Constr Approx (2010) 32: 619–648
DOI 10.1007/s00365-010-9084-9

CONSTRUCTIVE APPROXIMATION

# Overlaying Classifiers: A Practical Approach to Optimal Scoring

**Stéphan Clémençon · Nicolas Vayatis**

Received: 21 November 2008 / Revised: 30 June 2009 / Accepted: 9 September 2009 /
Published online: 13 February 2010


**Abstract** The Receiver Operating Characteristic (ROC) curve is one of the most widely used visual tools 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 of 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.

Communicated by G. Kerkyacharian.

S. Clémençon (✉)
TSI, Telecom ParisTech, 46, rue Barrault, 75634 Paris cedex 13, France
e-mail: stephan.clemencon@telecom-paristech.fr

N. Vayatis
CMLA—Ecole Normale Supérieure de Cachan, 61 avenue du président Wilson, 94235 Cachan, France
e-mail: nicolas.vayatis@cmla.ens-cachan.fr

**Keywords** Statistical learning · Bipartite ranking · ROC curve · Piecewise linear approximation · Minimum volume set estimation · Density level set · Scoring function · AUC criterion · Sup-norm

**Mathematics Subject Classification (2000)** Primary 62G99 · Secondary 68Q32 · 41A15 · 62G10

## 1 Introduction

In recent 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 the *Receiver Operating Characteristic* (ROC), see [19, 33], or the *Precision-Recall curve* (see [12] and the references therein), allow us 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, judging 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, e.g., [1, 8, 20, 32]), 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 supremum norm. At 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. Since each of these level sets can be seen as the solution of a binary classification problem with asymmetric misclassification costs, the bipartite ranking problem may be viewed as a 'continuum' of binary classification problems. The core of the method described relies on three steps: (i) discretization of the problem, (ii) estimation of a few well-chosen level sets of the regression function where the levels are fixed in advance, and (iii) combination of the estimated classification rules to build a scoring rule with nearly optimal ROC curve. The key point of level sets estimation may be addressed in several ways. In a nonparametric setup, regression or density level sets can be estimated with plug-in methods (see, e.g., [2, 6, 24, 34]). Here, we follow the work in [29] 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. The method leads to a practical ranking algorithm taking advantage of the 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 Sect. 2, we present the scoring problem and recall key notions of ROC analysis. In Sect. 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 Sect. 4 and derive the rate of convergence of an empirical estimate to the optimal ROC curve. In Sect. 5, we consider the subproblem of constrained empirical risk minimization. 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.

## 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 will be needed throughout the paper.

### 2.1 Notations and Setup

Let $\mathcal{X}$ be a measurable space which can be thought of 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 $\mu$ is the marginal distribution of $X$ and $\eta$ is the *regression function* (up to an affine transformation): $\eta(x) = \mathbb{P}\{Y = 1 \mid X = x\}$, $x \in \mathcal{X}$. We will denote by $p = \mathbb{P}\{Y = +1\}$ the expected proportion of positive labels. We denote by $G(dx)$ and $H(dx)$ the conditional distributions of the random variable $X$ given $Y = +1$ and given $Y = -1$, respectively. Hereafter, 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 to ranking the objects $x \in \mathcal{X}$ is to map them onto $\mathbb{R}$ through a certain measurable function $s : \mathcal{X} \to \mathbb{R}$ 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 $\mathcal{D}_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ which mimics the ranking induced by the regression function $\eta$. Hence, ideally, the higher the score $s(X)$ is, the more likely one should 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 $\eta$.

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

$$\mathcal{S}^* = \left\{ s^* = T \circ \eta \;\middle|\; T : [0, 1] \to \mathbb{R} \text{ strictly increasing} \right\}.$$

The statistical problem consists in finding a scoring function as "close" as possible to the class $\mathcal{S}^*$ from the i.i.d. sample $\mathcal{D}_n$. Here, the notion of "closeness" should translate into "induce similar rankings." This notion cannot be captured by usual distances in the functional space where the scoring functions live. Hopefully, the concept of ROC ANALYSIS provides a means for measuring the quality of a scoring function. Hence, two scoring functions will be considered as being "close" if their ROC curves are close in the usual $L_p$ distances in the functional space where the ROC curves live.

*ROC analysis* We now recall the concept of the 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

$$G_s(z) = \mathbb{P}\big\{s(X) \le z \;\big|\; Y = +1\big\},$$
$$H_s(z) = \mathbb{P}\big\{s(X) \le z \;\big|\; Y = -1\big\},$$

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 residual conditional cumulative distribution functions (cdf). 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 $G^*$, $H^*$, $\bar{G}^*$, and $\bar{H}^*$, 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 upon the event $Y = -1$. In particular, the following quantile will be of interest:

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

where we have used here the notion of generalized inverse $F^{-1}(z) = \inf\{t \in \mathbb{R} \mid F(t) \ge z\}$ 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

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

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

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

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

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 to be 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 $\alpha \in [0,1]$:

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

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 $\mathrm{ROC}(s,\alpha)$ representing simply the power of the test of exact level $\alpha$ 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*

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

*and*

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

*Regularity* In this paper, we will assume that $\mathrm{ROC}^*$ is twice differentiable with bounded second derivative. The assumption of twice differentiability of the optimal curve $\mathrm{ROC}^*$ can be translated in terms of the regularity of the conditional distributions of the random variable $\eta(X)$. Indeed, assume that the conditional cumulative distribution functions $G^*$ and $H^*$ of $\eta(X)$ are both differentiable and that $H^{*\prime}$ is continuous and bounded by below by some strictly positive constant on its support. Then, from Proposition 8 in [14], we have $\forall \alpha \in ]0,1]$:

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

In order to guarantee that $\mathrm{ROC}^*$ is differentiable at 0, we need to assume that the likelihood ratio $dG/dH(X)$ is upper bounded almost surely (recall that $G$ and $H$ are assumed absolutely continuous to each other in our setup). As we have

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

this assumption means 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., since the distributions $H$ and $\mu$ are equivalent). In addition, note that, under these assumptions, the distributions $G^*$ and $H^*$ are absolutely continuous with respect to each other. From Corollary 7 in [14], we also have

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

hence, the likelihood ratio $dG^*/dH^*(u)$ remains bounded in the neighborhood of $Q^*(0)$ under the hypothesis stipulated above. This condition also suffices to ensure that ROC$^*$ is twice differentiable on $[0,1]$, since $\forall \alpha \in [0,1]$,

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

We point out that the boundedness assumption on the likelihood ratio is a strong requirement and significantly restricts the range of distributions which can be considered for modeling the data. As shown above, it guarantees that ROC$^*$ is sufficiently regular so that it can be well approximated by a piecewise linear curve with breakpoints fixed in advance (see also Remark 10 for suggestions on how to weaken this assumption). (Relaxing this assumption and building consistent estimators of the optimal ROC curve is the subject of work in 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 that of the optimal elements in $\mathcal{S}^*$ in terms of closeness of their ROC curves. We then need to consider some metric in the space $\mathbb{D}([0,1])$ of càdlàg curves $f : [0,1] \to \mathbb{R}$. Let us denote by $d(s, s^*)$ the distance between ROC curves 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. In the latter problems, $d(s, s^*)$ can be written as a difference $A(s) - A(s^*)$ so that minimizing $d(s, s^*)$ is equivalent to minimizing $A(s)$. Then, if this decomposition holds, by $M$-estimation arguments which are now standard (see [5]), 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 $L_\infty$ cases. We shall see that the 'excess-risk' decomposition applies in the case of the $L_1$-distance, but not for $L_\infty$-distance.

*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

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

By Proposition 3, we have

$$d_1(s,\eta) = \left\|\mathrm{ROC}^*\right\|_1 - \left\|\mathrm{ROC}(s,.)\right\|_1.$$

In this case, minimizing $d(s,\eta)$ boils down to maximizing a popular quantity known as the *Area Under the* ROC *curve* (or AUC, see [22]):

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

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** ([7]) *For any scoring function $s$ such that $H_s$ and $G_s$ are continuous distribution functions, we have*

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

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

From this observation, ranking can indeed be interpreted as classification of pairs of observations. We refer to [8] for a systematic study of related empirical and convex risk minimization strategies which involve $U$-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, [15, 23, 35]). 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 $L_\infty$-distance* As an example of a strong notion of distance, we propose to study the distance induced by the $L_\infty$-norm:

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

The main difficulty in 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 $d_\infty$ 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 $L_\infty$-norm. The key idea is to combine such procedures with an adequate approximation stage, under very mild smoothness assumptions for the optimal ROC curve. More precisely, the ERM strategy will be applied here in order to recover an approximant of the target curve $\mathrm{ROC}^*$, involving a finite number of level sets of the regression function only. As a byproduct of the analysis we will also provide a statistical estimation of the optimal ROC curve which can also be of interest intrinsically.

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

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

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

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

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 $\nu$ is chosen to be the Lebesgue measure. The next proposition also illustrates this view of the problem. We set the notations

$$R^*_\alpha = \big\{x \in \mathcal{X} \,\big|\, \eta(x) > Q^*(\alpha)\big\},$$
$$R_{s,\alpha} = \big\{x \in \mathcal{X} \,\big|\, s(x) > Q\big(s(X),\alpha\big)\big\}.$$

The next result relates pointwise ROC curve maximization to a modified binary classification problem. This provides the main intuition for casting the problem of bipartite ranking as a 'continuum' of classification problems.

**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*

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

*where $\Delta$ 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 $R^*_\alpha$ through $R_{s,\alpha}$.

In contrast, standard binary classification amounts to recovering a single, very specific, $\eta$-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 $C : \mathcal{X} \to \{-1, +1\}$.

*Remark 1* (On the excess of risk) We point out that Proposition 5 generalizes the well-known relationship in the classification setup, see [17]:

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

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

*Plug-in scoring rules* The purpose of this paragraph is to show what to expect from a competitor method, namely the plug-in approach, of the one we develop in this paper, and to state the conditions under which plug-in scoring rules are consistent in the sense of the ROC criterion. Indeed, a possible angle to approximate optimal scoring rules is the *plug-in* approach, see [17] in the context of binary classification. 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 $\mathrm{ROC}(\hat{\eta},\cdot)$ and $\mathrm{ROC}^*$ are also close.

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

(i) *We have*

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

(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* $Q^*(\alpha) < 1$,

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

It clearly follows from (i) that a $L_1(\mu)$-consistent estimator, i.e., an estimator $\hat{\eta}_n(x)$ such that $\mathbb{E}(|\hat{\eta}_n(X) - \eta(X)|) \to 0$ as $n \to \infty$ with probability one, yields a consistent ranking rule in the AUC-sense. However, guaranteeing the pointwise convergence $\mathrm{ROC}^*(\alpha) - \mathrm{ROC}(\hat{\eta}_n,\alpha) \to 0$ is more difficult: in addition to $L_1(\mu)$-consistency, it would require that $\hat{\eta}_n(X)$ has a density uniformly bounded in $n$. We also point out that plug-in rules face computational difficulties when dealing with high-dimensional data (see, e.g., [21]). 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 $R^*_{\alpha_1},\ldots,R^*_{\alpha_K}$, where the subdivision $\sigma : 0 = \alpha_0 < \alpha_1 \leq \cdots \leq \alpha_K \leq \alpha_{K+1} = 1$ is fixed in advance and $K$ is a tuning parameter that controls the complexity of the method:

$$s^*_\sigma(x) = \sum_{i=1}^{K} \mathbb{I}\big\{x \in R^*_{\alpha_i}\big\}, \tag{2}$$

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

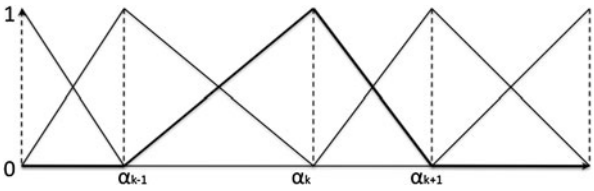


**Fig. 1** Graph of the "hat function" $\Phi_k^*$ (*bold line*)

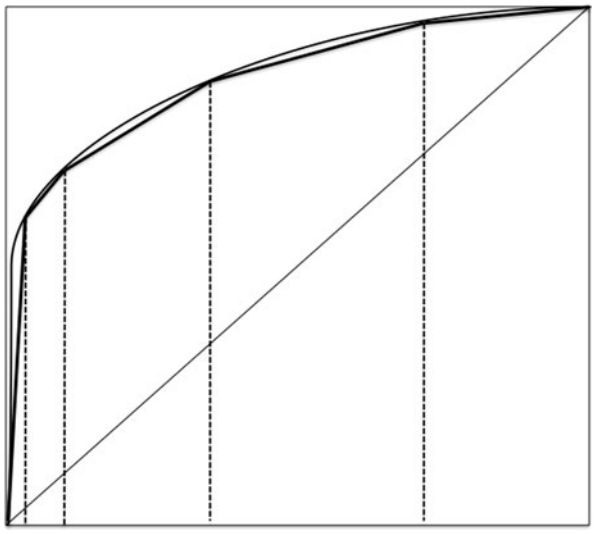

**Fig. 2** 2-spline approximant of the curve ROC*

### 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 $\{(\alpha_i, \mathrm{ROC}^*(\alpha_i)),\ 0 \le i \le K+1\}$. In order to make the latter explicit, we classically consider the "hat functions" related to the mesh grid $\{\alpha_i;\ 0 \le i \le K+1\}$ (see Fig. 1): $\forall i \in \{1, \ldots, K\}$, $\forall \alpha \in [0,1]$,

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

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

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

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

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

which may serve as a simple *approximant* of the optimal curve ROC*, see Fig. 2.

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

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

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

*Remark 2* (On adaptive approximation by 2-splines) 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 Chap. 12 in [16]). However, any practical procedure permitting the achievement of this approximation rate under weaker hypotheses would require choosing the breakpoints $\alpha_k$ depending on the properties of the target curve $\mathrm{ROC}^*$, instead of fixing them in advance. In order to consider a more general setup, including cases where the essential supremum of $\eta(X)$ is equal to 1 (i.e., $\lim_{\alpha \to 0} \mathrm{ROC}^{*\prime}(\alpha) = +\infty$), extensions of the new approach developed in this article will be tackled in a future paper, where the mesh grid is refined adaptively from the data. Incidentally, we point out that the spacings $\Delta_k$ between the breakpoints should be ideally chosen nondecreasing, given the geometry of the optimal ROC curve (concave and strictly increasing, see Fig. 2).

### 3.2 Empirical MV-set Estimation

In this section, we shall introduce a procedure for estimating the discrete scoring function $s_\sigma^*$ for a given mesh grid $\sigma$ of $[0, 1]$. This method will be based on the statistical estimation of the sets $R_\alpha^* = \{x \in \mathcal{X} \mid \eta(x) > Q^*(\alpha)\}$ for specific choices of $\alpha \in (0, 1)$, the points of the discrete grid $\sigma$. 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.

Interestingly, the level set $R_\alpha^*$ can be interpreted as the solution of the constrained optimization problem

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

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 $\mathcal{H}_0 : Y = -1$ against the alternative $\mathcal{H}_1 : Y = +1$ with a type I error equal to $\alpha$ and maximal power.

Note that this formulation is equivalent to the Minimum Volume (MV) set estimation framework (see [29] and references therein), since the complement $S_\alpha^* = \mathcal{X} \setminus R_\alpha^*$ may be seen as the solution of

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

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

$$\widehat{G}_n = \frac{1}{n_+}\sum_{i=1}^{n} \mathbb{I}\{Y_i = +1\} \cdot \delta_{X_i} \quad \text{and} \quad \widehat{H}_n = \frac{1}{n_-}\sum_{i=1}^{n} \mathbb{I}\{Y_i = -1\} \cdot \delta_{X_i},$$

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

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

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

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

### 3.3 The RankOver Algorithm

We now describe a very simple ranking procedure which builds an estimator of $s^*_\sigma$ 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 $\sigma_K$. 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 Sect. 5), we shall explore three possible strategies to solve this constrained classification problem: (i) empirical MV-set, (ii) threshold rules, and (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 Sect. 4. The other parameters of the algorithm are the partition $\sigma_K : \alpha_0 = 0 < \alpha_1 < \cdots < \alpha_K < \alpha_{K+1} = 1$, $K \geq 1$, and the tolerance parameter denoted by $\phi$.

Statistical performance of this procedure will be discussed later. 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 $(\widehat{R}_k)_{k\geq 1}$ could be to proceed in a top-down

THE RANKOVER ALGORITHM

**Input.** Mesh grid $\sigma_K$, tolerance parameter $\phi$, class $\mathcal{R}$ of sets

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

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

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

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

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

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

manner. Start with $\widetilde{R}_{K+1} = \mathcal{X}$ and $\widetilde{R}_k = \widetilde{R}_{k+1} \cap \widehat{R}_k$ for $k = K, \ldots, 1$. Results similar to those established in this paper could easily be derived from 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: $\sum_{k=1}^{K} (\alpha_k - \alpha_{k-1})\, \mathbb{I}\{x \in \widetilde{R}_k\}$. The latter quantity can be seen as a Riemann sum which approximates the integral (1) when $\nu$ is taken to be the Lebesgue measure.

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 $(\tilde{\alpha}_k, \tilde{\beta}_k) = (\widehat{H}_n(\widetilde{R}_k), \widehat{G}_n(\widetilde{R}_k))$ for all $k \in \{0, \ldots, K+1\}$, where by convention $\widetilde{R}_0 = \emptyset$ and $\widetilde{R}_{K+1} = \mathcal{X}$. We point out that the empirical ROC curve of the scoring function built by the RANKOVER algorithm is the piecewise linear function

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

where $\tilde{\Phi}_k = \Phi(\alpha, \tilde{\alpha}_{k-1}, \tilde{\alpha}_k) - \Phi(\alpha, \tilde{\alpha}_k, \tilde{\alpha}_{k+1})$ for all $k \in \{1, \ldots, K\}$ and $\tilde{\Phi}_{K+1}(\alpha) = \Phi(\alpha, \tilde{\alpha}_K, 1)$. Note that the quantity $\widetilde{\mathrm{ROC}}(s_K, \cdot)$ is the statistical version of $\mathrm{ROC}(s_K, \cdot)$ obtained by replacing $G$ and $H$ in Definition 2 by their respective empirical counterparts. Moreover, it follows from this expression that the corresponding empirical AUC, i.e., the area under the empirical curve $\widetilde{\mathrm{ROC}}(s_K, .)$, is given by

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

### 3.4 Algorithmic Approaches to Scoring

Until recently, the one and only known method for ranking/scoring binary-valued data was *logistic regression* and its numerous variants. Problems with high-dimensional data such 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 [20]) exporting the boosting approach of the 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 [9] and [11]). Indeed, many empirical summaries of the ROC curve such as the AUC ([22]), the local AUC ([9]), and the $p$-norm push ([25]) 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 these empirical risk maximization strategies require the control of a new class of stochastic processes, called rank processes (see [11] for preliminary results).
- **Partitioning methods** We have developed various partitioning methods for bipartite ranking. In [13], 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 [10, 14]).

Instead of partitioning the input space, the approach taken in this paper consists of taking a partition of the $x$-axis 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 Sect. 4). The main question for practitioners would be how to implement the **Optimization** step. For some clues on practical strategies devoted to this problem, see Sect. 7 of [29] 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. 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 $R_\alpha^*$ 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*

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

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

$$\left\| \mathrm{ROC}^*(.) - \mathrm{ROC}(s_{K_n}, .) \right\|_\infty \le cn^{-1/3}.$$

We first discuss the nature of the assumptions. A few remarks are in order.

*Remark 5* (On the class of candidate level sets containing the target set) The assumption $R_\alpha^* \in \mathcal{R}$ for all $\alpha \in (0,1)$ is a heavy assumption. However, it could easily be relaxed at the cost of additional technicalities. Indeed, one could define the following set:

$$\check{R}_\alpha = \arg\max_{R \in \mathcal{R}} G(R) \quad \text{s. t. } H(R) \le \alpha.$$

Then the result is true even if $R_\alpha^* \notin \mathcal{R}$, but we have to replace

$$G\big(\hat{R}_\alpha\big) \ge G\big(R_\alpha^*\big) - 2\phi \quad \text{by } G\big(\hat{R}_\alpha\big) \ge G\big(\check{R}_\alpha\big) - 2\phi$$

(and, as a particular case, we get $\check{R}_\alpha = R_\alpha^*$ if $R_\alpha^* \in \mathcal{R}$). Then it would remain to control the measure of the symmetric difference between $\check{R}_\alpha$ and $R_\alpha^*$ with respect to the measures $G$ and $H$. Assumptions on the regularity of the boundary of the level sets similar to those in [30] will be required to state the corresponding result.

*Remark 6* (Choice of the penalty) The issue of penalty calibration has been a topic of intensive research in recent years (see [5] and references therein). We do not include 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 [5]).

*Remark 7* (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 8* (Calibration of the size $K$ of the mesh grid) The proof of the theorem reveals the intuitive trade-off involving the optimal calibration of $K = K_n$. On the one hand, it is necessary to have $K_n$ tending to infinity in order to reduce the approximation error between the optimal curve which is smooth (twice differentiable) and its piecewise linear approximation of dimension $K_n + 1$. On the other hand, overlaying classifiers induces stacking of the errors committed at various levels, so that $K_n$ should not grow too fast with $n$ to infinity.

The previous theorem could be extended in two directions as illustrated in the following comments.

*Remark 9* (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 in progress.

*Remark 10* (Adaptivity of the partition) A natural extension of the approach would be to consider a flexible mesh grid $\{\alpha_k\}$ 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 ([14]), but the rates of convergence obtained in the present paper are faster.

Finally, we also mention a connection to previous work in the context of classification.

*Remark 11* (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 [3] and [4].

### 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 empirical estimate of the optimal ROC curve that achieves a faster rate of convergence than the empirical ROC curve $\widetilde{\mathrm{ROC}}(s_K, .)$, which suffers from the loss of pointwise accuracy due to the **Monotonicity** step (see the proof of Theorem 8). We introduce some notations. Set $\forall k \in \{0, \ldots, K+1\}, \hat{\alpha}_k = \widehat{H}_n(\widehat{R}_k)$, and $\hat{\beta}_k = \widehat{G}(\widehat{R}_k)$. The broken line that connects the knots $\{(\hat{\alpha}_k, \hat{\beta}_k);\ 0 \le k \le K+1\}$ provides an empirical counterpart of the piecewise linear approximant of the optimal ROC$^*$. We also introduce the "hat functions" defined by

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

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

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

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* $K = K_n \sim n^{1/4}$. *Then there exists a constant* $c = c(\delta)$ *such that, with probability at least* $1 - \delta$,

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

*Remark 12* (A nondecreasing estimate of ROC*) Notice that the curve $\widehat{\mathrm{ROC}^*}(.)$ is not necessarily increasing, in contrast to the empirical ROC curve $\widetilde{\mathrm{ROC}}(s_K, .)$. However, the accuracy of the latter estimate is worst, of the order $O_{\mathbb{P}}((\log(n)/n)^{1/3})$ at best, with a number of knots of the order $n^{1/6}$ only, while $K_n$ is of the order $n^{1/4}$ in the theorem above. As may be shown by a careful examination of the proof of Theorem 8, this is a consequence of the monotonicity condition set in the second step of the RANKOVER algorithm and the resulting pile-up error phenomenon. We also point out that the graph of the mapping $\overline{\mathrm{ROC}^*} : \alpha \in (0,1) \mapsto \sum_{k=1}^{K+1} \hat{\beta}_k \Phi_k^*(\alpha)$ provides a consistent estimate of ROC*, achieving the same rate of convergence as $\widehat{\mathrm{ROC}^*}$.

## 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 $R_{\alpha_k}^*$ through solving the following problem:

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

In other words, the goal is to select a critical region $R \in \mathcal{R}$ in order to construct a decision rule based on the observation $X$, i.e., a classifier $C(X) = 2 \cdot \mathbb{I}\{X \in R\} - 1$, for testing the null hypothesis $\mathcal{H}_0 : Y = -1$ with type I error $\alpha$ 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 [29] on learning minimum volume sets, but we also refer to [28] and [26]. We extend their result to the case

where the reference measure is unknown and provide fast rates of convergence of MV-set estimators. At 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 $\widehat{R}_\alpha$ 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 $\widehat{R}_\alpha$ with a simultaneous control of the type I error. A similar result was also obtained in [27].

The main assumptions for consistency results to hold concern the complexity of the collection $\mathcal{R}$ of candidate sets, as well as its capacity to represent the target set $R^*_\alpha$. 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 $R^*_\alpha$ 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 $1-\delta$: $\forall n \geq 1$,*

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

*Remark 13* (On recovering a point on the optimal ROC curve) When the cdf $H^*$ (respectively $G^*$) is continuous at $Q^*(\alpha)$, the point $(H(R^*_\alpha), G(R^*_\alpha))$ naturally coincides with the point on $(\alpha, \mathrm{ROC}^*(\alpha))$ of the optimal curve. As may be shown by examining Theorem 10's proof, the euclidean distance in the ROC space of the point $(\widehat{H}_n(\widehat{R}_\alpha), \widehat{G}_n(\widehat{R}_\alpha))$ determined by solving the constrained ERM problem (3) to $(\alpha, \mathrm{ROC}^*(\alpha))$ is then of order $O_{\mathbb{P}}(1/\sqrt{n})$. This will be exploited later when constructing an estimate of the curve $\mathrm{ROC}^*$ with a controlled approximation error.

### 5.2 Fast (but not so Fast) Rates of Convergence

We now show assumptions under which faster rates of convergence can be attained. In [31], 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 $a \in (0,1)$ and $D > 0$ such that $\forall t \geq 0$,

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

We point out that this assumption corresponds to the one introduced in [31], except that here the quantile $Q^*(\alpha)$ replaces $1/2$.

*Remark 14* (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 $a = 1/2$ and $D = \sup_t f^*(t)$. Indeed, the finite increments theorem yields

$$\mathbb{P}\{|\eta(X) - Q^*(\alpha)| \le t\} = F^*(Q^*(\alpha)+t) - F^*(Q^*(\alpha)-t) \le 2Dt.$$

The next result describes an important consequence of this condition.

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

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

*Then we have*

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

**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* $1-\delta$: $\exists C = C(\mathcal{R}, \delta, \alpha, p), n_0, \forall n \ge n_0$,

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

*Remark 15* (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 $\{R \in \mathcal{R} : H(R) \le \alpha\}$, the rate of the excess of risk (in terms of type II error) is of the order of $O(n^{-1/(2-a)})$. 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 $a = 1/2$.

*Remark 16* (Fast estimation of the curve ROC$^*$) One may easily see that, under the additional assumption that $F^*$'s density is bounded, the estimate $\widehat{\mathrm{ROC}^*}$ of the optimal curve introduced in Remark 12 inherits the $n^{-5/8}$ convergence rate obtained for the pointwise deviation considered above. However, as may be shown by a careful examination of Theorem 9's proof in the Appendix section, this is not true for the estimator $\widehat{\mathrm{ROC}^*}$ due to the impact of the deviations $|\hat{\alpha}_k - \alpha_k|$ involved in the control of the error.

### 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 [9]. 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 $R_\alpha^*$. 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 that the constraint is satisfied, we could consider sets of the form $R_\alpha(s) = \{x \in \mathcal{X} :\ s(x) \geq Q(s(X), \alpha)\}$, 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 $\widehat{Q}_n(s, \alpha) = \widehat{H}_n^{-1}(1-\alpha)$, which leads to the consideration actually of the set

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

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 $n \geq 1$:

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

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 $V$. These, for all $\delta \in (0,1)$, we have with probability at least $1-\delta$,*

$$\sup_{s \in \mathcal{S}} H\big(\widehat{R}_\alpha(s)\big) \leq \alpha + \widetilde{\phi}(\delta, n).$$

*Remark 17* (On the complexity assumption) For further details on the terminology of major sets and major classes one may refer to [18]. These 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 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

$$\hat{s}_n = \arg\max_{s \in \mathcal{S}} \widehat{G}_n\big(R_\alpha(s)\big).$$

**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 $1-\delta$:*

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

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

*Remark 18* (On fast rates) We also point out that this result may be viewed as a variant of Theorem 5 in [9], related to the so-called *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 $\mu$. In addition, it is noteworthy that, under Theorem 12's conditions combined with the assumption that the cdfs $H_s$ 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 [9].

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

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

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 $\eta$-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* $C \subset \mathcal{X}$,

$$L_\omega\big(C^*_\omega\big) \leq L_\omega(C).$$

*Also the optimal error is given by*

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

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

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

*where* $\Delta$ *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 the consideration of the *weighted empirical risk minimizer* over a class $\mathcal{R}$ of candidate sets:

$$\hat{C}_\omega = \operatorname*{arg\,min}_{C \in \mathcal{R}} \hat{L}_\omega(C).$$

The next result provides rates of convergence of the weighted empirical risk minimizer $\hat{C}_\omega$ 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 $C^*_\omega$. 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 $\omega$ such that, with probability at least $1-\delta$,*

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

*The same result also holds for the excess risk of $\hat{C}_\omega$ 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 usual classification methods can readily be used for recovering a collection of $\eta$-level sets, the empirical MV-set approach should be preferred regarding the rate of convergence.

## 6 Conclusion

In this paper, we propose 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, and weighted empirical risk minimization. Several issues remain open, including optimality of convergence rate bounds, adaptive grid for approximation, and practical implementations of empirical MV-set estimation. Their investigation is undertaken through ongoing projects.

**Acknowledgements** We are grateful to the anonymous referees for their comments, which helped us to improve the clarity of presentation.

## Appendix: Proof Section

*Proof of Proposition 5* First, we observe that, for any measurable function $h$, we have, by a change of probability, that

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

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

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

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

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

we get

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

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

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

*Proof of Proposition 6* We recall (see [8]) that

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

where

$$\Gamma = \big\{(x,x') : \mathrm{sgn}\big(\hat\eta(X)-\hat\eta\big(X'\big)\big) \neq \mathrm{sgn}\big(\eta(X)-\eta\big(X'\big)\big)\big\}.$$

But one may easily check that

$$\begin{aligned} &\text{if } \mathrm{sgn}\big(\hat\eta(X)-\hat\eta\big(X'\big)\big) \neq \mathrm{sgn}\big(\eta(X)-\eta\big(X'\big)\big), \quad \text{then} \\ &\big|\eta(X)-\eta\big(X'\big)\big| \leq \big|\eta(X)-\hat\eta(X)\big| + \big|\eta\big(X'\big)-\hat\eta\big(X'\big)\big|, \end{aligned}$$

which gives the first part of the result.

Turning to the second assertion, consider the event

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

Notice first that, after Proposition 5, we have

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

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

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

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

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

which, combined with the previous bound, proves the second part. $\square$

*Proof of Theorem 8* We note $\tilde{\alpha}_i = H(\widetilde{R}_i)$, $\tilde{\beta}_i = G(\widetilde{R}_i)$ and also $\widetilde{\Phi}_i(\cdot) = \Phi(\cdot(\widetilde{\alpha}_{i-1}, \widetilde{\alpha}_i)) - \Phi(\cdot;\ (\widetilde{\alpha}_i, \widetilde{\alpha}_{i+1}))$. We then have

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

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

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

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

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

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

$$\big|\mathrm{ROC}^*(\widetilde{\alpha}_i) - \widetilde{\beta}_i\big| \le \sup_{\alpha \in [0,1]} \frac{d}{d\alpha} \mathrm{ROC}^*(\alpha) \cdot |\widetilde{\alpha}_i - \alpha_i| + \big|\mathrm{ROC}^*(\alpha_i) - \widetilde{\beta}_i\big|.$$

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

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

$$H\big(\widetilde{R}_k\big) = H\big(\widehat{R}_{\alpha_k}\big) + (k-1)\, O_{\mathbb{P}}\big(\phi(\delta, n)\big),$$

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

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

$$|\widetilde{\alpha}_{i+1} - \widetilde{\alpha}_i| \le \max_{1 \le k \le K} \big|H\big(\widehat{R}_{\alpha_k}\big) - H\big(\widehat{R}_{\alpha_{k-1}}\big)\big| + K\, O_{\mathbb{P}}\big(\phi(\delta, n)\big).$$

We have that

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

As before, we have that the first term is of the order $\phi(\delta/K, n)$, 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 $K^2A_n^2 + K^2n^{-1} + K^{-2} + A_n^2 + \log K/n + KA_n + Kn^{-1/2} + A_n + (\log K/n)^{1/2}$. 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 $H(\widetilde{R}_2) = H(\widehat{R}_{\alpha_2}) + H(\widehat{R}_{\alpha_1} \setminus \widehat{R}_{\alpha_2})$. Therefore, since $R^*_{\alpha_1} \subset R^*_{\alpha_2}$ and observing that

$$H\big(\widehat{R}_{\alpha_1} \setminus \widehat{R}_{\alpha_2}\big) = H\Big(\big(\big(\widehat{R}_{\alpha_1} \setminus R^*_{\alpha_1}\big) \cup \big(\widehat{R}_{\alpha_1} \cap R^*_{\alpha_1}\big)\big) \setminus \big(\big(\widehat{R}_{\alpha_2} \setminus R^*_{\alpha_2}\big) \cup \big(\widehat{R}_{\alpha_2} \cap R^*_{\alpha_2}\big)\big)\Big),$$

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

*Proof of Theorem 9* We use the following decomposition, for any fixed $\alpha \in (0, 1)$:

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

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

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

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

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

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

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

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

We recall that $\max_{1\le i\le K} |\hat{\alpha}_i - \bar{\alpha}_i| = O_{\mathbb{P}}(Kn^{-1/2})$. Moreover, $\max_{0\le i\le K}\{\alpha_{i+1} - \alpha_i\}$ is of the order of $K^{-1}$. And with probability at least $1 - \delta$, we have that $\max_{1\le i\le K} |\alpha_i - \bar{\alpha}_i|$ is bounded as in Theorem 10, except that $\delta$ is replaced by $\delta/K$ 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

$$\widehat{\mathcal{R}}_\alpha = \big\{R \in \mathcal{R} : \widehat{H}_n(R) \le \alpha + \phi(\delta/2, n)\big\},$$

so that one may write

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

We shall consider the following events:

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

as well as

$$\Omega_H = \Big\{\sup_{R \in \mathcal{R}} \big|\widehat{H}_n(R) - H(R)\big| > \phi(\delta/2, n)\Big\} \quad \text{and}$$
$$\Omega_G = \Big\{\sup_{R \in \mathcal{R}} \big|\widehat{G}_n(R) - G(R)\big| > \phi(\delta/2, n)\Big\}.$$

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 [5] for details). We have, indeed, that, for all $\delta \in (0, 1)$, the event $\Omega_H$ (respectively, the event $\Omega_G$) 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 $\Omega_H^c$ we have

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

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

$$\begin{aligned} G\big(\widehat{R}_\alpha\big) &= \big(G\big(\widehat{R}_\alpha\big) - \widehat{G}_n\big(\widehat{R}_\alpha\big)\big) + \big(\widehat{G}_n\big(\widehat{R}_\alpha\big) - \widehat{G}_n\big(R^*_\alpha\big)\big) \\ &\quad + \big(\widehat{G}_n\big(R^*_\alpha\big) - G\big(R^*_\alpha\big)\big) + G\big(R^*_\alpha\big), \end{aligned}$$

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

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

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 [5]) to an arbitrary level. We use the following equivalent condition: there exists a positive constant $c$ such that, for any set $R$, we have

$$\mathbb{E}\big(\mathbb{I}\big\{X \in R_\alpha^* \Delta R\big\}\big) \le c\big(F\big(R_\alpha^*\big) - F(R)\big)^a,$$

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

$$s_\alpha^2(R) \le c\big(\mathbb{E}\big(\big|\eta(X) - Q^*(\alpha)\big| \cdot \mathbb{I}\big\{X \in R_\alpha^* \Delta R\big\}\big)\big)^a,$$

by Hölder's inequality.

Observe also that

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

and

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

This yields

$$\begin{aligned} \mathbb{E}\big[\big|\eta(X) - Q^*(\alpha)\big| \cdot \mathbb{I}\big\{X \in R_\alpha^* \Delta R\big\}\big] = p\big(1 - Q^*(\alpha)\big)\big(G\big(R_\alpha^*\big) - G(R)\big) \\ + (1-p)Q^*(\alpha)\big(H(R) - H\big(R_\alpha^*\big)\big). \end{aligned}$$

Combined with the previous bound, this entails that

$$\begin{aligned} s_\alpha^2(R) \le \frac{c}{1 - 2Q^*(\alpha)}\big(p\big(1 - Q^*(\alpha)\big)\big(G\big(R_\alpha^*\big) - G(R)\big) \\ + (1-p)Q^*(\alpha)\big(H(R) - H\big(R_\alpha^*\big)\big)\big)^a, \end{aligned}$$

which concludes the proof. □

*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 $1-\delta$,

$$H\big(\widehat{R}_\alpha\big) \le \alpha + 2\phi(\delta/2, n), \qquad G\big(\widehat{R}_\alpha\big) \ge G\big(R_\alpha^*\big) - 2\phi(\delta/2, n) \quad \text{and} \quad R_\alpha^* \in \widehat{\mathcal{R}}_\alpha.$$

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

$$P_n(R) = \frac{n_+}{n}\big\{\widehat{G}_n\big(R_\alpha^*\big) - \widehat{G}_n(R)\big\},$$

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

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

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

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

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

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

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

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

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

$$\Delta G \le Cn^{-5/8}. \qquad \square$$

*Proof of Lemma 13* For all $(s,t) \in \mathcal{S} \times \mathbb{R}$, set $R_{s,t} = \{x \in \mathcal{X} : s(x) \ge t\}$. For all $s_0 \in \mathcal{S}$, we have

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

As noticed in Remark 17, the collection of sets $\{R_{s,t}\}_{(s,t)\in\mathcal{S}\times\mathbb{R}}$ has finite VC-dimension. This observation permits us to conclude the proof. $\square$

*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

$$\ell_\omega(\alpha) = L_\omega\big(R_\alpha^*\big) = 2(1-p)\omega\alpha + 2p(1-\omega)\big(1 - \mathrm{ROC}^*(\alpha)\big).$$

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

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

Using also the fact that $\mathrm{ROC}^*$ dominates any other curve of the ROC space, we have $\forall C \subset \mathcal{X}$ measurable, $G(C) \le \mathrm{ROC}^*(H(C))$. Also, by assumption, there exists $m$

such that $\forall \alpha \in [0,1]$, $\frac{d^2}{d\alpha^2}\mathrm{ROC}^*(\alpha) \geq -m$. Hence, since $\ell_\omega(H(\hat{C}_\omega)) = L_\omega(\hat{C}_\omega)$, we have

$$\big(H\big(\hat{C}_\omega\big) - H\big(C^*_\omega\big)\big)^2 \leq \frac{1}{mp(1-\omega)}\big(L_\omega\big(\hat{C}_\omega\big) - L_\omega\big(C^*_\omega\big)\big).$$

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

Now set $F^* = pG^* + (1-p)H^*$. We observe that $\forall t > 0$, $\mathbb{P}(|\eta(X) - \omega| \leq t) = F^*(\omega + t) - F^*(\omega - t) \leq 2t \sup_u (F^*)'(u)$. We have thus shown that the distribution satisfies a modified margin condition [31], for all $\omega \in [0,1]$, of the form

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

with $\gamma = 1/2$ and $D = 2\sup_u (F^*)'(u)$. Adapting slightly the argument used in [5, 31], we have that, under the modified margin condition, there exists a constant $c$ such that, with probability $1-\delta$,

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

□

## References


1. Agarwal, S., Graepel, T., Herbrich, R., Har-Peled, S., Roth, D.: Generalization bounds for the area under the ROC curve. J. Mach. Learn. Res. **6**, 393–425 (2005)
2. Audibert, J.-Y., Tsybakov, A.: Fast learning rates for plug-in classifiers. Ann. Stat. **35**(2), 608–633 (2007)
3. Bach, F.R., Heckerman, D., Horvitz, E.: Considering cost asymmetry in learning classifiers. J. Mach. Learn. Res. **7**, 1713–1741 (2006)
4. Barreno, M., Cardenas, A.A., Tygar, J.D.: Optimal ROC curve for a combination of classifiers. In: NIPS'07 (2007)
5. Boucheron, S., Bousquet, O., Lugosi, G.: Theory of classification: a survey of some recent advances. ESAIM: Probab. Stat. **9**, 323–375 (2005)
6. Cavalier, L.: Nonparametric estimation of regression level sets. Statistics **29**, 131–160 (1997)
7. Clémençon, S., Lugosi, G., Vayatis, N.: Ranking and scoring using empirical risk minimization. In: Auer, P., Meir, R. (eds.) Proceedings of COLT 2005. Lecture Notes in Computer Science, vol. 3559, pp. 1–15. Springer, Berlin (2005)
8. Clémençon, S., Lugosi, G., Vayatis, N.: Ranking and empirical risk minimization of U-statistics. Ann. Stat. **36**(2), 844–874 (2008)
9. Clémençon, S., Vayatis, N.: Ranking the best instances. J. Mach. Learn. Res. **8**, 2671–2699 (2007)
10. Clémençon, S., Vayatis, N.: Approximation of the optimal ROC curve and a tree-based ranking algorithm. In: Proceedings of ALT'08 (2008)
11. Clémençon, S., Vayatis, N.: Empirical performance maximization for linear rank statistics. In: Koller, D., Schuurmans, D., Bengio, Y., Bottou, L. (eds.) Advances in Neural Information Processing Systems 21. Proceedings of NIPS'08, pp. 305–312 (2009)
12. Clémençon, S., Vayatis, N.: Nonparametric estimation of the precision-recall curve. In: Proceedings of the 26th International Conference on Machine Learning, ICML (2009)
13. Clémençon, S., Vayatis, N.: On partitioning rules for bipartite ranking. In: Koller, D., Schuurmans, D., Bengio, Y., Bottou, L. (eds.) Proceedings of the Twelfth International Conference on Artificial Intelligence and Statistics (AISTATS), JMLR: W&CP 5, pp. 97–104 (2009)
14. Clémençon, S., Vayatis, N.: Tree-based ranking methods. IEEE Trans. Inf. Theory **55**(9), 4316–4336 (2009)

15. Cortes, C., Mohri, M.: AUC optimization vs. error rate minimization. In: Thrun, S., Saul, L., Schölkopf, B. (eds.) Advances in Neural Information Processing Systems, vol. 16. MIT Press, Cambridge (2004)
16. Devore, R., Lorentz, G.: Constructive Approximation. Springer, Berlin (1993)
17. Devroye, L., Györfi, L., Lugosi, G.: A Probabilistic Theory of Pattern Recognition. Springer, Berlin (1996)
18. Dudley, R.M.: Uniform Central Limit Theorems. Cambridge University Press, Cambridge (1999)
19. Egan, J.P.: Signal Detection Theory and ROC Analysis. Academic Press, San Diego (1975)
20. Freund, Y., Iyer, R.D., Schapire, R.E., Singer, Y.: An efficient boosting algorithm for combining preferences. J. Mach. Learn. Res. **4**, 933–969 (2003)
21. Györfi, L., Köhler, M., Krzyzak, A., Walk, H.: A Distribution-Free Theory of Nonparametric Regression. Springer, Berlin (2002)
22. Hanley, J.A., McNeil, J.: The meaning and use of the area under a ROC curve. Radiology **143**, 29–36 (1982)
23. Rakotomamonjy, A.: Optimizing area under ROC curve with svms. In: Proceedings of the First Workshop on ROC Analysis in AI (2004)
24. Rigollet, P., Vert, R.: Fast rates for plug-in estimators of density level sets. Technical Report (2006). arXiv:math/0611473v2
25. Rudin, C.: Ranking with a P-norm push. In: Simon, H.U., Lugosi, G. (eds.) Proceedings of COLT 2006. Lecture Notes in Computer Science, vol. 4005, pp. 589–604. Springer, Berlin (2006)
26. Scott, C.: Performance measures for Neyman-Pearson classification. Technical report, Department of Statistics, Rice University (2005)
27. Scott, C., Bellala, G., Willett, R.: Generalization error analysis for FDR controlled classification. Technical report, Department of Electrical Engineering and Computer Science, University of Michigan (2008)
28. Scott, C., Nowak, R.: A Neyman-Pearson approach to statistical learning. IEEE Trans. Inf. Theory **51**(11), 3806–3819 (2005)
29. Scott, C., Nowak, R.: Learning minimum volume sets. J. Mach. Learn. Res. **7**, 665–704 (2006)
30. Tsybakov, A.: On nonparametric estimation of density level sets. Ann. Stat. **25**(3), 948–969 (1997)
31. Tsybakov, A.: Optimal aggregation of classifiers in statistical learning. Ann. Stat. **32**(1), 135–166 (2004)
32. Usunier, N., Amini, M., Gallinari, P.: A data-dependent generalisation error bound for the AUC. In: ROCML ICML 2005 Workshop (2005)
33. van Trees, H.L.: Detection, Estimation, and Modulation Theory, Part I. Wiley, New York (1968)
34. Willett, R., Nowak, R.: Minimax optimal level set estimation. IEEE Trans. Image Process. **16**(12), 2965–2979 (2007)
35. Yan, L., Dodier, R.H., Mozer, M., Wolniewicz, R.H.: Optimizing classifier performance via an approximation to the Wilcoxon–Mann–Whitney statistic. In: Fawcett, T., Mishra, N. (eds.) Proceedings of the Twentieth International Conference on Machine Learning (ICML 2003), pp. 848–855 (2003)