, Xiang Li, and Mario Sznaier. Probabilistically constrained linear programs and risk-adjusted controller design. *SIAM J. Optim.*, 15(3):938–951 (electronic), 2005. ISSN 1052-6234. doi: 10.1137/S1052623403430099.

E. L. Lehmann and Joseph P. Romano. *Testing statistical hypotheses*. Springer Texts in Statistics. Springer, New York, third edition, 2005. ISBN 0-387-98864-5.

Arkadi Nemirovski and Alexander Shapiro. Convex approximations of chance constrained programs. *SIAM J. Optim.*, 17(4):969–996, 2006. ISSN 1052-6234. doi: 10.1137/050622328.

András Prékopa. *Stochastic programming*, volume 324 of *Mathematics and its Applications*. Kluwer Academic Publishers Group, Dordrecht, 1995. ISBN 0-7923-3482-5.

Philippe Rigollet and Xin Tong. Neyman-pearson classification, convexity and stochastic constraints. arXiv:1102.5750, February 2011.

R. Tyrrell Rockafellar. *Convex analysis*. Princeton Landmarks in Mathematics. Princeton University Press, Princeton, NJ, 1997. ISBN 0-691-01586-4. Reprint of the 1970 original, Princeton Paperbacks.

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

C. Scott. Comparison and design of neyman-pearson classifiers. 2005.

C. Scott. Performance measures for neyman-pearson classification. *IEEE Transactions on Information 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.