

Overlaying classifiers: a practical approach to optimal scoring

Stéphan Cléménçon*

LTCI - UMR Telecom ParisTech & CNRS

Nicolas Vayatis

CMLA - Ecole Normale Supérieure de Cachan, CNRS & UniverSud

November 21, 2008

Abstract

The ROC curve is one of the most widely used visual tool to evaluate performance of scoring functions regarding their capacities to discriminate between two populations. It is the goal of this paper to propose a statistical learning method for constructing a scoring function with nearly optimal ROC curve. In this bipartite setup, the target is known to be the regression function up to an increasing transform and solving the optimization problem boils down to recovering the collection of level sets of the latter, which we interpret here as a continuum of imbricated classification problems. We propose a discretization approach, consisting in building a finite sequence of N classifiers by *constrained empirical risk minimization* and then constructing a piecewise constant scoring function $s_N(x)$ by overlaying the resulting classifiers. Given the functional nature of the ROC criterion, the accuracy of the ranking induced by $s_N(x)$ can be conceived in a variety of ways, depending on the distance chosen for measuring closeness to the optimal curve in the ROC space. By relating the ROC curve of the resulting scoring function to piecewise linear approximates of the optimal ROC curve, we establish the consistency of the method as well as rate bounds to control its generalization ability in sup-norm. Eventually, we also highlight the fact that, as a byproduct, the algorithm proposed provides an accurate estimate of the optimal ROC curve.

Keywords: Statistical learning, bipartite ranking, ROC curve, piecewise linear approximation, minimum volume set estimation, density level set, scoring function, AUC criterion, sup-norm.

*Address of corresponding author: Stéphan Cléménçon, TSI - Telecom ParisTech - 46, rue Barrault - 75634 Paris cedex 13, France - Email: stephan.clemencon@telecom-paristech.fr

1 Introduction

In the past years, statistical learning theory has witnessed impressive developments. This approach was mainly developed through the study of empirical risk minimization procedures and algorithms for standard problems such as classification and regression problems. A learning method can be described by a risk measure and some data-based optimization principle. In classification, the optimization criteria are risk functionals, such as the classification error or its convex surrogates, which take scalar values. However, in many important applications such as medical diagnosis, credit-risk screening or information retrieval, performance is monitored by a function-valued criterion. Function-like performance measures, such as *Receiver Operating Characteristic* (ROC), see [vT68], [Ega75] or the *Precision-Recall curve* (Refs.), allow to take into account various constraints in the decision process. In the present paper, we focus on scoring applications where the problem is to rank the data from binary label information. This problem is also known as the bipartite ranking problem in the machine learning literature. We will also focus on the ROC curve which permits, through a graphical display, to judge rapidly how a scoring rule discriminates the two populations (positive against negative). A scoring rule whose ROC curve is close to the diagonal line does not discriminate at all, while the one lying above all others is the best possible choice. From a statistical learning perspective, risk minimization (or performance maximization) strategies for bipartite ranking have been based mostly on a popular summary of the ROC curve known as the *Area Under the ROC Curve* (AUC - see [CLV08], [FISS03], [AGH⁺05], [UAG05], ...) which corresponds to the L_1 -metric on the space of ROC curves.

In the present paper, we propose a statistical methodology to estimate the optimal ROC curve in a stronger sense than the AUC, namely in the the supremum norm. In the same time, we will explain how to build a nearly optimal scoring function. Our approach is based on a simple observation: optimal scoring functions can be represented from the collection of level sets of the regression function. Hence, the bipartite ranking problem may be viewed as a 'continuum' of classification problems with asymmetric costs where the targets are the level sets. In a nonparametric setup, regression or density level sets can be estimated with plug-in methods ([Cav97], [RV06], [AA07], [WN07], ...). Here, we follow the work in [SN06] on minimum-volume set estimation and adapt it to our problem. We provide rates of convergence with which a point of the optimal ROC curve can be recovered according to this principle. We also develop a practical ranking method based on a discretization of the original problem. From the resulting classifiers and their related empirical errors, we show how to build a linear-by-part estimate of the optimal ROC curve and a quasi-optimal piecewise constant scoring function. Rate bounds in terms of sup-norm in the ROC space for these procedures are also established.

The rest of the paper is organized as follows. In Section 2, we present the scoring problem and recall key notions of ROC analysis. In Section 3, we describe the approach of overlaying classifiers used to approximate optimal scoring rules and introduce the RANKOVER

algorithm. We study statistical performance of the output of this algorithm in Section 4 and derive the rate of convergence of an empirical estimate to the optimal ROC curve. In Section 5, we consider the subproblem of constrained empirical risk minimization problem. The main strategy described as empirical minimum-volume set estimation is provided, fast rates of convergence are established and alternative methods are also discussed. Proofs are postponed to the Appendix section.

2 The scoring approach to bipartite ranking

In this section, we first set out the notations and recall the key concepts related to the bipartite ranking problem that shall be needed throughout the paper.

2.1 Notations and setup

Let \mathcal{X} be a measurable space which can be thought as a high-dimensional Euclidean space. Consider a random pair (X, Y) over $\mathcal{X} \times \{-1, +1\}$ where X is called the descriptor and Y is the binary label. We denote by $P = (\mu, \eta)$ the distribution of (X, Y) , where μ is the marginal distribution of X and η is the *regression function* (up to an affine transformation): $\eta(x) = \mathbb{P}\{Y = 1 \mid X = x\}$, $x \in \mathcal{X}$. We will also denote by $p = \mathbb{P}\{Y = +1\}$ the expected proportion of positive labels. We also denote by $G(dx)$ and $H(dx)$ the conditional distributions of the random variable X given $Y = +1$ and given $Y = -1$ respectively. In the sequel, we assume that these distributions are equivalent and absolutely continuous with respect to Lebesgue measure. We point out that, equipped with these notations, one may write $\mu = pG + (1 - p)H$ and $dG/dH(x) = (1 - p)\eta(x)/(p(1 - \eta(x)))$.

The scoring problem. A possible and natural approach in order to rank the objects $x \in \mathcal{X}$ is to map them onto \mathbb{R} through a certain measurable function $s : \mathcal{X} \rightarrow \mathbb{R}$ and use the natural order on the real line. We call such a function s a *scoring function* and the statistical challenge is to build an s from sampling data $\mathcal{D}_n = \{(X_1, Y_1), \dots, (X_n, Y_n)\}$ which mimics the ranking induced by the regression function η . Hence, ideally, the higher the score $s(X)$ is, the more likely should one observe $Y = +1$. We naturally define the class of optimal scoring functions for bipartite ranking as the class of strictly increasing transforms of the regression function η .

Definition 1 (Optimal scoring functions) *The class of optimal scoring functions is given by the set*

$$\mathcal{S}^* = \{ s^* = T \circ \eta \mid T : [0, 1] \rightarrow \mathbb{R} \text{ strictly increasing} \}.$$

The statistical problem consists in finding a scoring function as "close" as possible to the class \mathcal{S}^* from the i.i.d. sample \mathcal{D}_n . The concept of ROC ANALYSIS provides a means of measuring the quality of a scoring function through distances in a function space and, thus, a constructive meaning to the notion of close scoring functions.

ROC analysis. We now recall the concept of ROC curve and explain why it is a natural choice of performance measure for the ranking problem with classification data. We consider here *true* ROC curves which correspond to the situation where the underlying distribution is known. First, we need to introduce some notations. For a given scoring rule s , the conditional cumulative distribution functions of the random variable $s(X)$ are denoted by

$$\begin{aligned} G_s(z) &= \mathbb{P}\{s(X) \leq z \mid Y = +1\} , \\ H_s(z) &= \mathbb{P}\{s(X) \leq z \mid Y = -1\} , \end{aligned}$$

for all $z \in \mathbb{R}$. We also set, $\bar{G}_s(z) = 1 - G_s(z)$ and $\bar{H}_s(z) = 1 - H_s(z)$ to be $s(X)$'s the residual conditional cumulative distribution functions. The residual cdf \bar{G}_s is also called the *true positive rate* while \bar{H}_s is the *false positive rate*. When $s = \eta$, we shall denote the previous functions by G^* , H^* , \bar{G}^* , \bar{H}^* respectively. We introduce the notation $Q(Z, \alpha)$ to denote the quantile of order $1 - \alpha$ for the distribution of a random variable Z conditioned on the event $Y = -1$. In particular, the following quantile will be of interest:

$$Q^*(\alpha) = Q(\eta(X), \alpha) = \bar{H}^{*-1}(\alpha) ,$$

where we have used here the notion of generalized inverse $F^{-1}(z) = \inf\{t \in \mathbb{R} \mid F(t) \geq z\}$ of a càdlàg function F . We now turn to the definition of the ROC curve as the PP-plot of the true positive rate against the false positive rate.

Definition 2 (True ROC curve) *The ROC curve of a scoring function s is the parametric curve:*

$$z \mapsto (\bar{H}_s(z), \bar{G}_s(z))$$

for thresholds $z \in \mathbb{R}$. If H_s has no flat parts, the ROC curve can also be defined as the plot of the mapping:

$$\text{ROC}(s, \cdot) : \alpha \in [0, 1] \mapsto \bar{G}_s \circ \bar{H}_s^{-1}(\alpha) = \bar{G}_s(Q(s(X), \alpha)) .$$

For $s = \eta$, we take the notation $\text{ROC}^*(\alpha) = \text{ROC}(\eta, \alpha)$.

By convention, points of the curve corresponding to possible jumps are connected by line segments, so that the ROC curve is always continuous. We point out that, equipped with this usual convention, the ROC curve of any piecewise constant scoring function is linear-by-parts.

Optimality. As a functional criterion, the ROC curve induces a partial order over the space of all scoring functions. A scoring function $s_1(x)$ will be said more accurate than a competitor $s_2(x)$ if and only if its ROC curve is above the one of $s_2(x)$ everywhere, *i.e.* for all $\alpha \in [0, 1]$:

$$\text{ROC}(s_2, \alpha) \leq \text{ROC}(s_1, \alpha).$$

Equivalently, this condition means that the test defined by the statistic $s_1(X)$ for testing the null hypothesis $\mathcal{H}_0 : "Y = -1"$ against the alternative $\mathcal{H}_1 : "Y = +1"$ is *uniformly more powerful* than the one defined by $s_2(X)$, the quantity $\text{ROC}(s, \alpha)$ representing simply the power of the test of exact level α for testing \mathcal{H}_0 based on the diagnostic statistic $s(X)$. We expect optimal scoring functions to be those for which the ROC curve dominates all the others for all $\alpha \in (0, 1)$. The next proposition highlights the fact that the ROC curve is relevant when evaluating performance in the bipartite ranking problem.

Proposition 3 *The class \mathcal{S}^* of optimal scoring functions provides the best possible ranking with respect to the ROC curve. Indeed, for any scoring function s , we have:*

$$\forall \alpha \in (0, 1) , \quad \text{ROC}^*(\alpha) \geq \text{ROC}(s, \alpha) ,$$

and

$$\forall s^* \in \mathcal{S}^* , \forall \alpha \in (0, 1) , \quad \text{ROC}(s^*, \alpha) = \text{ROC}^*(\alpha) .$$

Regularity. In this paper, we will assume that ROC^* is twice differentiable with bounded second derivative. The assumption of twice differentiability of the optimal curve ROC^* can be translated in terms of the regularity of the conditional distributions of the random variable $\eta(X)$. Indeed, assume that the cumulative distribution functions G^* and H^* are both differentiable and that $H^{*'}$ is continuous and bounded by below by some strictly positive constant on its support. Then, from Proposition 8 in [CV08d], we have $\forall \alpha \in]0, 1[$:

$$(\text{ROC}^*)'(\alpha) = \frac{Q^*(\alpha)}{1 - Q^*(\alpha)} \cdot \frac{p}{1 - p} .$$

In order to guarantee that ROC^* is differentiable at 0, we need to assume that $1 - \eta(X)$ stays bounded away from zero almost-surely or equivalently that $Q^*(0) < 1$ (so that: $\eta(X) \leq Q^*(0) < 1$ a.s.). As we have:

$$\frac{dG}{dH}(X) = \frac{1 - p}{p} \cdot \frac{\eta(X)}{1 - \eta(X)} ,$$

this assumption simply means that the likelihood ratio is upper bounded almost surely, by $(1 - p)Q^*(0)/(p(1 - Q^*(0)))$ namely. In addition, as we also have:

$$\frac{dG}{dH}(X) = \frac{dG^*}{dH^*}(\eta(X)) ,$$

see Corollary 7 in [CV08d], this also means that the ratio $dG^*/dH^*(u)$ remains bounded in the neighborhood of $Q^*(0)$. This is also enough to ensure that ROC^* is twice differentiable on $[0, 1]$, since: $\forall \alpha \in [0, 1]$,

$$(\text{ROC}^*)''(\alpha) = \frac{(Q^*)'(\alpha)}{(1 - Q^*(\alpha))^2} \cdot \frac{p}{1 - p} .$$

We point out that these assumptions on the regularity of the ROC curve are strong requirements and restrict significantly the range of distributions which can be considered for modeling the data. Relaxing these assumptions and building consistent estimators of the optimal ROC curve is the subject of work under progress.

2.2 Metrics in ROC space, excess risk and optimal scoring functions

We may now compare the ranking performance of a given s to the one of the optimal elements in \mathcal{S}^* in terms of closeness of their ROC curves, closeness being possibly measured by any metric in the space $\mathbb{D}([0, 1])$ of càdlàg curves $f : [0, 1] \rightarrow \mathbb{R}$. Let us denote by $d(s, s^*)$ the distance describing the criterion of interest. The statistical problem consists in constructing a scoring function s_n based on the available data \mathcal{D}_n such that $d(s_n, s^*)$ can be upper bounded with high probability in terms of the sample size n , the level of confidence and possibly some structural parameters such as the complexity of the class \mathcal{S} of candidate scoring functions. In statistical learning theory, standard problems, such as classification or regression, benefit from the 'excess-risk' decomposition of the risk measure. Hence $d(s, s^*)$ can be written as a difference $A(s) - A(s^*)$ so that minimizing $d(s, s^*)$ is equivalent to minimizing $A(s)$. Then, by M -estimation arguments which are now standard (see [BBL05]), it is possible to show that strategies based on the minimization of an empirical counterpart of $A(s)$ can be efficient.

In the case of ranking/scoring applications, many different metrics can be introduced. Here we focus on the L_1 and L_∞ cases.

The L_1 -distance and the AUC criterion. Consider first the L_1 -distance between ROC curves as a measure of closeness for scoring functions. For any scoring function, we set:

$$d_1(s, \eta) = \|\text{ROC}(s, \cdot) - \text{ROC}^*(\cdot)\|_1 = \int_0^1 |\text{ROC}^*(\alpha) - \text{ROC}(s, \alpha)| \, d\alpha .$$

By Proposition 3, we have:

$$d_1(s, \eta) = \|\text{ROC}^*\|_1 - \|\text{ROC}(s, \cdot)\|_1 .$$

In this case, rendering $d(s, \eta)$ minimum boils down to maximizing a popular quantity known as the *Area Under the ROC curve* (or AUC, see [HM82]):

$$\text{AUC}(s) = \|\text{ROC}(s, \cdot)\|_1 = \int_0^1 \text{ROC}(s, \alpha) \, d\alpha .$$

In this particular case, the analysis of empirical risk minimization strategies is greatly facilitated by the fact that the AUC performance measure may be interpreted in a probabilistic fashion and natural estimates of the risk are of the form of a U -statistic.

Proposition 4 ([CLV05]) *For any scoring function s such that H_s and G_s are continuous distribution functions, we have:*

$$\begin{aligned} \text{AUC}(s) &= \mathbb{P}\{s(X) > s(X') \mid Y = 1, Y' = -1\} \\ &= \frac{1}{2p(1-p)} \mathbb{P}\{(s(X) - s(X'))(Y - Y') > 0\} . \end{aligned}$$

where (X, Y) and (X', Y') are i.i.d. copies.

From this observation, ranking can be indeed interpreted as classification of pairs of observations. We refer to [CLV08] for a systematic study of related empirical and convex risk minimization strategies which involve \mathbf{U} -statistics. From a machine learning perspective, there is a growing literature in which existing algorithms are adapted in order to perform AUC optimization (such as, for instance: [CM04], [Rak04], [YDMW03]).

However, the AUC as a summary criterion presents some drawbacks since two scoring functions can have the same AUC but behave very differently in the ROC space. Hence, more stringent notions of distance need to be considered.

The L_∞ -distance. As an example of a strong notion of distance, we propose to study the distance induced by the L_∞ -norm:

$$d_\infty(s, s^*) = \|\text{ROC}(s, \cdot) - \text{ROC}^*(\cdot)\|_\infty = \sup_{\alpha \in [0,1]} (\text{ROC}^*(\alpha) - \text{ROC}(s, \alpha)) .$$

The main difficulty for dealing with such a criterion from the perspective of empirical risk minimization is that there is no simple empirical counterpart. Indeed, in this case the usual 'excess-risk' decomposition of the form $d(s, s^*) = A^* - A(s)$ does not hold and it is not straightforward how to relate the empirical risk minimization (ERM) approach to the d_∞ criterion.

The goal of this paper is to show that empirical risk minimization procedures can be tailored for the ranking/scoring problem under the criterion induced by the L_∞ -norm. The ERM strategy will here be combined with an approximation stage, involving very mild smoothness assumptions for the optimal curve ROC^* . As a byproduct of the analysis we will also provide a statistical estimation of the optimal ROC curve which can also be of interest *per se*.

Optimal scoring functions as overlaid classification rules. From the angle embraced in this paper, ranking amounts to recovering the decreasing collection of level sets of the regression function $\eta(x)$:

$$\{\{x \in \mathcal{X} \mid \eta(x) > u\}, u \in [0, 1]\},$$

without necessarily disposing of the corresponding levels. Indeed, any scoring function of the form

$$s^*(x) = \int_0^1 \mathbb{I}\{\eta(x) > Q^*(\alpha)\} d\nu(\alpha), \tag{1}$$

where $\nu(d\alpha)$ is an arbitrary finite positive measure on $[0, 1]$ with same support as $H^*(d\alpha)$, is optimal with respect to the ROC criterion. Notice that $s^*(x) = H^*(\eta(x))$ when ν is chosen to be the Lebesgue measure. The next proposition also illustrates this view on the problem. We set the notations:

$$\begin{aligned} R_\alpha^* &= \{x \in \mathcal{X} \mid \eta(x) > Q^*(\alpha)\}, \\ R_{s,\alpha} &= \{x \in \mathcal{X} \mid s(x) > Q(s(X), \alpha)\}. \end{aligned}$$

Proposition 5 *Let s be a scoring function and $\alpha \in (0, 1)$ such that $Q^*(\alpha) < 1$. Suppose additionally that the cdf H_s (respectively, H^*) is continuous at $Q(s(X), \alpha)$ (resp. at $Q^*(\alpha)$). Then, we have:*

$$\text{ROC}^*(\alpha) - \text{ROC}(s, \alpha) = \frac{\mathbb{E}(|\eta(X) - Q^*(\alpha)| \mathbb{I}\{X \in R_\alpha^* \Delta R_{s, \alpha}\})}{p(1 - Q^*(\alpha))}$$

where Δ denotes the symmetric difference between sets.

This result shows that the pointwise difference between the dominating ROC curve and the one related to a candidate scoring function s may be interpreted as the error made in recovering the specific level set R_α^* through $R_{s, \alpha}$.

In contrast, standard binary classification amounts to recover a single, very specific, η -level set, namely $\{x \in \mathcal{X} \mid \eta(x) > 1/2\}$. It is well-known that the latter corresponds to the classifier $C^*(X) = 2 \cdot \mathbb{I}\{\eta(X) > 1/2\} - 1$ with minimum classification error $L(C) = \mathbb{P}\{Y \neq C(X)\}$ with $C : \mathcal{X} \rightarrow \{-1, +1\}$.

Remark 1 (ON THE EXCESS OF RISK) We point out that Proposition 5 generalizes the well-known relationship in the classification setup, see [DGL96]:

$$\begin{aligned} L(C) - L(C^*) &= p(G(C^*) - G(C)) + (1 - p)(H(C) - H(C^*)) \\ &= \mathbb{E}[|2\eta(X) - 1| \mathbb{I}\{X \in R^* \Delta R\}], \end{aligned}$$

where $R = \{x \in \mathcal{X} \mid C(x) = +1\}$ and $R^* = R_{\alpha^*}^*$ with $Q^*(\alpha^*) = 1/2$.

Plug-in scoring rules. A possible angle to approach optimal scoring rules is the *plug-in* approach, see [DGL96]. The idea of plug-in consists of using an estimate $\hat{\eta}(x)$ of the regression function as a scoring function. It is expected that, whenever $\hat{\eta}(x)$ is close to $\eta(x)$ in a certain sense, then $\text{ROC}(\hat{\eta}, \cdot)$ and ROC^* are also close.

Proposition 6 *Let $\hat{\eta}(x)$ be an approximant of $\eta(x)$ and suppose that $G_{\hat{\eta}}(dx)$ and $H_{\hat{\eta}}(dx)$ are continuous distribution functions.*

(i) *We have:*

$$\text{AUC}^* - \text{AUC}(\hat{\eta}) \leq \frac{1}{p(1 - p)} \mathbb{E}(|\hat{\eta}(X) - \eta(X)|).$$

(ii) *Assume in addition that H^* has a density which is bounded by below on $[0, 1]$: $\exists c > 0$ such that $\forall \alpha \in [0, 1]$, $\frac{dH^*}{d\alpha}(\alpha) \geq c^{-1}$. Then, we have: $\forall \alpha \in [0, 1]$ such that $Q^*(\alpha) < 1$,*

$$\text{ROC}^*(\alpha) - \text{ROC}(\hat{\eta}, \alpha) \leq \frac{c \mathbb{E}(|H^*(\eta(X)) - H_{\hat{\eta}}(\hat{\eta}(X))|)}{p(1 - Q^*(\alpha))}.$$

It clearly follows from (i) that a $L_1(\mu)$ -consistent estimator, *i.e.* an estimator $\hat{\eta}_n(x)$ such that $\mathbb{E}(|\hat{\eta}_n(X) - \eta(X)|) \rightarrow 0$ as $n \rightarrow \infty$ with probability one, yields a consistent ranking rule in the AUC-sense. It is however much more difficult to guarantee the pointwise convergence $\text{ROC}^*(\alpha) - \text{ROC}(\hat{\eta}_n, \alpha) \rightarrow 0$: this would be again implied by $L_1(\mu)$ -consistency provided that $\hat{\eta}_n(X)$ has a density uniformly bounded in n . In addition, plug-in rules face computational difficulties when dealing with high-dimensional data ([GKKW02]). These observations provide the motivation for exploring algorithms based on direct empirical risk minimization.

3 Ranking by overlaying classifiers

The approach considered in this paper consists of a discretization of the ranking problem. The main idea is to build a scoring function close to the one obtained by overlaying a finite collection of level sets $R_{\alpha_1}^*, \dots, R_{\alpha_K}^*$, where the subdivision $\sigma: 0 = \alpha_0 < \alpha_1 \leq \dots \leq \alpha_K \leq \alpha_{K+1} = 1$ is fixed in advance and K is a tuning parameter that controls the complexity of the method:

$$s_\sigma^*(x) = \sum_{i=1}^K \mathbb{I}\{x \in R_{\alpha_i}^*\}, \quad (2)$$

which may be seen as a discrete version of (1), where ν is taken as the *point measure* $\sum_{i=1}^K \delta_{\alpha_i}$, where δ_x denotes the Dirac mass at x .

3.1 Piecewise linear approximation of the optimal ROC curve

Observe that the ROC curve of the stepwise scoring function $s_\sigma^*(x)$ is the broken line that connects the knots $\{(\alpha_i, \text{ROC}^*(\alpha_i)), 0 \leq i \leq K+1\}$. In order to explicit the latter, we classically consider the "hat functions" related to the meshgrid $\{\alpha_i; 0 \leq i \leq K+1\}$: $\forall i \in \{1, \dots, K\}, \forall \alpha \in [0, 1]$,

$$\Phi_i^*(\alpha) = \Phi(\alpha, \alpha_{i-1}, \alpha_i) - \Phi(\alpha, \alpha_i, \alpha_{i+1}),$$

and $\Phi_{K+1}^*(\alpha) = \Phi(\alpha, \alpha_K, 1)$, where for all $\alpha' < \alpha''$,

$$\Phi(\alpha, \alpha', \alpha'') = \frac{\alpha - \alpha'}{\alpha'' - \alpha'} \mathbb{I}\{\alpha \in [\alpha', \alpha'']\}.$$

Equipped with these notations, the ROC curve of the piecewise constant scoring function (2) is the linear-by-parts curve given by:

$$\text{ROC}(s_\sigma^*, \cdot) = \sum_{i=1}^{K+1} \text{ROC}^*(\alpha_i) \Phi_i^*(\cdot),$$

which may serve as a simple *approximant* of the optimal curve ROC^* , see Fig. 3.1.

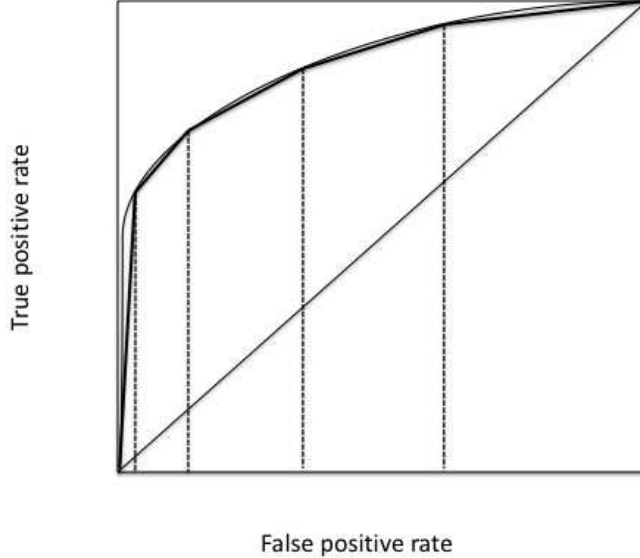


Figure 1: 2-spline approximant of the curve ROC^*

The next result, providing a bound for the corresponding approximation error, is well-known folklore in linear approximation theory.

Proposition 7 *Suppose that ROC^* is twice differentiable with bounded second derivative. In addition, set $\Delta = \max_{0 \leq i < K} \{\alpha_{i+1} - \alpha_i\}$. Then, we have:*

$$\|\text{ROC}(s_{\sigma}^*, \cdot) - \text{ROC}^*(\cdot)\|_{\infty} \leq -\frac{\Delta^2}{8} \inf_{\alpha \in [0,1]} \frac{d^2}{d\alpha^2} \text{ROC}^*(\alpha) .$$

Remark 2 (ON ADAPTIVE APPROXIMATION BY 2-SPLINES) Of course, considering approximation by piecewise linear functions with k pieces, the class of functions corresponding to the approximation order $O(k^{-1})$ in sup-norm is much larger than the collection of twice differentiable functions with bounded derivatives, see Chapter 12 in [DL93]. However, any practical procedure permitting to achieve this approximation rate under weaker hypotheses would require to choose the breakpoints α_k depending on the properties of the target curve ROC^* , not in advance anymore. In order to consider a more general setup, including cases where $\eta(X)$'s essential supremum is equal to 1 (*i.e.* $\lim_{\alpha \rightarrow 0} \text{ROC}^*(\alpha) = +\infty$), extensions of the premier approach developed in this article will be tackled in a future article, where the meshgrid is refined adaptively from the data. Incidentally, we point out that the spacings Δ_k between the breakpoints should be ideally chosen non decreasing given the geometry of the optimal ROC curve (concave and strictly increasing).

3.2 Empirical MV-set estimation

In the next section, we shall introduce a procedure for estimating the discrete scoring function s_{σ}^* for a given meshgrid σ . This method will be based on the statistical estimation

of the sets $\mathbf{R}_\alpha^* = \{\mathbf{x} \in \mathcal{X} \mid \eta(\mathbf{x}) > Q^*(\alpha)\}$ for specific choices of $\alpha \in (0, 1)$. This subproblem which is related to the design of statistical tests of *composite hypotheses* is interesting in itself. Applications include in particular anomaly/outlier detection, when the probability distribution corresponding to normal system activity is unknown or only partially known. More details will be provided in Section 5.

Interestingly, the level set \mathbf{R}_α^* can be interpreted as the solution of the constrained optimization problem:

$$\sup_{\mathbf{R} \in \mathcal{B}(\mathcal{X})} \mathbb{P}\{\mathbf{X} \in \mathbf{R} \mid \mathbf{Y} = +1\} \quad \text{subject to} \quad \mathbb{P}\{\mathbf{X} \in \mathbf{R} \mid \mathbf{Y} = -1\} \leq \alpha, \quad (3)$$

where the supremum is taken over the set $\mathcal{B}(\mathcal{X})$ of all measurable subsets of \mathcal{X} . This fact follows from Neyman-Pearson's lemma once the problem is cast as a hypothesis testing problem: test the null hypothesis $\mathcal{H}_0: \mathbf{Y} = -1$ against the alternative $\mathcal{H}_1: \mathbf{Y} = +1$ with a type I error equal to α and maximal power.

Note that this formulation is equivalent to the Minimum Volume (MV) set estimation framework (see [SN06] and references therein), since the complement set $\mathbf{S}_\alpha^* = \mathcal{X} \setminus \mathbf{R}_\alpha^*$ may be seen as the solution of:

$$\min_{\mathbf{S} \in \mathcal{B}(\mathcal{X})} G(\mathbf{S}) \quad \text{subject to} \quad H(\mathbf{S}) > 1 - \alpha,$$

the distribution G of positive instances being the *volume* to be minimized, while the distribution H of negative instances corresponds to the *reference measure*.

In our case, the major difference with the usual setting lies in the fact that the measure of reference H involved in the mass constraint is unknown, like G , and must be estimated from sampling data. A statistical search strategy based on the training sample \mathcal{D}_n could naturally consist of replacing the unknown probability distributions G and H by their empirical counterparts

$$\widehat{G}_n = \frac{1}{n_+} \sum_{i=1}^n \mathbb{I}\{\mathbf{Y}_i = +1\} \cdot \delta_{\mathbf{X}_i} \quad \text{and} \quad \widehat{H}_n = \frac{1}{n_-} \sum_{i=1}^n \mathbb{I}\{\mathbf{Y}_i = -1\} \cdot \delta_{\mathbf{X}_i}$$

with $n_+ = \sum_{i=1}^n \mathbb{I}\{\mathbf{Y}_i = +1\} = n - n_-$.

Let \mathcal{R} a class of measurable subsets of \mathcal{X} . We consider the following optimization problem as the empirical version of the previous one:

$$\sup_{\mathbf{R} \in \mathcal{R}} \widehat{G}_n(\mathbf{R}) \quad \text{subject to} \quad \widehat{H}_n(\mathbf{R}) \leq \alpha + \phi,$$

where ϕ is a complexity penalty, serving as a *tolerance parameter*. The success of this program in recovering a set close to \mathbf{R}_α^* will depend on both choices of the class \mathcal{R} and the parameter ϕ which will be discussed in Section 5.1.

3.3 The RANKOVER algorithm

We now describe a very simple ranking procedure which builds an estimator of s_{σ}^* in (2). The RANKOVER algorithm has two steps: **Optimization** and **Monotonicity**. The crucial part is the **Optimization** step. At each iteration, the procedure calls a classification algorithm which extracts, from the class \mathcal{R} of sets, the empirical counterpart of a level set of the regression function which contains a certain proportion of best instances. The grid of proportion levels depends on the partition σ_K . More precisely, if we set $u_k = \mathbb{P}\{\eta(X) \geq Q^*(\alpha_k)\}$, the method will successively target the best $(100u_1)\%$ among all instances, then the best $(100u_2)\%$, *etc.* Note that the classification algorithm invoked here is nonstandard since an additional constraint on the classifiers is involved. In this paper (see Section 5), we shall explore three possible strategies to solve this constrained classification problem: (i) empirical MV-set, (ii) threshold rules, (iii) weighted classification error. Here we only focus on statistical aspects, the design of practical techniques for empirical MV-set estimation such as grid methods will be investigated in a forthcoming paper. The **Monotonicity** step aims at deriving an increasing sequence of sets. This is a desirable property for estimators of the increasing sequence of the true level sets of the regression function. Additionally, this construction facilitates the analysis provided in Section 4. The other parameters of the algorithm are the partition $\sigma_K : \alpha_0 = 0 < \alpha_1 < \dots < \alpha_K < \alpha_{K+1} = 1$, $K \geq 1$, and the tolerance parameter denoted by ϕ .

THE RANKOVER ALGORITHM

Input. Meshgrid σ_K , tolerance parameter ϕ , class \mathcal{R} of sets

1. **Optimization.** For $k = 1, \dots, K$, compute:

$$\widehat{R}_k = \arg \max_{R \in \mathcal{R}} \widehat{G}_n(R) \quad \text{subject to} \quad \widehat{H}_n(R) \leq \alpha_k + \phi.$$

2. **Monotonicity.** Build recursively the increasing sequence $(\widetilde{R}_k)_{k \geq 1}$ through:

$$\widetilde{R}_1 = \widehat{R}_1 \text{ and } \widetilde{R}_{k+1} = \widetilde{R}_k \cup \widehat{R}_{k+1}, \quad \text{for all } k \in \{1, \dots, K-1\}.$$

Output. The piecewise constant scoring function obtained by overlaying the indicator functions of the sets \widetilde{R}_k :

$$s_K(x) = \sum_{k=1}^K \mathbb{I}\{x \in \widetilde{R}_k\}.$$

Statistical performance of this procedure will be discussed in the sequel. For now, we

provide some comments on possible modifications or additional outputs.

Remark 3 (BOTTOM-UP VS. TOP-DOWN.) Another strategy for constructing an increasing sequence of subsets from the collection $(\widetilde{\mathcal{R}}_k)_{k \geq 1}$ could be to proceed in a top-down manner. Start with $\widetilde{\mathcal{R}}_{K+1} = \mathcal{X}$ and $\widetilde{\mathcal{R}}_k = \widetilde{\mathcal{R}}_{k+1} \cap \widehat{\mathcal{R}}_k$ for $k = K, \dots, 1$. Similar results as those established in this paper could easily be derived for such a construction.

Remark 4 (PLUG-IN ESTIMATOR.) From Proposition 6, it turns out that a canonical scoring function would be $H^*(\eta(x))$. As a byproduct of the procedure, one may derive the following estimate of this function by reweighting the terms in the sum: $\sum_{k=1}^K (\alpha_k - \alpha_{k-1}) \mathbb{I}\{x \in \widetilde{\mathcal{R}}_k\}$.

Beyond the overlaid scoring function $s_K(x)$ resulting from the RANKOVER algorithm, additional outputs of the procedure are the estimates of the ROC curve and the AUC. Let $(\widetilde{\alpha}_k, \widetilde{\beta}_k) = (\widehat{H}_n(\widetilde{\mathcal{R}}_k), \widehat{G}_n(\widetilde{\mathcal{R}}_k))$ for all $k \in \{0, \dots, K+1\}$, where by convention $\widetilde{\mathcal{R}}_0 = \emptyset$ and $\widetilde{\mathcal{R}}_{K+1} = \mathcal{X}$. We point out that the empirical ROC curve of the scoring function output by the RANKOVER algorithm is the piecewise linear function:

$$\forall \alpha \in [0, 1], \quad \widetilde{\text{ROC}}(s_K, \alpha) = \sum_{k=1}^{K+1} \widetilde{\beta}_k \cdot \widetilde{\Phi}_k(\alpha),$$

where $\widetilde{\Phi}_k = \Phi(\alpha, \widetilde{\alpha}_{k-1}, \widetilde{\alpha}_k) - \Phi(\alpha, \widetilde{\alpha}_k, \widetilde{\alpha}_{k+1})$ for all $k \in \{1, \dots, K\}$ and $\widetilde{\Phi}_{K+1}(\alpha) = \Phi(\alpha, \widetilde{\alpha}_K, 1)$. Moreover, it follows from this expression that the corresponding empirical AUC is given by:

$$\widetilde{\text{AUC}}(s_K) = \frac{1}{2} \sum_{k=1}^K (\widetilde{\alpha}_{k+1} - \widetilde{\alpha}_{k-1}) \widetilde{\beta}_k.$$

3.4 Algorithmic approaches to scoring

Until recently the one and only known method for ranking/scoring binary-valued data was *logistic regression* or its numerous variants. Problems with high-dimensional data as those generated by the development of internet technologies naturally oriented the research of efficient ranking algorithms towards machine learning techniques. A nice illustration is the RankBoost algorithm (see [FISS03]) exporting the boosting approach of combination of weak learners to the problem of bipartite ranking. In a series of papers, we have developed our view of the ranking/scoring problem and proposed various approaches leading to or inspiring practical algorithms for doing the job:

- **W-ranking functionals.** This approach describes M-estimation strategies based on linear rank functionals (see [CV07] and [CV08b]). Indeed, many empirical summaries of the ROC curve such as the AUC ([HM82]), the local AUC ([CV07]), the p-norm push ([Rud06]), can be expressed as linear rank statistics. These statistics are

to be maximized over the functional class of scoring functions and the theoretical properties of this empirical risk maximization strategies require the control of a new class of stochastic processes, called rank processes (see [CV08b] for preliminary results).

- **Partitioning methods.** We have developed various partitioning methods for bipartite ranking. In [CV08c], we consider fixed partitions and histogram scoring rules for bipartite ranking. We also studied adaptive partitions based on decision trees in order to monitor the ranking performance in terms of the ROC curve (see [CV08a], [CV08d]).

Instead of partitioning the input space, the approach taken in this paper consists of taking a partition of the ROC space to build a finite-dimensional approximation of the optimal ROC curve. As illustrated above, the ranking problem reduces then to a collection of classification problems with an additional constraint. Solving each of these classification problems and then combining/overlaying their solutions through the RANKOVER algorithm leads to a scoring rule with good statistical performance (see Section 4). The main question for practitioners would be how to implement the **Optimization** step. For some clues on practical strategies devoted to this problem, we refer to Section 7 of [SN06] and the references therein.

4 Main results

4.1 Statistical properties of the RANKOVER algorithm

The next result offers a rate bound for the scoring function output by the RANKOVER algorithm in the ROC space, equipped with a sup-norm. Up to our knowledge, this is the first result on the generalization ability of decision rules in such a functional space. Given a class \mathcal{R} of sets in \mathcal{X} , we introduce the Rademacher average:

$$A_n = \mathbb{E} \left(\sup_{R \in \mathcal{R}} \frac{1}{n} \left| \sum_{i=1}^n \epsilon_i \mathbb{I}\{X_i \in R\} \right| \right)$$

where $(\epsilon_i)_{i \geq 1}$ forms an i.i.d. sequence which is independent of $(X_i)_{i \geq 1}$.

Theorem 8 *We consider a class \mathcal{R} of sets and we assume the following:*

- *the class \mathcal{R} of sets contains R_α^* for all $\alpha \in (0, 1)$,*
- *the Rademacher average A_n is of the order of $O(n^{-1/2})$,*
- *both G^* and H^* are twice continuously differentiable and have strictly positive first derivatives,*
- *the function ROC^* has a bounded second derivative.*

For all $k \in \{0, \dots, K + 1\}$, set $\alpha_k = k/(K + 1)$ and set the tolerance parameter to be:

$$\phi = \phi(\delta, n) = 2A_n + \sqrt{\frac{2 \log(1/\delta)}{n}} .$$

Denote by s_K the output of the RANKOVER algorithm with these parameters. If $K = K_n \sim n^{1/6}$ as $n \rightarrow \infty$, then there exists a constant $c = c(\delta)$ such that, with probability at least $1 - \delta$, we have, for n large enough:

$$\|\text{ROC}^*(.) - \text{ROC}(s_{K_n}, .)\|_\infty \leq cn^{-1/3} .$$

Remark 5 (CHOICE OF THE PENALTY.) The issue of penalty calibration has been a topic of intensive research in the last years (see [BBL05] and references therein). We do not enter in the subtleties related to this important question and we have chosen to use Rademacher averages as a complexity measure which covers most of the important examples of classes \mathcal{R} of sets (see [BBL05]).

Remark 6 (ON THE COMPLEXITY ASSUMPTION) The assumption on the Rademacher average A_n being of the order of $n^{-1/2}$ is fulfilled for instance if \mathcal{R} is a VC class. In that case, the constant c also depends on the VC dimension.

Remark 7 (OPTIMALITY ISSUE.) The rate of convergence in the theorem is not proved to be optimal. The investigation of lower bounds for this problem is the object of work under progress.

Remark 8 (PERFORMANCE OF CLASSIFIERS AND ROC CURVES.) In the present paper, we have adopted a scoring approach to ROC analysis which is somehow related to the evaluation of the performance of classifiers in ROC space. Using combinations of such classifiers to improve performance in terms of ROC curves has also been pointed out in [BDH06] and [BCT07].

Remark 9 (ADAPTIVITY OF THE PARTITION.) A natural extension of the approach would be to consider a flexible meshgrid $\{\alpha_k\}$ depending on the local smoothness of the optimal ROC curve. However, under the present assumptions, using an adaptive partition of $[0, 1]$ may yield sharper constants but will not improve the rate of convergence. We have investigated adaptive partitions of the interval $[0, 1]$ corresponding to *tree-structured* recursive approximation schemes of the optimal ROC curve elsewhere ([CV08d]), but the rates of convergence obtained in the present paper are faster.

4.2 Statistical estimation of the optimal ROC curve

We now show how to exploit the output of the **Optimization** step of the RANKOVER procedure in order to produce an accurate empirical estimate of the optimal ROC curve. We introduce some notations. Set: $\forall k \in \{0, \dots, K + 1\}$, $\hat{\alpha}_k = \hat{H}_n(\hat{R}_k)$ and $\hat{\beta}_k = \hat{G}(\hat{R}_k)$.

The broken line that connects the knots $\{(\hat{\alpha}_k, \hat{\beta}_k); 0 \leq k \leq K + 1\}$ provides an empirical counterpart of the piecewise linear approximant of the optimal ROC^* . We also introduce the "hat functions" defined by:

$$\forall k \in \{1, \dots, K - 1\}, \hat{\Phi}_k(\cdot) = \Phi(\cdot; (\hat{\alpha}_{k-1}, \hat{\alpha}_k)) - \Phi(\cdot; (\hat{\alpha}_k, \hat{\alpha}_{k+1})).$$

We also set $\hat{\Phi}_K(\cdot) = \Phi(\cdot; (\hat{\alpha}_K, 1))$ for notational convenience. The statistical estimate may be then written as:

$$\widehat{\text{ROC}}^*(\alpha) = \sum_{k=1}^K \hat{\beta}_k \hat{\Phi}_k(\alpha).$$

The next result takes the form of a deviation bound for the estimation of the optimal ROC curve. It quantifies the order of magnitude of a confidence band in supremum norm around an empirical estimate based on a statistical version of a simple finite elements method (FEM) approximation scheme.

Theorem 9 *Under the same assumptions as in Theorem 8, we set here $K = K_n \sim n^{1/4}$. Then, there exists a constant $c = c(\delta)$ such that, with probability at least $1 - \delta$:*

$$\|\widehat{\text{ROC}}^* - \text{ROC}^*\|_\infty \leq c \left(\frac{\log n}{n} \right)^{1/2}.$$

Remark 10 (A NONDECREASING ESTIMATE OF ROC^* .) Notice that the curve $\widehat{\text{ROC}}^*(\cdot)$ is not necessarily increasing, in contrast to the empirical ROC curve $\widetilde{\text{ROC}}^*$. However, the accuracy of the latter estimate is worst (only of order $O_{\mathbb{P}}((\log(n)/n)^{1/3})$) with the best possible choice of $K = K_n \sim n^{1/6}$.

5 On learning a statistical test of composite hypotheses

In this section, we focus on the statistical study of the subprocedure called the **Optimization** step in the RANKOVER algorithm. Recall that the purpose of this step is to estimate the sets $\mathcal{R}_{\alpha_k}^*$ through solving the following problem:

$$\sup_{\mathcal{R} \in \mathcal{R}} \widehat{G}_n(\mathcal{R}) \text{ subject to } \widehat{H}_n(\mathcal{R}) \leq \alpha + \phi.$$

In other words, the goal is to select a critical region $\mathcal{R} \in \mathcal{R}$ in order to construct a decision rule based on the observation X , *i.e.* a classifier $C(X) = 2 \cdot \mathbb{I}\{X \in \mathcal{R}\} - 1$, for testing the null hypothesis $\mathcal{H}_0 : Y = -1$ with type I error α and maximum power. As the distribution of the observation X is unknown under both hypotheses, this may be interpreted as the problem of learning an optimal statistical test of composite hypotheses. Even though this only corresponds to a step towards reaching the overall goal considered in this paper, this problem is interesting in itself. Our main approach in this section will follow the work of

Scott and Nowak ([SN06]) on learning minimum volume sets, but we also refer to [SN05], [Sco05]. We extend their result to the case where the reference measure is unknown and provide fast rates of convergence of MV-set estimators. In the end of the section, we also describe alternative methods to the MV-set approach and describe their statistical properties.

5.1 Rate bounds for empirical MV-set estimation

We denote by $\widehat{\mathbf{R}}_\alpha$ the solution to this problem. The next result can be interpreted as a rate bound, in terms of type II error, for the excess risk of the classifier defined by $\widehat{\mathbf{R}}_\alpha$ with a simultaneous control of the type I error. A similar result was also obtained in [SBW08].

The main assumptions for consistency results to hold concerns the complexity of the collection \mathcal{R} of candidate sets, as well as its capacity to represent the target set \mathbf{R}_α^* . For simplicity, we have chosen to describe the complexity in terms of the Rademacher average and we have also assumed that the class \mathcal{R} contains the optimal element.

Theorem 10 *Let $\alpha \in (0, 1)$. Assume that \mathbf{R}_α^* belongs to the set \mathcal{R} of region candidates. Suppose in addition that \mathcal{R} forms a class of subsets of \mathcal{X} with Rademacher average denoted by A_n . For all $(\delta, n) \in (0, 1) \times \mathbb{N}^*$, set*

$$\phi(\delta, n) = 2A_n + \sqrt{\frac{2 \log(1/\delta)}{n}}.$$

Then, for all $\delta > 0$, we simultaneously have with probability at least $1 - \delta$: $\forall n \geq 1$,

$$H(\widehat{\mathbf{R}}_\alpha) \leq \alpha + 2\phi(\delta/2, n) \text{ and } G(\widehat{\mathbf{R}}_\alpha) \geq G(\mathbf{R}_\alpha^*) - 2\phi(\delta/2, n).$$

Remark 11 (ON RECOVERING A POINT ON THE OPTIMAL ROC CURVE) When the cdf H^* (respectively, G^*) is continuous at $Q^*(\alpha)$, the point $(H(\mathbf{R}_\alpha^*), G(\mathbf{R}_\alpha^*))$ naturally coincides with the point on $(\alpha, \text{ROC}^*(\alpha))$ of the optimal curve. As may be shown by examining Theorem 10's proof, the euclidean distance in the ROC space of the point $(\widehat{H}_n(\widehat{\mathbf{R}}_\alpha), \widehat{G}_n(\widehat{\mathbf{R}}_\alpha))$ determined by solving the constrained ERM problem (3) to $(\alpha, \text{ROC}^*(\alpha))$ is then of order $O_{\mathbb{P}}(1/\sqrt{n})$. This will be exploited later when constructing an estimate of the curve ROC^* with a controlled approximation error.

5.2 Fast (but not so fast) rates of convergence

We now exhibit assumptions under which faster rates of convergence can be attained. In [Tsy04], conditions leading to rate bounds faster than $n^{-1/2}$ have been examined in the binary classification setting. It is the purpose of this subsection to adapt the latter to the hypothesis testing setup.

Noise assumption (NA). There exist constants $\alpha \in (0, 1)$ and $D > 0$ such that: $\forall t \geq 0$,

$$\mathbb{P}\{|\eta(X) - Q^*(\alpha)| \leq t\} \leq D \cdot t^{\frac{\alpha}{1-\alpha}}.$$

We point out that this assumption corresponds to the one introduced in [Tsy04], except that here the quantile $Q^*(\alpha)$ replaces $1/2$.

Remark 12 (ON THE NOISE ASSUMPTION) It is noteworthy that as soon as $\eta(X)$'s distribution, namely $F^* = pG^* + (1-p)H^*$, has a bounded density f^* , this hypothesis is automatically fulfilled with $\alpha = 1/2$ and $D = \sup_t f^*(t)$. Indeed, the finite increments theorem yields:

$$\begin{aligned} \mathbb{P}\{|\eta(X) - Q^*(\alpha)| \leq t\} &= F^*(Q^*(\alpha) + t) - F^*(Q^*(\alpha) - t) \\ &\leq 2Dt. \end{aligned}$$

The next result describes an important consequence of this condition.

Lemma 11 (VARIANCE CONTROL) *Suppose that condition (NA) is fulfilled. Set for all $R \in \mathcal{R}$,*

$$s_\alpha^2(R) \stackrel{\text{def}}{=} \text{var}(\mathbb{I}\{Y = +1\}(\mathbb{I}\{X \in R_\alpha^*\} - \mathbb{I}\{X \in R\}))$$

Then, we have:

$$\forall r \in \mathcal{R}, \quad s_\alpha^2(R) \leq c(p(1 - Q^*(\alpha))(G(R_\alpha^*) - G(R)) + Q^*(\alpha)(1 - p)(H(R) - \alpha))^\alpha.$$

Theorem 12 (FAST RATES) *Assume that the assumptions of Theorem 10 are fulfilled. Suppose that, additionally, $\eta(X)$ has a bounded density and that $A_n = O(n^{-1/2})$. Then, for all $\delta > 0$, we simultaneously have with probability at least $1 - \delta$: $\exists C = C(\mathcal{R}, \delta, \alpha, p), n_0$, $\forall n \geq n_0$,*

$$H(\widehat{R}_\alpha) \leq \alpha + 2\phi(\delta/2, n) \text{ and } \text{ROC}^*(\alpha) - G(\widehat{R}_\alpha) \leq Cn^{-5/8}.$$

Remark 13 (MV-SET ESTIMATION WITH KNOWN REFERENCE MEASURE) We point out that it follows from the proof of Theorem 12 that, in the case where the reference measure is known, condition (NA) ensures that, when performing empirical risk minimization over the set $\{R \in \mathcal{R} : H(R) \leq \alpha\}$, the rate of the excess of risk (in terms of type II error) is of the order of $O(n^{-1/(2-\alpha)})$. Here, there is no guarantee that the H-term in the variance control bound can be either negative or neglected. Thus, we obtain a not so fast rate of the order of $n^{-5/8}$ instead of the expected $n^{-2/3}$ with $\alpha = 1/2$.

5.3 Alternative methods for solving the ERM under constraints

Here we consider alternatives to the empirical MV-set estimation method for solving the constrained classification problem. The first example consists of threshold rules which were introduced in [CV07]. We also consider an empirical risk minimization method based on a weighted classification error. The latter method is not a true competitor of the others in the sense that it does not lead to an estimator of the target level set \mathbf{R}_α^* . However, we present it for completeness as it could inspire a similar overlaying scheme from a finite collection of level sets of the regression function.

Threshold rules. In order to guarantee the constraint to be satisfied, we could consider sets of the form $\mathbf{R}_\alpha(s) = \{x \in \mathcal{X} : s(x) \geq Q(s(X), \alpha)\}$, where s belongs to a collection \mathcal{S} of scoring functions. However, since the distribution H is unknown, the quantile $Q(s, \alpha)$ has to be replaced by its empirical counterpart $\hat{Q}_n(s, \alpha) = \hat{H}_n^{-1}(1 - \alpha)$, which leads to consider actually the set

$$\hat{\mathbf{R}}_\alpha(s) \stackrel{\text{def}}{=} \left\{ x \in \mathcal{X} : s(x) \geq \hat{Q}_n(s, \alpha) \right\}.$$

The next result shows that, under basic complexity assumptions, the type I errors are uniformly controlled over $s \in \mathcal{S}$. We introduce a different penalty based on Vapnik-Chervonenkis (VC) type characterization, for $\delta > 0$ and $n \geq 1$:

$$\tilde{\phi}(\delta, n) = 2\sqrt{\frac{2V \log(n+1)}{n}} + \sqrt{\frac{2 \log(1/\delta)}{n}}$$

where V is the VC dimension of the underlying functional class.

Lemma 13 (TYPE I ERROR - UNIFORM BOUND) *Suppose that \mathcal{S} is a major VC class of functions with finite VC dimension V . Then, for all $\delta \in (0, 1)$, we have with probability at least $1 - \delta$,*

$$\sup_{s \in \mathcal{S}} H(\hat{\mathbf{R}}_\alpha(s)) \leq \alpha + \tilde{\phi}(\delta, n).$$

Remark 14 (ON THE COMPLEXITY ASSUMPTION) For further details on the terminology of major sets and major classes one may refer to [Dud99]. This notions determine the combinatorial complexity of sets of the form $\{x \in \mathcal{X} : s(x) \leq t\}$ or $\{x \in \mathcal{X} : s(x) \geq t\}$. The complexity assumption involved in Lemma 13 ensures that the collection of sets and indexed by $(s, t) \in \mathcal{S} \times \mathbb{R}$ form a VC class of sets.

Let us investigate the performance of the test with maximum power, which corresponds to the test function:

$$\hat{s}_n = \arg \max_{s \in \mathcal{S}} \hat{G}_n(\mathbf{R}_\alpha(s)).$$

Theorem 14 *Suppose that $\mathcal{S} \cap \mathcal{S}^* \neq \emptyset$. Under the assumptions of Lemma 13, for all $\delta \in (0, 1)$, we have with probability at least $1 - \delta$:*

$$H(\widehat{R}_\alpha(\widehat{s}_n)) \leq \alpha + 2\widetilde{\phi}(\delta/2, n) \text{ and } G(\widehat{R}_\alpha(\widehat{s}_n)) \geq \text{ROC}^*(\alpha) - 2\widetilde{\phi}(\delta/2, n)$$

As it immediately follows from Lemma 13 combined with the proof argument of Theorem 10, the proof is omitted.

Remark 15 (ON FAST RATES) We also point out that this result may be viewed as a variant of Theorem 5 in [CV07], related to the so-termed *classification problem with mass-constraint*. The difference with the present setting lies in the fact that the 'volume' to be minimized is a signed measure up to an additive constant, namely the classification error $\mathbb{P}\{Y \neq 2\mathbb{I}\{X \in R\} - 1\} = pH(R) + (1 - p)(1 - G(R))$, and the reference measure involved in the constraint is the marginal distribution μ . In addition, it is noteworthy that, under Theorem 12's conditions combined with the assumption that the cdf's H_s and G_s are both twice differentiable at $Q(s(X), \alpha)$ for all $s \in \mathcal{S}$, the rate $n^{-2/3}$ for the excess of type II error can be achieved, see Theorem 10 in [CV07].

Classification with asymmetric costs. For any measurable set $C \subset \mathcal{X}$, we define the *weighted classification error*:

$$L_\omega(C) = 2p(1 - \omega) (1 - G(C)) + 2(1 - p)\omega H(C) ,$$

with $\omega \in (0, 1)$ being the asymmetry factor. For $\omega = 1/2$ one recovers the standard classification error $L(C) = \mathbb{P}\{C(X) \neq Y\}$. As shown by the next result, the minimizers of this collection of risk measures coincide with the η -level sets. The proof is left to the reader.

Proposition 15 *The optimal set for this error measure is $C_\omega^* = \{x : \eta(x) > \omega\}$. We have indeed, for all $C \subset \mathcal{X}$:*

$$L_\omega(C_\omega^*) \leq L_\omega(C) .$$

Also the optimal error is given by:

$$L_\omega(C_\omega^*) = 2\mathbb{E} \min\{\omega(1 - \eta(X)), (1 - \omega)\eta(X)\} .$$

The excess risk for an arbitrary set C can be written:

$$L_\omega(C) - L_\omega(C_\omega^*) = 2\mathbb{E} (|\eta(X) - \omega| \mathbb{I}\{X \in C \Delta C_\omega^*\}) ,$$

where Δ stands for the symmetric difference between sets.

The empirical counterpart of the weighted classification error can be defined as:

$$\hat{L}_\omega(C) = \frac{2\omega}{n} \sum_{i=1}^n \mathbb{I}\{Y_i = -1, X_i \in C\} + \frac{2(1-\omega)}{n} \sum_{i=1}^n \mathbb{I}\{Y_i = +1, X_i \notin C\}.$$

This leads to consider the *weighted empirical risk minimizer* over a class \mathcal{R} of candidate sets:

$$\hat{C}_\omega = \arg \min_{C \in \mathcal{R}} \hat{L}_\omega(C).$$

The next result provides rates of convergence of the weighted empirical risk minimizer \hat{C}_ω to the best set in the class in terms of the two types of error.

Theorem 16 *Let $\omega \in (0, 1)$. Assume that \mathcal{R} is of finite VC dimension V and contains C_ω^* . Suppose also that both G^* and H^* are twice continuously differentiable with strictly positive first derivatives and that ROC^* has a bounded second derivative. Then, for all $\delta > 0$, there exist constants $c(V)$ independent of ω such that, with probability at least $1 - \delta$:*

$$|\mathbb{H}(\hat{C}_\omega) - \mathbb{H}(C_\omega^*)| \leq \frac{c(V)}{\sqrt{p(1-\omega)}} \cdot \left(\frac{\log(1/\delta)}{n} \right)^{\frac{1}{3}}.$$

The same result also holds for the excess risk of \hat{C}_ω in terms of the true positive rate with a factor term of $\sqrt{(1-p)\omega}$ in the denominator instead.

It is noteworthy that, while convergence in terms of classification error is expected to be of the order of $n^{-1/2}$, its two components corresponding to the rate of false positive and true positive present slower rates. Hence, even though through usual classification methods can readily be used for recovering a collection of η -level sets, the empirical MV set approach should be preferred regarding the rate of convergence.

6 Conclusion

In the paper, we proposed a ranking/scoring algorithm based on the resolution of a collection of constrained classification problems. Statistical performance in terms of the convergence towards the optimal ROC curve in supremum norm is studied. We also consider various strategies for solving the constrained classification problem: empirical MV-set approach, threshold rules, weighted empirical risk minimization. Several issues remain open: optimality of convergence rate bounds, adaptive grid for approximation, practical implementations of empirical MV-set estimation. Their investigation is undertaken through ongoing projects.

Appendix - Proof section

Proof of Proposition 5

First, we observe that, for any measurable function h , we have, by a change of probability, that:

$$\mathbb{E}(h(X) \mid Y = +1) = \frac{1-p}{p} \mathbb{E} \left(\frac{\eta(X)}{1-\eta(X)} h(X) \mid Y = -1 \right) .$$

We apply this to $h(X) = \mathbb{I}\{X \in \mathcal{R}_{\alpha}^*\} - \mathbb{I}\{X \in \mathcal{R}_{s,\alpha}\}$ in order to get:

$$\text{ROC}^*(\alpha) - \text{ROC}(s, \alpha) = \frac{1-p}{p} \mathbb{E} \left(\frac{\eta(X)}{1-\eta(X)} h(X) \mid Y = -1 \right) .$$

Then we add and subtract $Q^*(\alpha)/(1-Q^*(\alpha))$ and using the fact that

$$\alpha = \mathbb{P}\{X \in \mathcal{R}_{s,\alpha} \mid Y = -1\} = \mathbb{P}\{X \in \mathcal{R}_{\alpha}^* \mid Y = -1\} ,$$

we get:

$$\text{ROC}^*(\alpha) - \text{ROC}(s, \alpha) = \left(\frac{1-p}{p} \right) \mathbb{E} \left(\left(\frac{\eta(X)}{1-\eta(X)} - \frac{Q^*(\alpha)}{1-Q^*(\alpha)} \right) h(X) \mid Y = -1 \right) .$$

We remove the conditioning with respect to $Y = -1$ and using then conditioning on X , we obtain:

$$\text{ROC}^*(\alpha) - \text{ROC}(s, \alpha) = \frac{1}{p} \mathbb{E} \left(\left(\frac{\eta(X) - Q^*(\alpha)}{1-Q^*(\alpha)} \right) h(X) \right) . \blacksquare$$

Proof of Proposition 6

We recall (see [CLV08]) that:

$$\text{AUC}^* - \text{AUC}(\hat{\eta}) = \frac{\mathbb{E} (|\eta(X) - \eta(X')| \mathbb{I}\{(X, X') \in \Gamma\})}{2p(1-p)} .$$

where

$$\Gamma = \{(x, x') : \text{sgn}(\hat{\eta}(x) - \hat{\eta}(x')) \neq \text{sgn}(\eta(x) - \eta(x'))\}$$

But, one may easily check that:

if $\text{sgn}(\hat{\eta}(X) - \hat{\eta}(X')) \neq \text{sgn}(\eta(X) - \eta(X'))$, then

$$|\eta(X) - \eta(X')| \leq |\eta(X) - \hat{\eta}(X)| + |\eta(X') - \hat{\eta}(X')| ,$$

which gives the first part of the result.

Turning to the second assertion, consider the event

$$\mathcal{E} = \{X \in \mathcal{R}_{\alpha}^* \Delta \mathcal{R}_{\hat{\eta}, \alpha}\} .$$

Notice first that, after Proposition 5, we have:

$$\begin{aligned} \text{ROC}^*(\alpha) - \text{ROC}(\hat{\eta}, \alpha) &= \frac{\mathbb{E}(|\eta(X) - Q^*(\alpha)| \mathbb{I}_{\mathcal{E}})}{p(1 - Q^*(\alpha))} \\ &\leq \frac{c\mathbb{E}(|H^*(\eta(X)) - 1 + \alpha| \mathbb{I}_{\mathcal{E}})}{p(1 - Q^*(\alpha))} \end{aligned}$$

by virtue of the finite increments theorem. Now, observing that

$$\mathcal{E} = \{\text{sgn}(H^*(\eta(X)) - 1 + \alpha) \neq \text{sgn}(H_{\hat{\eta}}(\hat{\eta}(X)) - 1 + \alpha)\},$$

we have in a similar fashion as above: if $X \in \mathbb{R}_{\alpha}^* \Delta \mathbb{R}_{\hat{\eta}, \alpha}$, then

$$|H^*(\eta(X)) - 1 + \alpha| \leq |H^*(\eta(X)) - H_{\hat{\eta}}(\hat{\eta}(X))|,$$

which, combined to the previous bound, proves the second part. \blacksquare

Proof of Theorem 8

We note $\tilde{\alpha}_i = H(\tilde{\mathbb{R}}_i)$, $\tilde{\beta}_i = G(\tilde{\mathbb{R}}_i)$ and also $\tilde{\Phi}_i(\cdot) = \Phi(\cdot; (\tilde{\alpha}_{i-1}, \tilde{\alpha}_i)) - \Phi(\cdot; (\tilde{\alpha}_i, \tilde{\alpha}_{i+1}))$. We then have

$$\text{ROC}(\tilde{s}_{\sigma_K}, \alpha) = \sum_{i=1}^K \tilde{\beta}_i \tilde{\Phi}_i(\alpha)$$

and we can use the following decomposition, for any $\alpha \in [0, 1]$:

$$\text{ROC}^*(\alpha) - \text{ROC}(\tilde{s}_{\sigma_K}, \alpha) = \left(\text{ROC}^*(\alpha) - \sum_{i=1}^K \text{ROC}^*(\tilde{\alpha}_i) \tilde{\Phi}_i(\alpha) \right) + \sum_{i=1}^K (\text{ROC}^*(\tilde{\alpha}_i) - \tilde{\beta}_i) \tilde{\Phi}_i(\alpha).$$

From Proposition 7 we can bound the first term (which is positive), $\forall \alpha \in [0, 1]$, by:

$$-\frac{1}{8} \inf_{\alpha \in [0, 1]} \frac{d^2}{d\alpha^2} \text{ROC}^*(\alpha) \cdot \max_{0 \leq i \leq K} (\tilde{\alpha}_{i+1} - \tilde{\alpha}_i)^2.$$

Now, to control the second term, we upper bound the following quantity:

$$|\text{ROC}^*(\tilde{\alpha}_i) - \tilde{\beta}_i| \leq \sup_{\alpha \in [0, 1]} \frac{d}{d\alpha} \text{ROC}^*(\alpha) \cdot |\tilde{\alpha}_i - \alpha_i| + |\text{ROC}^*(\alpha_i) - \tilde{\beta}_i|$$

We further bound: $|\tilde{\alpha}_i - \alpha_i| \leq |\tilde{\alpha}_i - \bar{\alpha}_i| + |\bar{\alpha}_i - \alpha_i|$ where $\bar{\alpha}_i = H(\hat{\mathbb{R}}_{\alpha_i})$. In order to deal with the first term, the next lemma will be needed:

Lemma 17 *We have, for all $k \in \{1, \dots, K\}$:*

$$H(\tilde{\mathbb{R}}_k) = H(\hat{\mathbb{R}}_{\alpha_k}) + (k-1)O_{\mathbb{P}}(\phi(\delta, n)).$$

where the notation $O_{\mathbb{P}}(1)$ is used for a r.v. which is bounded in probability.

From the lemma, it follows that: $\max_{1 \leq i \leq K} |\tilde{\alpha}_i - \bar{\alpha}_i| = O_{\mathbb{P}}(K\phi(\delta, \mathbf{n}))$. We can then use Theorem 10 with δ replaced by δ/K to get that $\max_{1 \leq i \leq K} |\bar{\alpha}_i - \alpha_i| = O_{\mathbb{P}}(\phi(\delta/K, \mathbf{n}))$. The same inequalities hold with the β 's. It remains to control the quantity $\tilde{\alpha}_{i+1} - \tilde{\alpha}_i$. We have:

$$|\tilde{\alpha}_{i+1} - \tilde{\alpha}_i| \leq \max_{1 \leq k \leq K} |H(\hat{R}_{\alpha_k}) - H(\hat{R}_{\alpha_{k-1}})| + K O_{\mathbb{P}}(\phi(\delta, \mathbf{n})) .$$

We have that:

$$\max_{1 \leq k \leq K} |H(\hat{R}_{\alpha_k}) - H(\hat{R}_{\alpha_{k-1}})| \leq 2 \max_{1 \leq k \leq K} |H(\hat{R}_{\alpha_k}) - \alpha_k| + \max_{1 \leq k \leq K} |\alpha_k - \alpha_{k-1}| .$$

As before, we have that the first term is of the order $\phi(\delta/K, \mathbf{n})$ and since the second derivative of the optimal ROC curve is bounded, the second term is of the order K^{-1} . Eventually, we choose K in order to optimize the quantity: $K^2 A_n^2 + K^2 n^{-1} + K^{-2} + A_n^2 + \log K/n + K A_n + K n^{-1/2} + A_n + (\log K/n)^{1/2}$. Using the assumption on the rate of A_n , the optimization in K leads to the choice of $K = K_n \sim n^{1/6}$.

Proof of Lemma 17.

We have that $H(\tilde{R}_2) = H(\hat{R}_{\alpha_2}) + H(\hat{R}_{\alpha_1} \setminus \hat{R}_{\alpha_2})$. Therefore, since $R_{\alpha_1}^* \subset R_{\alpha_2}^*$ and observing that

$$H(\hat{R}_{\alpha_1} \setminus \hat{R}_{\alpha_2}) = H(((\hat{R}_{\alpha_1} \setminus R_{\alpha_1}^*) \cup (\hat{R}_{\alpha_1} \cap R_{\alpha_1}^*)) \setminus ((\hat{R}_{\alpha_2} \setminus R_{\alpha_2}^*) \cup (\hat{R}_{\alpha_2} \cap R_{\alpha_2}^*))) ,$$

it suffices to use the additivity of the probability measure $H(\cdot)$ to get: $H(\tilde{R}_2) = H(\hat{R}_{\alpha_2}) + O_{\mathbb{P}}(\phi(\delta, \mathbf{n}))$. Eventually, errors are stacked and we obtain the result.

Proof of Theorem 9.

We use the following decomposition, for any fixed $\alpha \in (0, 1)$:

$$\widehat{ROC}^*(\alpha) - ROC^*(\alpha) = \left(\widehat{ROC}^*(\alpha) - \sum_{i=1}^K ROC^*(\hat{\alpha}_i) \hat{\Phi}_i(\alpha) \right) + \left(\sum_{i=1}^K ROC^*(\hat{\alpha}_i) \hat{\Phi}_i(\alpha) - ROC^*(\alpha) \right) .$$

Therefore, we have by a triangular inequality: $\forall \alpha \in [0, 1]$,

$$\left| \widehat{ROC}^*(\alpha) - \sum_{i=1}^K ROC^*(\hat{\alpha}_i) \hat{\Phi}_i(\alpha) \right| \leq \max_{1 \leq i \leq K} |\hat{\beta}_i - \bar{\beta}_i| + |\bar{\beta}_i - ROC^*(\alpha_i)| + |ROC^*(\alpha_i) - ROC^*(\hat{\alpha}_i)| ,$$

where $\bar{\beta}_i = G(\hat{R}_{\alpha_i})$ for $i \in \{1, \dots, K\}$. And, by the finite increments theorem, we have:

$$|ROC^*(\alpha_i) - ROC^*(\hat{\alpha}_i)| \leq \left(\sup_{\alpha \in [0, 1]} \frac{d}{d\alpha} ROC^*(\alpha) \right) (|\alpha_i - \bar{\alpha}_i| + |\bar{\alpha}_i - \hat{\alpha}_i|) .$$

For the other term, we use the same result on approximation as in the proof of Theorem 8:

$$\left| \sum_{i=1}^K \text{ROC}^*(\hat{\alpha}_i) \hat{\phi}_i(\alpha) - \text{ROC}^*(\alpha) \right| \leq \frac{1}{8} \inf_{\alpha \in [0,1]} \frac{d^2}{d\alpha^2} \text{ROC}^*(\alpha) \cdot \max_{0 \leq i \leq K} (\hat{\alpha}_{i+1} - \hat{\alpha}_i)^2$$

$$\max_{0 \leq i \leq K} (\hat{\alpha}_{i+1} - \hat{\alpha}_i) \leq \max_{0 \leq i \leq K} (\alpha_{i+1} - \alpha_i) + 2 \max_{1 \leq i \leq K} |\alpha_i - \bar{\alpha}_i| + 2 \max_{1 \leq i \leq K} |\hat{\alpha}_i - \bar{\alpha}_i| .$$

We recall that: $\max_{1 \leq i \leq K} |\hat{\alpha}_i - \bar{\alpha}_i| = O_{\mathbb{P}}(K n^{-1/2})$. Moreover, $\max_{0 \leq i \leq K} \{\alpha_{i+1} - \alpha_i\}$ is of the order of K^{-1} . And with probability at least $1 - \delta$, we have that $\max_{1 \leq i \leq K} |\alpha_i - \bar{\alpha}_i|$ is bounded as in Theorem 10, except that δ is replaced by δ/K in the bound. Eventually, we get the generalization bound: $K^{-2} + (\log K/n)^{1/2}$, which is optimal for a number of knots: $K \sim n^{1/4}$.

Proof of Theorem 10

In order to prove the desired result, we introduce further notation, namely

$$\hat{\mathcal{R}}_{\alpha} = \left\{ \mathbf{R} \in \mathcal{R} : \hat{H}_n(\mathbf{R}) \leq \alpha + \phi(\delta/2, n) \right\} ,$$

so that one may write

$$\hat{\mathbf{R}}_{\alpha} = \arg \max_{\mathbf{R} \in \hat{\mathcal{R}}_{\alpha}} \hat{G}_n(\mathbf{R}) .$$

We shall consider the following events:

$$\Theta_H = \left\{ H(\hat{\mathbf{R}}_{\alpha}) > \alpha + 2\phi(\delta/2, n) \right\} \text{ and } \Theta_G = \left\{ G(\hat{\mathbf{R}}_{\alpha}) < G(\mathbf{R}_{\alpha}^*) - 2\phi(\delta/2, n) \right\} ,$$

as well as

$$\Omega_H = \left\{ \sup_{\mathbf{R} \in \mathcal{R}} |\hat{H}_n(\mathbf{R}) - H(\mathbf{R})| > \phi(\delta/2, n) \right\} \text{ and } \Omega_G = \left\{ \sup_{\mathbf{R} \in \mathcal{R}} |\hat{G}_n(\mathbf{R}) - G(\mathbf{R})| > \phi(\delta/2, n) \right\} .$$

The complementary event of any event E will be denoted by E^c . The matter is to establish a lower bound for the probability of occurrence of the complementary event of $\Theta_H \cup \Theta_G$. We shall prove that

$$\Theta_H \cup \Theta_G \subset \Omega_H \cup \Omega_G. \tag{4}$$

and the result will then follow from the union bound combined with McDiarmid's concentration inequality and the control of empirical process by a Rademacher average through a double symmetrization argument (see [BBL05] for details). We have, indeed, that, for all $\delta \in (0, 1)$, the event Ω_H (respectively, the event Ω_G) occurs with probability less than $\delta/2$.

Observe first that $\Omega_H^c \cap \Omega_G^c \subset \Theta_G^c$. As a matter of fact, on the event Ω_H^c we have

$$\widehat{H}_n(\mathbf{R}_\alpha^*) - \alpha \leq \sup_{\mathbf{R} \in \mathcal{R}} |\widehat{H}_n(\mathbf{R}) - H(\mathbf{R})| \leq \phi(\delta/2, \mathbf{n}),$$

so that we have $\mathbf{R}_\alpha^* \in \widehat{\mathcal{R}}_\alpha$ and thus, $\widehat{G}_n(\widehat{\mathbf{R}}_\alpha) \geq \widehat{G}_n(\mathbf{R}_\alpha^*)$. In addition, since

$$G(\widehat{\mathbf{R}}_\alpha) = (G(\widehat{\mathbf{R}}_\alpha) - \widehat{G}_n(\widehat{\mathbf{R}}_\alpha)) + (\widehat{G}_n(\widehat{\mathbf{R}}_\alpha) - \widehat{G}_n(\mathbf{R}_\alpha^*)) + (\widehat{G}_n(\mathbf{R}_\alpha^*) - G(\mathbf{R}_\alpha^*)) + G(\mathbf{R}_\alpha^*),$$

on the event of $\Omega_H^c \cap \Omega_G^c$ we have $G(\widehat{\mathbf{R}}_\alpha) \geq G(\mathbf{R}_\alpha^*) - 2\phi(\delta/2, \mathbf{n})$, and the latter event corresponds to Θ_G^c . Eventually, on the event Ω_H^c , we have

$$\begin{aligned} H(\mathbf{R}_\alpha^*) &\leq \widehat{H}_n(\mathbf{R}_\alpha^*) + \sup_{\mathbf{R} \in \mathcal{R}} |H(\mathbf{R}) - \widehat{H}_n(\mathbf{R})| \\ &\leq \alpha + 2\phi(\delta/2, \mathbf{n}), \end{aligned}$$

so that $\Omega_H^c \subset \Theta_H^c$. ■

Proof of Lemma 11

It is straightforward to extend the equivalent statements of the noise assumption (NA) in the standard classification setup (see [BBL05]) to an arbitrary level. We use the following equivalent condition: there exists a positive constant c such that, for any set \mathbf{R} , we have

$$\mathbb{E}(\mathbb{I}\{X \in \mathbf{R}_\alpha^* \Delta \mathbf{R}\}) \leq c(F(\mathbf{R}_\alpha^*) - F(\mathbf{R}))^\alpha,$$

where $F = pG + (1-p)H$. From there, we can deduce the next bound:

$$s_\alpha^2(\mathbf{R}) \leq c(\mathbb{E}(|\eta(X) - Q^*(\alpha)| \cdot \mathbb{I}\{X \in \mathbf{R}_\alpha^* \Delta \mathbf{R}\}))^\alpha,$$

by Hölder's inequality.

Observe also that

$$\begin{aligned} p(G(\mathbf{R}_\alpha^*) - G(\mathbf{R})) &= \mathbb{E}[(\eta(X) - Q^*(\alpha)) \cdot (\mathbb{I}\{X \in \mathbf{R}_\alpha^*\} - \mathbb{I}\{X \in \mathbf{R}\})] \\ &\quad + Q^*(\alpha)(\mathbb{P}\{X \in \mathbf{R}_\alpha^*\} - \mathbb{P}\{X \in \mathbf{R}\}), \end{aligned}$$

and

$$\begin{aligned} (1-p)(H(\mathbf{R}_\alpha^*) - H(\mathbf{R})) &= \mathbb{E}[(Q^*(\alpha) - \eta(X)) \cdot (\mathbb{I}\{X \in \mathbf{R}_\alpha^*\} - \mathbb{I}\{X \in \mathbf{R}\})] \\ &\quad + (1 - Q^*(\alpha))(\mathbb{P}\{X \in \mathbf{R}_\alpha^*\} - \mathbb{P}\{X \in \mathbf{R}\}). \end{aligned}$$

This yields

$$\mathbb{E}[|\eta(X) - Q^*(\alpha)| \cdot \mathbb{I}\{X \in \mathbf{R}_\alpha^* \Delta \mathbf{R}\}] = p(1 - Q^*(\alpha))(G(\mathbf{R}_\alpha^*) - G(\mathbf{R}))$$

$$+ (1 - p)Q^*(\alpha) (H(\mathbf{R}) - H(\mathbf{R}_\alpha^*)).$$

Combined with the previous bound, this entails that

$$s_\alpha^2(\mathbf{R}) \leq \frac{c}{1 - 2Q^*(\alpha)} (p(1 - Q^*(\alpha)) (G(\mathbf{R}_\alpha^*) - G(\mathbf{R})) + (1 - p)Q^*(\alpha) (H(\mathbf{R}) - H(\mathbf{R}_\alpha^*)))^\alpha,$$

which concludes the proof. \blacksquare

Proof of Theorem 12

For simplicity, we provide the proof for a finite class \mathcal{R} with cardinality N . First observe that from Theorem 10 and its proof, we have, with probability larger than $1 - \delta$:

$$H(\widehat{\mathbf{R}}_\alpha) \leq \alpha + 2\phi(\delta/2, n), \quad G(\widehat{\mathbf{R}}_\alpha) \geq G(\mathbf{R}_\alpha^*) - 2\phi(\delta/2, n) \quad \text{and} \quad \mathbf{R}_\alpha^* \in \widehat{\mathcal{R}}_\alpha.$$

For all $\mathbf{R} \in \mathcal{R}$, we set:

$$P_n(\mathbf{R}) = \frac{n_+}{n} \left\{ \widehat{G}_n(\mathbf{R}_\alpha^*) - \widehat{G}_n(\mathbf{R}) \right\},$$

so that we have $P_n(\widehat{\mathbf{R}}_\alpha) \leq 0$ since $\mathbf{R}_\alpha^* \in \widehat{\mathcal{R}}_\alpha$. We also introduce:

$$P(\mathbf{R}) = \mathbb{E} \left(\frac{1}{p} P_n(\mathbf{R}) \right) = G(\mathbf{R}_\alpha^*) - G(\mathbf{R}).$$

Now from Bernstein's inequality and the union bound, we have, with probability larger than $1 - \delta$:

$$\forall \mathbf{R} \in \mathcal{R}, \quad pP(\mathbf{R}) \leq P_n(\mathbf{R}) + \sqrt{\frac{2s_\alpha^2(\mathbf{R}) \log(N/\delta)}{n}} + \frac{4 \log(N/\delta)}{3n}.$$

Using this inequality for $\mathbf{R} = \widehat{\mathbf{R}}_\alpha$, we get:

$$p(G(\mathbf{R}_\alpha^*) - G(\widehat{\mathbf{R}}_\alpha)) \leq \sqrt{\frac{2s_\alpha^2(\widehat{\mathbf{R}}_\alpha) \log(N/\delta)}{n}} + \frac{4 \log(N/\delta)}{3n}.$$

We set the notations: $\Delta G = G(\mathbf{R}_\alpha^*) - G(\widehat{\mathbf{R}}_\alpha)$, $\Delta H = H(\widehat{\mathbf{R}}_\alpha) - H(\mathbf{R}_\alpha^*)$ and $q = Q^*(\alpha)$. Then, from the variance control lemma with $\alpha = 1/2$, we get:

$$p\Delta G \leq \sqrt{\frac{2c \log(N/\delta)}{n(1 - 2q)}} \left((p(1 - q)\Delta G)^{1/4} + ((1 - p)q\Delta H)^{1/4} \right) + \frac{4 \log(N/\delta)}{3n}.$$

Eventually, using the control on ΔG and ΔH from Theorem 10, we obtain that there exists a constant $C = C(N, \delta, \alpha, p)$:

$$\Delta G \leq Cn^{-5/8}. \quad \blacksquare$$

Proof of Lemma 13

For all $(s, t) \in \mathcal{S} \times \mathbb{R}$, set $\mathbf{R}_{s,t} = \{x \in \mathcal{X} : s(x) \geq t\}$. For all $s_0 \in \mathcal{S}$, we have

$$\begin{aligned} \mathbf{H}(\widehat{\mathbf{R}}_\alpha(s_0)) &\leq \sup_{(s,t) \in \mathcal{S} \times \mathbb{R}} |\mathbf{H}(\mathbf{R}_{s,t}) - \widehat{\mathbf{H}}_n(\mathbf{R}_{s,t})| + \widehat{\mathbf{H}}_n(\widehat{\mathbf{R}}_\alpha(s_0)) \\ &\leq \sup_{(s,t) \in \mathcal{S} \times \mathbb{R}} |\mathbf{H}(\mathbf{R}_{s,t}) - \widehat{\mathbf{H}}_n(\mathbf{R}_{s,t})| + \alpha + \frac{1}{n}. \end{aligned}$$

As noticed in Remark 14, the collection of sets $\{\mathbf{R}_{s,t}\}_{(s,t) \in \mathcal{S} \times \mathbb{R}}$ has finite VC-dimension, this observation permits to conclude the proof.

Proof of Theorem 16.

The idea of the proof is to relate the excess risk in terms of type I error to the excess risk in terms of weighted classification error. First we re-parameterize the weighted classification error. Set:

$$\ell_\omega(\alpha) = \mathbf{L}_\omega(\mathbf{R}_\alpha^*) = 2(1-p)\omega \alpha + 2p(1-\omega)(1 - \text{ROC}^*(\alpha))$$

Since ROC^* is assumed to be differentiable, it is easy to check that the value $\alpha^* = \mathbf{H}(\mathbf{C}_\omega^*)$ minimizes $\ell_\omega(\alpha)$. Denote by $\ell_\omega^* = \ell_\omega(\alpha^*)$. It follows from a Taylor expansion of $\ell_\omega(\alpha)$ around α^* at the second order that there exists $\alpha_0 \in [0, 1]$ such that:

$$\ell_\omega(\alpha) = \ell_\omega^* - p(1-\omega) \frac{d^2}{d\alpha^2} \text{ROC}^*(\alpha_0) (\alpha - \alpha^*)^2$$

Using also the fact that ROC^* dominates any other curve of the ROC space, we have: $\forall C \subset \mathcal{X}$ measurable, $\mathbf{G}(C) \leq \text{ROC}^*(\mathbf{H}(C))$. Also, by assumption, there exists m such that: $\forall \alpha \in [0, 1]$, $\frac{d^2}{d\alpha^2} \text{ROC}^*(\alpha) \geq -m$. Hence, since $\ell_\omega(\mathbf{H}(\widehat{\mathbf{C}}_\omega)) = \mathbf{L}_\omega(\widehat{\mathbf{C}}_\omega)$, we have:

$$(\mathbf{H}(\widehat{\mathbf{C}}_\omega) - \mathbf{H}(\mathbf{C}_\omega^*))^2 \leq \frac{1}{mp(1-\omega)} (\mathbf{L}_\omega(\widehat{\mathbf{C}}_\omega) - \mathbf{L}_\omega(\mathbf{C}_\omega^*)) .$$

We have obtained the desired inequality. It remains to get the rate of convergence for the weighted empirical risk.

Now set: $F^* = p\mathbf{G}^* + (1-p)\mathbf{H}^*$. We observe that: $\forall t > 0$, $\mathbb{P}(|\eta(X) - \omega| \leq t) = F^*(\omega + t) - F^*(\omega - t) \leq 2t \sup_{\mathbf{u}} (F^*)'(\mathbf{u})$. We have thus shown that the distribution satisfies a modified margin condition [Tsy04], for all $\omega \in [0, 1]$, of the form:

$$\mathbb{P}(|\eta(X) - \omega| \leq t) \leq D t^{\frac{\gamma}{1-\gamma}} .$$

with $\gamma = 1/2$ and $D = 2 \sup_{\mathbf{u}} (F^*)'(\mathbf{u})$. Adapting slightly the argument used in [Tsy04], [BBL05], we have that, under the modified margin condition, there exists a constant c such that, with probability $1 - \delta$:

$$\mathbf{L}_\omega(\widehat{\mathbf{C}}_\omega) - \mathbf{L}_\omega^*(\mathbf{C}_\omega^*) \leq c \left(\frac{\log(1/\delta)}{n} \right)^{\frac{1}{2-\gamma}} .$$

References

- [AA07] J.-Y. Audibert and A. Tsybakov. Fast learning rates for plug-in classifiers. *Annals of statistics*, 35(2):608–633, 2007.
- [AGH⁺05] S. Agarwal, T. Graepel, R. Herbrich, S. Har-Peled, and D. Roth. Generalization bounds for the area under the ROC curve. *Journal of Machine Learning Research*, 6:393–425, 2005.
- [BBL05] S. Boucheron, O. Bousquet, and G. Lugosi. Theory of Classification: A Survey of Some Recent Advances. *ESAIM: Probability and Statistics*, 9:323–375, 2005.
- [BCT07] M. Barreno, A.A. Cardenas, and J.D. Tygar. Optimal ROC curve for a combination of classifiers. In *NIPS'07*, 2007.
- [BDH06] F.R. Bach, D. Heckerman, and Eric Horvitz. Considering cost asymmetry in learning classifiers. *Journal of Machine Learning Research*, 7:1713–1741, 2006.
- [Cav97] L. Cavalier. Nonparametric estimation of regression level sets. *Statistics*, 29:131–160, 1997.
- [CLV05] S. Cléménçon, G. Lugosi, and N. Vayatis. Ranking and scoring using empirical risk minimization. In P. Auer and R. Meir, editors, *Proceedings of COLT 2005*, volume 3559 of *Lecture Notes in Computer Science*, pages 1–15. Springer, 2005.
- [CLV08] S. Cléménçon, G. Lugosi, and N. Vayatis. Ranking and empirical risk minimization of U-statistics. *The Annals of Statistics*, 36(2):844–874, 2008.
- [CM04] C. Cortes and M. Mohri. AUC optimization vs. error rate minimization. In S. Thrun, L. Saul, and B. Schölkopf, editors, *Advances in Neural Information Processing Systems 16*. MIT Press, Cambridge, MA, 2004.
- [CV07] S. Cléménçon and N. Vayatis. Ranking the best instances. *Journal of Machine Learning Research*, 8:2671–2699, 2007.
- [CV08a] S. Cléménçon and N. Vayatis. Approximation of the optimal roc curve and a tree-based ranking algorithm. In *Proceedings of ALT'08*, 2008.
- [CV08b] S. Cléménçon and N. Vayatis. Empirical performance maximization for linear rank statistics. In *Proceedings of NIPS'08*, 2008.
- [CV08c] S. Cléménçon and N. Vayatis. On partitioning rules for bipartite ranking. Technical report, Submitted, 2008.

- [CV08d] S. Cléménçon and N. Vayatis. Tree-structured ranking rules and approximation of the optimal ROC curve. Technical Report hal-00268068, HAL, 2008.
- [DGL96] L. Devroye, L. Györfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition*. Springer, 1996.
- [DL93] R. Devore and G. Lorentz. *Constructive Approximation*. Springer, 1993.
- [Dud99] R.M. Dudley. *Uniform Central Limit Theorems*. Cambridge University Press, 1999.
- [Ega75] J.P. Egan. *Signal Detection Theory and ROC Analysis*. Academic Press, 1975.
- [FISS03] Y. Freund, R. D. Iyer, R. E. Schapire, and Y. Singer. An efficient boosting algorithm for combining preferences. *Journal of Machine Learning Research*, 4:933–969, November 2003.
- [GKKW02] L. Györfi, M. Köhler, A. Krzyzak, and H. Walk. *A Distribution-Free Theory of Nonparametric Regression*. Springer, 2002.
- [HM82] J.A. Hanley and J. McNeil. The meaning and use of the area under a ROC curve. *Radiology*, (143):29–36, 1982.
- [Rak04] A. Rakotomamonjy. Optimizing area under roc curve with svms. In *Proceedings of the First Workshop on ROC Analysis in AI*, 2004.
- [Rud06] C. Rudin. Ranking with a P-Norm Push. In H.U. Simon and G. Lugosi, editors, *Proceedings of COLT 2006*, volume 4005 of *Lecture Notes in Computer Science*, pages 589–604, 2006.
- [RV06] P. Rigollet and R. Vert. Fast rates for plug-in estimators of density level sets. Technical Report arXiv:math/0611473v2, arXiv:math/0611473v2, 2006.
- [SBW08] C. Scott, G. Bellala, and R. Willett. Generalization error analysis for FDR controlled classification. Technical report, Department of Electrical Engineering and Computer Science, University of Michigan, 2008.
- [Sco05] C. Scott. Performance measures for Neyman-Pearson classification. Technical report, Department of Statistics, Rice University, 2005.
- [SN05] C. Scott and R. Nowak. A Neyman-Pearson approach to statistical learning. *IEEE Transactions on Information Theory*, 51(11):3806–3819, November 2005.

- [SN06] C. Scott and R. Nowak. Learning minimum volume sets. *Journal of Machine Learning Research*, 7:665–704, April 2006.
- [Tsy04] A. Tsybakov. Optimal aggregation of classifiers in statistical learning. *Annals of Statistics*, 32(1):135–166, 2004.
- [UAG05] N. Usunier, M. Amini, and P. Gallinari. A data-dependent generalisation error bound for the AUC. In *ROCML ICML 2005 Workshop*, 2005.
- [vT68] H.L. van Trees. *Detection, Estimation, and Modulation Theory, Part I*. John Wiley, 1968.
- [WN07] R. Willett and R. Nowak. Minimax optimal level set estimation. *IEEE Transactions on Image Processing*, 16(12):2965–2979, 2007.
- [YDMW03] L. Yan, R.H. Dodier, M. Mozer, and R.H. Wolniewicz. Optimizing classifier performance via an approximation to the wilcoxon-mann-whitney statistic. In Tom Fawcett and Nina Mishra, editors, *Proceedings of the Twentieth International Conference on Machine Learning (ICML 2003)*, pages 848–855, 2003.