Optimal rates for F-score binary classification

Evgenii Chzhen

Optimal rates for F-score binary classification

Evgenii Chzhen

Abstract We study the minimax settings of binary classification with F-score under the β-smoothness assumptions on the regression function $\eta(x) = \mathbb{P}(Y = 1|X = x)$ for $x \in \mathbb{R}^d$. We propose a classification procedure which under the α-margin assumption achieves the rate $\mathcal{O}(n^{-(1+\alpha)\beta/(2\beta+d)})$ for the excess F-score. In this context, the Bayes optimal classifier for the F-score can be obtained by thresholding the aforementioned regression function η on some level θ^* to be estimated. The proposed procedure is performed in a semi-supervised manner, that is, for the estimation of the regression function we use a labeled dataset of size $n \in \mathbb{N}$ and for the estimation of the optimal threshold θ^* we use an unlabeled dataset of size $N \in \mathbb{N}$. Interestingly, the value of $N \in \mathbb{N}$ does not affect the rate of convergence, which indicates that it is "harder" to estimate the regression function η than the optimal threshold θ^*. This further implies that the binary classification with F-score behaves similarly to the standard settings of binary classification. Finally, we show that the rates achieved by the proposed procedure are optimal in the minimax sense up to a constant factor.

1 Introduction

The problem of binary classification is among the most basic and well-studied problems in statistics and machine learning [22, 24, 3, 1, 13, 2]. Until very recently, theoretical guarantees were almost exclusively formulated in terms of the probability of miss-classification (a.k.a accuracy) as the measure of the risk. This choice of the risk is practically suitable in the case of the "well-balanced" distributions and datasets, that is, the probabilities to observe both classes are similar.

Once this assumption fails to be satisfied, classifiers based on the accuracy might perform poorly in practice. One possible approach to treat such a situation is to modify the measure to be optimized in an appropriate way. A popular choice of such measure is the F-score, whose roots can be tracked back to the information retrieval literature [21, 11]. From the statistical point of view there are two alternative approaches [25, 7] to the theoretical

LAMA, Université Paris-Est
Cité Descartes
5 boulevard Descartes
77454 Marne-la-Vallée cedex 2
E-mail: evgenii.chzhen@univ-paris-est.fr

treatment of the F-score: Population Utility (PU) and Expected Test Utility (ETU). In this work we follow the PU approach which, as noted in [7], has stronger roots in classical statistics. Our goal is to provide minimax analysis of the binary classification with F-score under non-parametric assumptions.

2 The problem formulation

We first introduce some notation that is used throughout this work. For any two real numbers $a, b \in \mathbb{R}$ we denote by $a \wedge b$ (resp. $a \vee b$) the minimum (resp the maximum) between a and b. The standard Euclidean norm in $\mathbb{R}^d$ is denoted by $\|\cdot\|_2$ and a ball centered at $x \in \mathbb{R}^d$ of radius r is denoted by $\mathcal{B}(x, r)$. For positive real valued sequences $a_n, b_n : \mathbb{N} \mapsto \mathbb{R}_+$ we say that $a_n = \mathcal{O}(b_n)$ if there exists some positive constant $M > 0$ such that for all $n \in \mathbb{N}$ it holds that $a_n / b_n \leq M$. We consider a random couple (X, Y) taking values in $\mathbb{R}^d \times \{0, 1\}$ with joint distribution $\mathbb{P}$. The vector $X \in \mathbb{R}^d$ is the feature vector and the binary variable $Y \in \{0, 1\}$ is the label, in what follows we assume that $\mathbb{P}(Y = 1) \neq 0$. We denote by $\mathbb{P}_X$ the marginal distribution of the feature vector $X \in \mathbb{R}^d$ and by $\eta(X) \coloneqq \mathbb{P}(Y = 1|X)$ the regression function. A classifier is any measurable function $g : \mathbb{R}^d \mapsto \{0, 1\}$ and the set of all such functions is denoted by $\mathcal{G}$.

We assume that we have access to two datasets: the first dataset $\mathcal{D}_n = \{(X_i, Y_i)\}_{i=1}^n$ consists of $n \in \mathbb{N}$ *i.i.d.* copies of $(X, Y) \sim \mathbb{P}$; and the second dataset $\mathcal{D}_N = \{X_i\}_{i=n+1}^{n+N}$ consists of $N \in \mathbb{N}$ independent copies of $X \sim \mathbb{P}_X$. Denote by $\mathbb{P}^{\otimes n}$ and $\mathbb{P}_X^{\otimes N}$ the distributions of $\mathcal{D}_n$ and $\mathcal{D}_N$ respectively. Moreover, we denote by $\mathbb{E}_{(\mathcal{D}_n, \mathcal{D}_N)}$ the expectation with respect to the distribution of $(\mathcal{D}_n, \mathcal{D}_N)$, that is, with respect to $\mathbb{P}^{\otimes n} \otimes \mathbb{P}_X^{\otimes N}$ on the space $\big(\mathbb{R}^d \times \{0, 1\}\big)^n \times \big(\mathbb{R}^d\big)^N$. We additionally assume that the size of the unlabeled dataset is not smaller that the size of the labeled dataset[1], that is, $N \geq n$. For a given classifier $g : \mathbb{R}^d \mapsto \{0, 1\}$ we define its F_b-score[2] for any $b > 0$ by

$$F_b(g) \coloneqq \frac{\mathbb{P}(Y = 1, g(X) = 1)}{b^2 \mathbb{P}(Y = 1) + \mathbb{P}(g(X) = 1)} \ .$$

A Bayes-optimal classifier $g^* : \mathbb{R}^d \mapsto \{0, 1\}$ is any classifier that maximizes the F-score over all classifiers $\mathcal{G}$, that is,

$$g^* \in \operatorname*{arg\,max}_{g \in \mathcal{G}} F_b(g) \ .$$

It was established by [26] that a maximizer of the F_1-score can be obtained by comparing the regression function $\eta(X)$ with a threshold $\theta^* \in [0, 1]$. Importantly, this threshold depends explicitly on the distribution $\mathbb{P}$ and can be obtained as unique root of

$$\theta \mapsto \theta \mathbb{P}(Y = 1) - \mathbb{E}(\eta(X) - \theta)_+ \ .$$

One of the contributions of this work is extension of the result of [26, Section 6] for an arbitrary value of $b > 0$.

[1] Note that one can always satisfy this assumption by augmenting $\mathcal{D}_N$ using a portion of $\mathcal{D}_n$ and erasing labels. Typically, in practice it is easier to gather the unlabeled data then labeled, that is why this assumption is rather a formality.

[2] We decided to divide the classical definition of the F_b-score by the factor $1 + b^2$ to simplify the notation, thus, it is sufficient to multiply the obtained results by $1 + b^2$, to recover the results on the classical definition of the F_b-score.

Theorem 1 *A Bayes-optimal classifier g^* can be obtained point-wise for all $x \in \mathbb{R}^d$ as*

$$g^*(x) = \mathbb{1}_{\{\eta(x) > \theta^*\}} \ , \tag{1}$$

where $\theta^* \in [0,1]$ is a threshold which satisfies

$$b^2\theta^*\mathbb{P}(Y=1) = \mathbb{E}(\eta(X) - \theta^*)_+ \ .$$

Moreover, the classifier g^* satisfies $F_b(g^*) = \theta^*$.

The proof can be found in Appendix A Notice that if the optimal threshold $\theta^* \in [0,1]$ is known a priori, the problem of binary classification with the F-score is no harder than the standard settings of binary classification with the accuracy as the measure of performance. As the threshold $\theta^* \in [0,1]$ depends on the distribution $\mathbb{P}$, it could be estimated using data. Theorem 1 allows to obtain a trivial upper bound on the threshold θ^*, indeed, since $\theta^* = F_b(g^*)$ and for any classifier $g \in \mathcal{G}$ the F_b-score is upper bounded by $1/(1+b^2)$ we have $\theta^* \in [0, 1/(1+b^2)]$.

For any classifier $g : \mathbb{R}^d \mapsto \{0,1\}$ we define its excess score as

$$\mathcal{E}_b(g) \coloneqq F_b(g^*) - F_b(g), \quad \text{(excess score)} \ .$$

The excess score is the central object of our analysis and one of our goals is to provide an estimator whose excess score is as small as possible. Using Theorem 1 we can show that the excess score of any classifier $g : \mathbb{R}^d \mapsto \{0,1\}$ can be written in a simple form.

Lemma 1 *Let $g : \mathbb{R}^d \mapsto \{0,1\}$ be any classifier and assume that $\mathbb{P}(Y=1) \neq 0$, then*

$$\mathcal{E}_b(g) = \frac{\mathbb{E}\,|\eta(X) - \theta^*|\,\mathbb{1}_{\{g^*(X) \neq g(X)\}}}{b^2\mathbb{P}(Y=1) + \mathbb{P}(g(X)=1)} \ .$$

In general the Bayes optimal rule is not unique, Theorem 1 only states that one of the optimal classifiers has the form described by its statement. Even though, the function $\theta \mapsto b^2\theta\mathbb{P}(Y=1) - \mathbb{E}(\eta(X)-\theta)_+$ has unique root (see Appendix C for the proof), other thresholds may result in the same Bayes rule. Indeed, consider a simple example with $\eta(x) \equiv 1/2$, $b=1$, then it is easy to see that the solution θ^* of $\theta/2 = (1/2 - \theta)_+$ is exactly $1/3$, and every Bayes optimal classifier predicts one almost surely. Clearly, any threshold $\theta \in [0, 1/2)$ of the regression function η results in the same classifier. Importantly, Lemma 1 and the equality $\arg\max_{g \in \mathcal{G}} F_1(g) = \theta^*$ are valid *only* for the threshold $\theta^* = 1/3$. In this work, we shall always refer to θ^* being the solution of $b^2\theta\mathbb{P}(Y=1) = \mathbb{E}(\eta(X) - \theta)_+$ and we call this threshold as the *optimal* threshold.

Remark 1 For the rest of the paper, we focus our attention only on the value $b = 1$ to simplify the presentation. It will be clear from our arguments that the generalization of the theoretical results of the paper to an arbitrary value $b > 0$ follows straightforwardly from our analysis.

Interestingly, the results above demonstrate that the problem of binary classification with F-score has a lot in common with the standard settings. Indeed, in both cases the Bayes optimal classifier is obtained via thresholding of the regression function and the expression for the excess risk is also similar. Consequently, in this work we address the following questions

Q1.: Is the problem of binary classification with F-score harder than its more known counterpart? In particular, can the minimax analysis of [2] be extended to these settings and what is an optimal algorithm?

Q2.: In view of recent results of [5], we wonder if the introduction of unlabeled dataset can improve classification algorithms in the context of F-score.

Let us point out, that Lemma 1 is crucial for our analysis as it allows to adapt the scheme provided by [2] for the standard setting of the binary classification. Nevertheless, as the threshold $\theta^* \in [0,1]$ is unknown beforehand, this machinery cannot be applied in a straightforward way and some effort is required. In this work, we pose similar assumptions on the distribution $\mathbb{P}$ to the ones used in [2].

Assumption 1 (α-margin assumption) *We say that the distribution $\mathbb{P}$ of the pair $(X,Y) \in \mathbb{R}^d \times \{0,1\}$ satisfies the α-margin assumption if there exist constants $C_0 > 0$, $\delta_0 \in (0, 1/12]$ and $\alpha > 0$ such that for every positive $\delta \le \delta_0$ we have*

$$\mathbb{P}_X(0 < \left|\eta(X) - \theta^*\right| \le \delta) \le C_0 \delta^\alpha \ .$$

The case of "$\alpha = \infty$" is understood in the following manner [13]: there exists a constant $\delta_0 \in (0,1]$ such that

$$\mathbb{P}_X(0 < \left|\eta(X) - \theta^*\right| \le \delta_0) = 0 \ ,$$

typically this is the most advantageous situation for the binary classification, as the regression function η is separated from the optimal threshold θ^*. Assumption 1 specifies the concentration rate of the regression function η around the optimal threshold θ^*. Notice, if Assumption 1 is satisfied, it holds that for all $\delta > 0$

$$\mathbb{P}_X(0 < \left|\eta(X) - \theta^*\right| \le \delta) \le c_0 \delta^\alpha \ ,$$

where $c_0 = C_0 \vee \delta_0^{-\alpha}$. This condition is tightly related to the rate of convergence in the case of the binary classification [2,13]. The classification algorithm that is proposed in this work is based on a direct estimation of the regression function η and the optimal threshold θ^*.

In the sequel, we consider the case of non-parametric estimation, that is we assume that the regression function $\eta : \mathbb{R}^d \mapsto \{0,1\}$ lies in some class of β-smooth functions and the marginal density $\mathbb{P}_X$ of $X \in \mathbb{R}^d$ admits density *w.r.t.* to the Lebesgue measure supported on a well-behaved compact set and uniformly lower- and upper bounded. The exact formal description of these assumptions is given in Section 4.2, where we prove optimality of our rates. As for now, it is sufficient to assume that there exists a good estimator $\hat{\eta}$ based on the labeled set $\mathcal{D}_n$ of the regression function η.

Assumption 2 (Existence of estimator) *There exists an estimator $\hat{\eta}$ based on $\mathcal{D}_n$ which satisfies for all $t > 0$*

$$\mathbb{P}^{\otimes n}(|\hat{\eta}(x) - \eta(x)| \ge t) \le C_1 \exp(-C_2 a_n t^2) \text{ a.s. } \mathbb{P}_X \ ,$$

for some universal constants $C_1, C_2 > 0$ and an increasing sequence $a_n : \mathbb{N} \mapsto \mathbb{R}_+$.

For instance, in the case of β-smooth regression function[3] $\eta : \mathbb{R}^d \mapsto [0,1]$, a typical non-parametric rate is $a_n = n^{2\beta/(2\beta+d)}$ and it can be achieved by the local polynomial estimator, see [2, Theorem 3.2]. Finally, in this work we assume that the probability $\mathbb{P}(Y = 1)$ is lower bounded by some constant which can be arbitrary small but fixed.

[3] Typically, one also need to assume that the marginal distribution $\mathbb{P}_X$ is well behaved, see Section 4.2.

Assumption 3 (Lower bounded $\mathbb{P}(Y=1)$) *We assume that there exists a positive constant p such that $p \leq \mathbb{P}(Y=1)$.*

It is assumed that the constants C_0, C_1, C_2, p are independent of both $n, N \in \mathbb{N}$, however these constants can depend on the dimension of the problem d, on the value of $\alpha > 0$ as well as on each other. The values of the constants C_0, C_1, C_2, p are not going to impact the rates of convergence, though they might and will enter as numerical constants in front of the rate. In contrast, the value of α in the margin assumption will explicitly appear in the obtained rates.

3 Related works and contributions

Literature on the binary classification with F-score is rather broad, it spans both applied and theoretical studies of the problem. It should be noted that our work falls into the Population Utility (PU) approach [7], that is, the expectation is taken in the numerator and the denominator of the F-score simultaneously. This approach should not be confused with the Expected Test Utility (ETU) approach, for which a non-asymptotic behavior can differ significantly. We refer the reader to [7,25] where the PU vs. ETU tale is discussed in depth and their asymptotic equivalency is established. Let us mention that, the asymptotic statistical theory of the binary classification with F-score has been studied in the prior literature [10, 15,14,25]. Bellow, we summarize our contributions and highlight the improvements with respect to the previous results on the non-asymptotic analysis of the binary classification with F-score.

- We propose a two-step estimator, which first estimates the regression function η and then the optimal threshold θ^*. This type of two-step estimators, which involve an explicit thresholds tuning, are well-known in the literature and demonstrate a promising empirical performance [10,9]. An important novelty introduced here is the semi-supervised nature of the procedure which can exploit the unlabeled data. It is already a well established fact that the semi-supervised methods might [18] or not [16] improve supervised estimation from statistical point of view. However, let us point out, that from the practical point of view, typically the most expensive part of the data gathering process is the correct labeling. Thus, one may assume that the unlabeled dataset $\mathcal{D}_N$ is always available in reality and the settings $N \gg n$ are satisfied. Our analysis implies that in the setting of binary classification with F-score the semi-supervised techniques are not superior to the supervised ones. In contrast, in [5] the authors showed that in the context of confidence set classification semi-supervised classifiers might outperform it supervised counterparts.
- From the theoretical point of view, the most relevant reference is a recent work of [23], where the authors studied a rather broad class of performance measures for the problem of binary classification, namely Karmic measures. This class includes the F-score, considered in the present manuscript. Under similar, though stronger assumptions on the distribution[4] of the pair $(X, Y) \in \mathbb{R}^d \times \{0,1\}$ they proposed an algorithm whose rate of convergence is at most $\mathcal{O}(a_n^{-(1+1\wedge\alpha)/2})$. This rate is rather counter intuitive, since it suggests that if the constant α in the margin assumption is large it does not affect the rate of convergence. In contrast, here we show that for the proposed algorithm the rate of convergence is of order $\mathcal{O}(a_n^{-(1+\alpha)/2})$. That is, it strictly improves upon the results

[4] The authors additionally require that the random variable $\eta(X)$ on $[0,1]$ admits bounded density.

in [23] whenever the constant $\alpha > 1$. However, it should be noted, that the authors of [23] study a much more general family of the score functions and the sub-optimal rate can result from such a generality.

– We show that the constructed estimator is optimal in the minimax sense over the class of Hölder smooth regression functions. Let us mention that the optimality of the bound is expected, as in the classical work of [2] the authors showed that the minimax risk in the standard binary classification settings is of order $a_n^{-(1+\alpha)/2}$, and it is achieved by a plug-in rule classifier. Clearly, it is hard to expect that the rate in a more difficult situation can be improved. Nevertheless, to the best of our knowledge, the minimax optimality in the context of binary classification with F-score have not been considered before.

The paper is organized as follows: in Section 4 we present the semi-supervised classification algorithm; in Section 4.1 we establish an upper bound on the excess F-score under the margin assumption; in Section 4.2 we introduce the class of distributions considered in this work and establish a minimax lower bound on the excess F-score.

4 Main results

In this section we describe the proposed procedure $\hat{g}$ to estimate the Bayes optimal classifier g^*, this procedure is performed in two steps. On the first step we estimate the regression function $\eta : \mathbb{R}^d \mapsto \{0, 1\}$ using the labeled data $\mathcal{D}_n$ and on the second step we estimate the optimal threshold θ^* based on the unlabeled data $\mathcal{D}_N$ and the estimator $\hat{\eta}$ provided by the first step. This procedure falls into the category of plug-in type classifiers, that is, we formally replace all the unknown quantities in the Bayes rule by its estimates. That is, the classifier $\hat{g}$ is defined as

$$\hat{g}(x) = \mathbb{1}_{\{\hat{\eta}(x) > \hat{\theta}\}} \ ,$$

where $\hat{\eta}$ is any estimator satisfying Assumption 2 and $\hat{\theta}$ is the unique solution of

$$\theta \frac{1}{N} \sum_{X_i \in \mathcal{D}_N} \hat{\eta}(X_i) = \frac{1}{N} \sum_{X_i \in \mathcal{D}_N} (\hat{\eta}(X_i) - \theta)_+ \ . \tag{2}$$

In practice one can use a simple bisection algorithm [6, Algorithm 3.1] or its more sophisticated modifications (regula falsi or the secant method) to approximate $\hat{\theta}$ with any given precision. For our theoretical analysis we assume that Equation (2) is solved exactly. However a simple modification of our arguments can handle the situation when the threshold $\hat{\theta}$ is known up to an additive factor $\epsilon_n = \mathcal{O}(a_n^{-1/2})$.

4.1 Upper bound

The main result of this subsection is an upper bound on excess score of the proposed procedure. Here we provide two theorems, the first one gives an upper bound on the expected difference between the optimal threshold θ^* and its estimate $\hat{\theta}$. The second one gives an upper bound on the excess F-score.

Theorem 2 *If there exists an estimator $\hat{\eta}$ of the regression function η which satisfies Assumption 2, then there exists a constant $C > 0$ which depends on C_0, C_1, C_2, p such that, the threshold $\hat{\theta}$ defined in Eq.* (2) *satisfies*

$$\mathbb{E}_{(\mathcal{D}_n, \mathcal{D}_N)} |\theta^* - \hat{\theta}| \leq C \left(a_n^{-1/2} + N^{-1/2} \right) \ .$$

Theorem 3 *If the distribution $\mathbb{P}$ of (X, Y) satisfies the α-margin assumption for some $C_0 > 0$ and $\alpha \geq 0$ and there exists an estimator $\hat{\eta}$ of the regression function η which satisfies Assumption 2, then there exists a constant $C > 0$ which depends on α, C_0, C_1, C_2, p such that*

$$\mathbb{E}_{(\mathcal{D}_n, \mathcal{D}_N)} \mathcal{E}_1(\hat{g}) \leq C \left(a_n^{-\frac{1+\alpha}{2}} + N^{-\frac{1+\alpha}{2}} \right) \ ,$$

where $\hat{g}(x) = \mathbb{1}_{\left\{\hat{\eta}(x) > \hat{\theta}\right\}}$ with the threshold $\hat{\theta}$ defined in Equation (2)*.*

Before proceeding to the proofs let us discuss the implications of these results. First of all, there are two regimes in the bound of Theorems 3, the first one is $N \geq a_n$, in this regime, the dominant term is $a_n^{-(1+\alpha)/2}$ which is the classical rate of convergence in the standard settings of binary classification with the α-margin assumption. The second regime is when $N < a_n$, then the dominating term of the bound is $N^{-(1+\alpha)/2}$. However, let us recall that one can always augment the second unlabeled dataset $\mathcal{D}_N$ by dividing $\mathcal{D}_n$ into two independent parts. It implies that the second regime never occurs in our theoretical analysis of the excess score and the upper bound is actually independent of N. Similar reasoning holds for the case of the optimal threshold estimation in Theorem 2. Once it is clear that the obtained upper bounds are actually independent of the size of the unlabeled dataset $\mathcal{D}_N$ it is interesting to notice that the dependence on n is the same as in the standard case of the binary classification [2]. That is, similarly to the standard settings, the binary classification with F-score can achieve fast (faster than $1/\sqrt{n}$) and even super-fast (faster than $1/n$) rate depending on the interplay of α, β, d.

Proofs of both theorems relies on the following lemma, provided in Appendix B, which relates the difference of the thresholds to the difference of the cumulative distribution function empirical of (CDF) η and empirical CDF of $\hat{\eta}$.

Lemma 2 *Let $\hat{\theta} \in [0, 1]$ be the threshold which satisfies Equation 2, then*

$$\left| \hat{\theta} - \theta^* \right| \mathbb{P}(Y = 1) \leq \int_0^1 \left| \mathbb{P}_X(\eta(X) \leq t) - \frac{1}{N} \sum_{X_i \in \mathcal{D}_N} \mathbb{1}_{\{\hat{\eta}(X_i) \leq t\}} \right| dt \ .$$

This result is the main reason why our conclusions on the semi-supervised estimation is different from the ones in [5, 18]. For instance, in [5] the authors also obtain a final decision rule by thresholding on some estimated level. However, in the present work the difference between θ^* and $\hat{\theta}$ is controlled via ℓ_1-norm of difference of CDF's, whereas in [5] they control a similar quantity through Wassertstein infinity distance.

The complete proof of Theorems 2 and 3 can be found in Appendix C, we only sketch the steps which are different from the analysis of [2]. Recall, that due to Lemma 1 we have the following bound for the excess score $\mathcal{E}_1$

$$\mathbb{E}_{(\mathcal{D}_n, \mathcal{D}_N)} \frac{\mathbb{E} \left| \eta(X) - \theta^* \right| \mathbb{1}_{\{g^*(X) \neq \hat{g}(X)\}}}{\mathbb{P}(Y = 1) + \mathbb{P}(\hat{g}(X) = 1)} \leq \frac{1}{p} \mathbb{E}_{(\mathcal{D}_n, \mathcal{D}_N)} \mathbb{E} \left| \eta(X) - \theta^* \right| \mathbb{1}_{\{g^*(X) \neq \hat{g}(X)\}} \ .$$

First of all, notice that if for some $x \in \mathbb{R}^d$ the event $g^*(x) \neq \hat{g}(x)$ occurs, than we have

$$\left|\eta(x) - \theta^*\right| \leq |\eta(x) - \hat{\eta}(x)| + |\theta^* - \hat{\theta}| \ ,$$

which further implies that at least one of the following inequalities hold for this $x \in \mathbb{R}^d$

$$\begin{aligned}
&\left|\eta(x) - \theta^*\right| \leq 2\,|\eta(x) - \hat{\eta}(x)| \ , \\
&\left|\eta(x) - \theta^*\right| \leq 2|\theta^* - \hat{\theta}| \ .
\end{aligned}$$

Thus, we can upper bound the excess risk as

$$\begin{aligned}
\mathcal{E}_1(\hat{g}) \leq & \underbrace{\frac{1}{p}\mathbb{E}\left|\eta(X) - \theta^*\right| \mathbb{1}_{\{|\eta(X)-\theta^*| \leq 2|\eta(X)-\hat{\eta}(X)|\}}}_{T_1} \\
& + \underbrace{\frac{1}{p}\mathbb{E}\left|\eta(X) - \theta^*\right| \mathbb{1}_{\left\{|\eta(X)-\theta^*| \leq 2\left|\theta^*-\hat{\theta}\right|\right\}}}_{T_2} \ .
\end{aligned}$$

The first term on the right hand side (T_1) of the inequality can be handled by the peeling technique used in [2, Lemma 3.1.], which implies that, there exists a constant $C' = C'(p, \alpha, C_0, C_1, C_2) > 0$ such that

$$\mathbb{E}_{(\mathcal{D}_n, \mathcal{D}_N)} T_1 \leq C' a_n^{-\frac{1+\alpha}{2}} \ .$$

Hence, it remains to upper bound the second term on the right hand side (T_2) of the inequality. Using Lemma 2 we can upper bound T_2 as

$$T_2 \leq \frac{1}{p}\mathbb{E}\left|\eta(X) - \theta^*\right| \mathbb{1}_{\{E\}} \ ,$$

with $E = \left\{p\,|\eta(X) - \theta^*| \leq 2\int_0^1 \left|\mathbb{P}_X(\eta(X) \leq t) - \frac{1}{N}\sum_{X_i \in \mathcal{D}_N} \mathbb{1}_{\{\hat{\eta}(X_i) \leq t\}}\right| dt\right\}$. Finally, we upper bound the indicator $\mathbb{1}_{\{E\}}$ by the indicators of two events E^1 and E^2 which are defined as

$$\begin{aligned}
E^1 &= \left\{p\left|\eta(X) - \theta^*\right| \leq 4 \sup_{t \in [0,1]} \left|\mathbb{P}_X(\hat{\eta}(X) \leq t) - \frac{1}{N}\sum_{X_i \in \mathcal{D}_N} \mathbb{1}_{\{\hat{\eta}(X_i) \leq t\}}\right|\right\} \ , \\
E^2 &= \left\{p\left|\eta(X) - \theta^*\right| \leq 4\int_0^1 |\mathbb{P}_X(\hat{\eta}(X) \leq t) - \mathbb{P}_X(\eta(X) \leq t)|\, dt\right\} \ .
\end{aligned}$$

Thus, we have the following upper bound on T_2

$$T_2 \leq \underbrace{\frac{1}{p}\mathbb{E}\left|\eta(X) - \theta^*\right| \mathbb{1}_{\{E^1\}}}_{T_2^1} + \underbrace{\frac{1}{p}\mathbb{E}\left|\eta(X) - \theta^*\right| \mathbb{1}_{\{E^2\}}}_{T_2^2} \ ,$$

Notice that thanks to the Dvoretzky-Kiefer-Wolfowitz inequality [8, 12] the term

$$\sup_{t \in [0,1]} \left|\mathbb{P}_X(\hat{\eta}(X) \leq t) - \frac{1}{N}\sum_{X_i \in \mathcal{D}_N} \mathbb{1}_{\{\hat{\eta}(X_i) \leq t\}}\right| \ ,$$

conditionally on $\mathcal{D}_n$ admits an exponential concentration with the rate $N^{-1/2}$. Hence, using the margin assumption, one can effortlessly show there exists a constant $C'' = C''(p, \alpha, C_0) > 0$ such that

$$\mathbb{E}_{(\mathcal{D}_n, \mathcal{D}_N)} T_2^1 \leq C'' N^{-\frac{1+\alpha}{2}} \quad .$$

For the second term T_2^2 we proceed as follows

$$T_2^2 \leq \frac{4}{p^2} \int_0^1 \left| \mathbb{P}_X(\hat{\eta}(X) \leq t) - \mathbb{P}_X(\eta(X) \leq t) \right| dt \mathbb{P}(E^2) \ ,$$

thus, using the α-margin assumption we get

$$T_2^2 \leq \frac{C_0 4^{1+\alpha}}{p^{2+\alpha}} \left(\int_0^1 \left| \mathbb{P}_X(\hat{\eta}(X) \leq t) - \mathbb{P}_X(\eta(X) \leq t) \right| dt \right)^{1+\alpha} \ ,$$

the integral on the right hand side of the bound corresponds to the 1-Wasserstein distance on the real line, see for instance [4, Theorem 2.9] or [20] for the proof, and can be further upper bounded by the L_1 norm between $\hat{\eta}$ and η, that is

$$T_2^2 \leq \frac{C_0 4^{1+\alpha}}{p^{2+\alpha}} \left(\mathbb{E}_{\mathbb{P}_X} \left| \eta(X) - \hat{\eta}(X) \right| \right)^{1+\alpha} \quad .$$

Since the estimator $\hat{\eta}$ satisfies Assumption 2, one can show that there exists a constant $C''' = C'''(p, \alpha, C_0, C_1, C_2) > 0$ such that

$$\mathbb{E}_{(\mathcal{D}_n, \mathcal{D}_N)} T_2^2 \leq C''' a_n^{-\frac{1+\alpha}{2}} \quad .$$

Combination of all the inequalities yields the result of Theorem 3. Notice that the same reasoning starting from Lemma 2 implies the upper bound on the threshold estimation, that is, Theorem 2.

4.2 Lower bound

In the beginning of the section we state the class of distribution $\mathcal{P}_\Sigma$ of the random pair $(X, Y) \in \mathbb{R}^d \times \{0, 1\}$ considered in this work. The first assumption is made on smoothness of the regression function $\eta : \mathbb{R}^d \mapsto [0, 1]$.

Definition 1 (Hölder smoothness) Let $L > 0$ and $\beta > 0$. The class of function $\Sigma(\beta, L, \mathbb{R}^d)$ consists of all functions $h : \mathbb{R}^d \mapsto [0, 1]$ such that for all $x, x' \in \mathbb{R}^d$, we have

$$\left| h(x) - h_x(x') \right| \leq L \left\| x - x' \right\|_2^\beta \ ,$$

where $h_x(\cdot)$ is the Taylor polynomial of h at point x of degree $\lfloor \beta \rfloor$.

Assumption 4 ((β, L)-Hölder regression function) *The distribution $\mathbb{P}$ of the pair $(X, Y) \in \mathbb{R}^d \times \{0, 1\}$ is such that $\eta \in \Sigma(\beta, L, \mathbb{R}^d)$ for some positive β, L.*

Assumption 4 is usually not sufficient to guarantee the existence of an estimator $\hat{\eta}$ satisfying Assumption 2: extra assumptions are required on the marginal distribution $\mathbb{P}_X$ of the vector $X \in \mathbb{R}^d$.

Definition 2 A Lebesgue measurable set $A \subset \mathbb{R}^d$ is said to be (c_0, r_0)-regular for some constants $c_0 > 0, r_0 > 0$ if for every $x \in A$ and every $r \in (0, r_0]$ we have

$$\lambda\left(A \cap \mathcal{B}(x, r)\right) \geq c_0 \lambda\left(\mathcal{B}(x, r)\right) \ ,$$

where λ is the Lebesgue measure and $\mathcal{B}(x, r)$ is the Euclidean ball of radius r centered at x.

Assumption 5 (Strong density assumption) *We say that the marginal distribution $\mathbb{P}_X$ of the vector $X \in \mathbb{R}^d$ satisfies the strong density assumption if*

- *$\mathbb{P}_X$ is supported on a compact (c_0, r_0)-regular set $A \subset \mathbb{R}^d$,*
- *$\mathbb{P}_X$ admits a density μ* w.r.t. *the Lebesgue measure uniformly lower- and upper-bounded by $\mu_{\min} > 0$ and $\mu_{\max} > 0$ respectively.*

If the regression function $\eta : \mathbb{R}^d \mapsto [0, 1]$ is (β, L)-Hölder and the marginal distribution satisfies the strong density assumption, one can state the following result due to [2].

Theorem 4 ([2]) *Let $\mathcal{P}$ be a class of distributions on $\mathbb{R}^d \times \{0, 1\}$ such that the regression function $\eta \in \Sigma(\beta, L, \mathbb{R}^d)$ and the marginal distribution $\mathbb{P}_X$ satisfies the strong density assumption. Then, there exists an estimator $\hat{\eta}$ of the regression function satisfying*

$$\sup_{\mathbb{P} \in \mathcal{P}} \mathbb{P}^{\otimes n}(|\hat{\eta}(x) - \eta(x)| \geq t) \leq C_1 \exp\left(-C_2 n^{\frac{2\beta}{2\beta + d}} t^2\right) \text{ a.s. } \mathbb{P}_X \ ,$$

for come constants C_1, C_2 depending on β, d, L, c_0, r_0.

Consider a class of distribution $\mathcal{P}_\Sigma$ for which Assumptions 1, 3, 4, 5 are satisfied, then Theorem 4 and Theorems 2, 3 imply the following corollary.

Corollary 1 *There exist constants $C, B > 0$ which depend only on $\alpha, p, d, C_0, C_1, C_2$ such that for any $n > 1, N > 1$ we have*

$$\inf_{\hat{g}} \sup_{\mathbb{P} \in \mathcal{P}_\Sigma} \mathbb{E}_{(\mathcal{D}_n, \mathcal{D}_N)} \mathcal{E}_1(\hat{g}) \leq C n^{-\frac{(1+\alpha)\beta}{2\beta + d}} \ , \tag{3}$$

$$\inf_{\hat{\theta}} \sup_{\mathbb{P} \in \mathcal{P}_\Sigma} \mathbb{E}_{(\mathcal{D}_n, \mathcal{D}_N)} \left|\theta^* - \hat{\theta}\right| \leq B n^{-\frac{\beta}{2\beta + d}} \ . \tag{4}$$

where the infima are taken over all estimators $\hat{g}$ and $\hat{\theta}$ respectively.

The next theorem states that the upper bounds of the previous corollary are optimal up to a constant multiplicative factor.

Theorem 5 *If $\alpha\beta \leq d$, there exists constants $c > 0$ such that for any $n > 1, N > 1$ we have the following lower-bound on the minimax risk*

$$\inf_{\hat{g}} \sup_{\mathbb{P} \in \mathcal{P}_\Sigma} \mathbb{E}_{(\mathcal{D}_n, \mathcal{D}_N)} \mathcal{E}_1(\hat{g}) \geq c n^{-\frac{(1+\alpha)\beta}{2\beta + d}} \ , \tag{5}$$

where the infimum is taken over all estimators $\hat{g}$.

The proof of the lower bound can be found in Appendix D, it follows standard information-theoretic arguments using reduction of the minimax risk to a Bayes risk. The construction of the distributions is inspired by both [17] and [2], and the actual proof relies on [1, Lemma 5.1.], which is based on the Assouad's lemma, see for instance [19, Lemma 2.12].

5 Conclusion

In this work we proposed a semi-supervised plug-in type algorithm for the problem of binary classification with F-score. The proposed algorithm can leverage an unlabeled dataset for the estimation of the optimal threshold. Under the margin assumption it is shown that the proposed algorithm is optimal in the minimax sense and can achieve fast rates of convergence. Further development of the binary classification with F-score will be devoted to empirical risk minimization rules.

Acknowledgements This work was partially supported by "Labex Bézout" of Université Paris-Est. Besides, we would like to thank Joseph Salmon and Mohamed Hebiri for their thoughtful remarks.

A Bayes classifier and Lemma 1

For the rest of this section the parameter $b > 0$ is assumed to be fixed and known. Let us first recall the definition of the F_b-score

$$F_b(g) = (1+b^2)\frac{\mathbb{P}(Y=1, g(X)=1)}{b^2\mathbb{P}(Y=1)+\mathbb{P}(g(X)=1)} \ ,$$

and an optimal classifier is defined as

$$g^* \in \arg\max_{g\in\mathcal{G}} F_b(g) \ .$$

In this section we would like to show that a classifier defined for all $x \in \mathbb{R}^d$ as

$$g_*(x) = \mathbb{1}_{\{\eta(x)\geq\theta^*\}} \ ,$$

with θ^* being a root of

$$\theta \mapsto b^2\mathbb{P}(Y=1)\theta - \mathbb{E}(\eta(X)-\theta)_+ \ .$$

Let us first show that θ^* is well-defined, that is, it exists and is unique for every distribution with $\mathbb{P}(Y=1) \neq 0$. Hence, we would like to study solutions of the following equation

$$b^2\mathbb{P}(Y=1)\theta = \mathbb{E}(\eta(X)-\theta)_+ \ .$$

Clearly, the mapping $\theta \mapsto b^2\mathbb{P}(Y=1)\theta$ is continuous and strictly increasing on $[0,1]$ and the mapping $\theta \mapsto \mathbb{E}(\eta(X)-\theta)_+$ is non-increasing on $[0,1]$. Thus, it is sufficient to demonstrate that the mapping $\theta \mapsto \mathbb{E}(\eta(X)-\theta)_+$ is continuous, indeed, let $\theta, \theta' \in [0,1]$, then, due to the Lipschitz continuity of $(\cdot)_+$ we can write

$$\left|\mathbb{E}(\eta(X)-\theta)_+ - \mathbb{E}(\eta(X)-\theta')_+\right| \leq \mathbb{E}\left|(\eta(X)-\theta)_+ - (\eta(X)-\theta')_+\right| \leq \left|\theta-\theta'\right| \ .$$

This implies that the mapping $\theta \mapsto \mathbb{E}(\eta(X)-\theta)_+$ is a contraction and thus is continuous. Hence, the threshold θ^* is well-defined, that is, it exists and is unique. Consequently, the classifier $x \mapsto \mathbb{1}_{\{\eta(x)\geq\theta^*\}}$ is well-defined.

Now, we are interested in the value $F_b(g_*)$, we can write

$$\begin{aligned}
F_b(g_*) &= (1+b^2)\frac{\mathbb{P}(Y=1, g_*(X)=1)}{b^2\mathbb{P}(Y=1)+\mathbb{P}(g_*(X)=1)} \\
&= (1+b^2)\frac{\mathbb{E}[\eta(X)\mathbb{1}_{\{\eta(X)\geq\theta^*\}}]}{b^2\mathbb{P}(Y=1)+\mathbb{P}(\eta(X)\geq\theta^*)} \\
&= (1+b^2)\frac{\mathbb{E}[(\eta(X)-\theta^*)\mathbb{1}_{\{\eta(X)\geq\theta^*\}}]+\theta^*\mathbb{E}\mathbb{1}_{\{\eta(X)\geq\theta^*\}}}{b^2\mathbb{P}(Y=1)+\mathbb{P}(\eta(X)\geq\theta^*)} \\
&= (1+b^2)\frac{\mathbb{E}(\eta(X)-\theta^*)_+ + \theta^*\mathbb{P}(\eta(X)\geq\theta^*)}{b^2\mathbb{P}(Y=1)+\mathbb{P}(\eta(X)\geq\theta^*)} \ ,
\end{aligned}$$

using the definition of θ^* we continue as

$$F_b(g_*) = (1+b^2)\frac{\mathbb{E}(\eta(X)-\theta^*)_+ + \theta^*\mathbb{P}(\eta(X)\geq\theta^*)}{b^2\mathbb{P}(Y=1)+\mathbb{P}(\eta(X)\geq\theta^*)}$$
$$= (1+b^2)\frac{\theta^* b^2\mathbb{P}(Y=1)+\theta^*\mathbb{P}(\eta(X)\geq\theta^*)}{b^2\mathbb{P}(Y=1)+\mathbb{P}(\eta(X)\geq\theta^*)} = (1+b^2)\theta^* \ .$$

To conclude the optimality of g_* we prove Lemma 1.

Proof Fix an arbitrary measurable function $g:\mathbb{R}^d\mapsto\{0,1\}$, then by the definition of the excess score we have

$$\begin{aligned}
\mathcal{E}_b(g) &\coloneqq \frac{\mathbb{P}(Y=1,g^*(X)=1)}{b^2\mathbb{P}(Y=1)+\mathbb{P}(g^*(X)=1)} - \frac{\mathbb{P}(Y=1,g(X)=1)}{b^2\mathbb{P}(Y=1)+\mathbb{P}(g(X)=1)} \\
&= \frac{\mathbb{E}\eta(X)\mathbb{1}_{\{\eta(X)>\theta^*\}}}{b^2\mathbb{P}(Y=1)+\mathbb{P}(g^*(X)=1)} - \frac{\mathbb{E}\eta(X)\mathbb{1}_{\{g(X)=1\}}}{b^2\mathbb{P}(Y=1)+\mathbb{P}(g(X)=1)} \\
&= \frac{\mathbb{E}\eta(X)\mathbb{1}_{\{\eta(X)>\theta^*\}} - \mathbb{E}\eta(X)\mathbb{1}_{\{g(X)=1\}}}{b^2\mathbb{P}(Y=1)+\mathbb{P}(g^*(X)=1)} \\
&\quad + \frac{\mathbb{E}\eta(X)\mathbb{1}_{\{g(X)=1\}}}{b^2\mathbb{P}(Y=1)+\mathbb{P}(g(X)=1)}\left(\frac{\mathbb{P}(g(X)=1)-\mathbb{P}(g^*(X)=1)}{b^2\mathbb{P}(Y=1)+\mathbb{P}(g^*(X)=1)}\right) \\
&= \frac{\mathbb{E}(\eta(X)-\theta^*)(\mathbb{1}_{\{\eta(X)>\theta^*\}}-\mathbb{1}_{\{g(X)=1\}}) + \theta^*(\mathbb{P}(g^*(X)=1)-\mathbb{P}(g(X)=1))}{b^2\mathbb{P}(Y=1)+\mathbb{P}(g^*(X)=1)} \\
&\quad + F_b(g)\left(\frac{\mathbb{P}(g(X)=1)-\mathbb{P}(g^*(X)=1)}{b^2\mathbb{P}(Y=1)+\mathbb{P}(g^*(X)=1)}\right) \\
&= \frac{\mathbb{E}\,|\eta(X)-\theta^*|\,\mathbb{1}_{\{g^*(X)\neq g(X)\}}}{b^2\mathbb{P}(Y=1)+\mathbb{P}(g^*(X)=1)} + (\theta^*-F_b(g))\frac{\mathbb{P}(g^*(X)=1)-\mathbb{P}(g(X)=1)}{b^2\mathbb{P}(Y=1)+\mathbb{P}(g^*(X)=1)} \ .
\end{aligned}$$

Using Theorem 1 we know that $\theta^* = F_b(g^*)$ and therefore

$$\mathcal{E}(g) = \frac{\mathbb{E}\,|\eta(X)-\theta^*|\,\mathbb{1}_{\{g^*(X)\neq g(X)\}}}{b^2\mathbb{P}(Y=1)+\mathbb{P}(g^*(X)=1)} + \mathcal{E}(g)\frac{\mathbb{P}(g^*(X)=1)-\mathbb{P}(g(X)=1)}{b^2\mathbb{P}(Y=1)+\mathbb{P}(g^*(X)=1)} \ .$$

We conclude by solving the previous equality for $\mathcal{E}(g)$. Thus, g_* is a Bayes optimal classifier and hence can be denoted by g^*.

B Proof of Lemma 2

Proof To prove this lemma, it is convenient to rewrite Equation 2 in terms of CDF. Let μ be an arbitrary probability measure on $\mathbb{R}^d$ and $p:\mathbb{R}^d\mapsto[0,1]$ be any measurable function, then using Fubini's theorem we can write

$$\int p(x)d\mu(x) = \int\int_0^1 \mathbb{1}_{\{p(x)>t\}}dtd\mu(x)$$
$$= \int_0^1 \mu(p(X)>t)dt \ ,$$

and for any $\theta\in[0,1]$, since $(p(x)-\theta)_+\in[0,1]$ we have

$$\int (p(x)-\theta)_+ d\mu(x) = \int\int_0^1 \mathbb{1}_{\{p(x)-\theta>t\}}dtd\mu(x)$$
$$= \int\int_\theta^{1+\theta} \mathbb{1}_{\{p(x)>t\}}dtd\mu(x)$$
$$= \int\int_\theta^1 \mathbb{1}_{\{p(x)>t\}}dtd\mu(x)$$
$$= \int_\theta^1 \mu(p(X)>t)dt \ .$$

Let us denote by $\mathbb{P}_{X,N} = \frac{1}{N}\sum_{X_i \in \mathcal{D}_N} \delta_{X_i}$ the empirical measure of the unlabeled dataset $\mathcal{D}_N$. Using these equalities, the thresholds $\theta^*, \hat{\theta} \in [0,1]$ satisfy

$$\hat{\theta} = \frac{\int_{\hat{\theta}}^1 \mathbb{P}_{X,N}(\hat{\eta}(X) > t)dt}{\int_0^1 \mathbb{P}_{X,N}(\hat{\eta}(X) > t)dt}, \quad \theta^* = \frac{\int_{\theta^*}^1 \mathbb{P}_X(\eta(X) > t)dt}{\int_0^1 \mathbb{P}_X(\eta(X) > t)dt} \ .$$

Now, we are in position to bound the difference $|\hat{\theta} - \theta^*|$, first assume that $\theta^* \geq \hat{\theta}$, then

$$\begin{aligned}
\theta^* - \hat{\theta} &= \frac{\int_{\theta^*}^1 \mathbb{P}_X(\eta(X) > t)dt}{\int_0^1 \mathbb{P}_X(\eta(X) > t)dt} - \frac{\int_{\hat{\theta}}^1 \mathbb{P}_{X,N}(\hat{\eta}(X) > t)dt}{\int_0^1 \mathbb{P}_{X,N}(\hat{\eta}(X) > t)dt} \\
&\leq \frac{\int_{\hat{\theta}}^1 \mathbb{P}_X(\eta(X) > t)dt}{\int_0^1 \mathbb{P}_X(\eta(X) > t)dt} - \frac{\int_{\hat{\theta}}^1 \mathbb{P}_{X,N}(\hat{\eta}(X) > t)dt}{\int_0^1 \mathbb{P}_{X,N}(\hat{\eta}(X) > t)dt} \\
&= \frac{\int_{\hat{\theta}}^1 (\mathbb{P}_X(\eta(X) > t) - \mathbb{P}_{X,N}(\hat{\eta}(X) > t))dt}{\int_0^1 \mathbb{P}_X(\eta(X) > t)dt} \\
&\quad + \frac{\int_{\hat{\theta}}^1 \mathbb{P}_{X,N}(\hat{\eta}(X) > t)dt}{\int_0^1 \mathbb{P}_{X,N}(\hat{\eta}(X) > t)dt} \frac{\int_0^1 (\mathbb{P}_{X,N}(\eta(X) > t) - \mathbb{P}_X(\hat{\eta}(X) > t))dt}{\int_0^1 \mathbb{P}_X(\eta(X) > t)dt} \\
&= \frac{\int_{\hat{\theta}}^1 (\mathbb{P}_X(\eta(X) > t) - \mathbb{P}_{X,N}(\hat{\eta}(X) > t))dt - \hat{\theta}\int_0^1 (\mathbb{P}_X(\hat{\eta}(X) > t) - \mathbb{P}_{X,N}(\eta(X) > t))dt}{\int_0^1 \mathbb{P}_X(\eta(X) > t)dt} \\
&\leq \frac{1}{\mathbb{P}(Y=1)} \int_0^1 \left| \mathbb{P}_X(\eta(X) > t) - \mathbb{P}_{X,N}(\hat{\eta}(X) > t) \right| dt \ .
\end{aligned}$$

Further, if $\hat{\theta} > \theta^*$ we can write

$$\begin{aligned}
\hat{\theta} - \theta^* &= \frac{\int_{\hat{\theta}}^1 \mathbb{P}_{X,N}(\hat{\eta}(X) > t)dt}{\int_0^1 \mathbb{P}_{X,N}(\hat{\eta}(X) > t)dt} - \frac{\int_{\theta^*}^1 \mathbb{P}_X(\eta(X) > t)dt}{\int_0^1 \mathbb{P}_X(\eta(X) > t)dt} \\
&\leq \frac{\int_{\theta^*}^1 \mathbb{P}_{X,N}(\hat{\eta}(X) > t)dt}{\int_0^1 \mathbb{P}_{X,N}(\hat{\eta}(X) > t)dt} - \frac{\int_{\theta^*}^1 \mathbb{P}_X(\eta(X) > t)dt}{\int_0^1 \mathbb{P}_X(\eta(X) > t)dt} \\
&= \frac{\int_{\theta^*}^1 (\mathbb{P}_{X,N}(\hat{\eta}(X) > t) - \mathbb{P}_X(\eta(X) > t))dt}{\int_0^1 \mathbb{P}_X(\eta(X) > t)dt} \\
&\quad + \frac{\int_{\theta^*}^1 \mathbb{P}_{X,N}(\hat{\eta}(X) > t)dt}{\int_0^1 \mathbb{P}_{X,N}(\hat{\eta}(X) > t)dt} \frac{\int_0^1 (\mathbb{P}_X(\eta(X) > t) - \mathbb{P}_{X,N}(\hat{\eta}(X) > t))dt}{\int_0^1 \mathbb{P}_X(\eta(X) > t)dt} \\
&\leq \frac{1}{\mathbb{P}(Y=1)} \int_0^1 \left| \mathbb{P}_X(\eta(X) > t) - \mathbb{P}_{X,N}(\hat{\eta}(X) > t) \right| dt \ ,
\end{aligned}$$

where the last inequality follows the same lines as for the case $\hat{\theta} \leq \theta^*$.

C Proof of the upper bound

Let $\hat{\eta}$ be an estimator of the regression function based on the labeled dataset $\mathcal{D}_n$ which satisfies Assumption 2. Recall, that the estimator $\hat{g}$ is defined for every $x \in \mathbb{R}^d$ as

$$\hat{g}(x) = \mathbb{1}_{\{\hat{\eta}(x) > \hat{\theta}\}} \ ,$$

with $\hat{\theta}$ being the unique solution of Eq. (2). Unless stated otherwise, we work conditionally on $(\mathcal{D}_n, \mathcal{D}_N)$. Using Lemma 1 we can express the excess score of $\hat{g}$ as

$$\mathcal{E}_1(\hat{g}) = \frac{\mathbb{E}\left|\eta(X) - \theta^*\right| \mathbb{1}_{\{g^*(X) \neq \hat{g}(X)\}}}{\mathbb{P}(Y=1) + \mathbb{P}(\hat{g}(X)=1)} \leq \frac{1}{p}\mathbb{E}\left|\eta(X) - \theta^*\right| \mathbb{1}_{\{g^*(X) \neq \hat{g}(X)\}} \ .$$

Clearly, on the event $\{g^*(X) \neq \hat{g}(X)\}$ it holds that $\left\{|\eta(X) - \theta^*| \leq |\hat{\eta}(X) - \eta(X)| + \left|\hat{\theta} - \theta^*\right|\right\}$, thus

$$\begin{aligned}
\mathcal{E}_1(\hat{g}) &\leq \frac{1}{p}\mathbb{E}\,|\eta(X) - \theta^*|\,\mathbb{1}_{\{|\eta(X)-\theta^*|\leq|\hat{\eta}(X)-\eta(X)|+|\hat{\theta}-\theta^*|\}} \\
&\leq \frac{1}{p}\mathbb{E}\,|\eta(X) - \theta^*|\,\mathbb{1}_{\{|\eta(X)-\theta^*|\leq 2|\hat{\theta}-\theta^*|\}} + \frac{1}{p}\mathbb{E}\,|\eta(X) - \theta^*|\,\mathbb{1}_{\{|\eta(X)-\theta^*|\leq 2|\hat{\eta}(X)-\eta(X)|\}}\ .
\end{aligned}$$

Using, Lemma 2 the excess risk can be further upper bounded as

$$\begin{aligned}
\mathcal{E}_1(\hat{g}) \leq& \frac{1}{p}\mathbb{E}\,|\eta(X) - \theta^*|\,\mathbb{1}_{\{|\eta(X)-\theta^*|\leq 2|\hat{\eta}(X)-\eta(X)|\}} \\
&+ \frac{1}{p}\mathbb{E}\,|\eta(X) - \theta^*|\,\mathbb{1}_{\left\{|\eta(X)-\theta^*|\leq \frac{2}{p}\int_0^1 \left|\mathbb{P}_X(\eta(X)\leq t) - \frac{1}{N}\sum_{X_i\in\mathcal{D}_N}\mathbb{1}_{\{\hat{\eta}(X_i)\leq t\}}\right| dt\right\}} \\
\leq& \frac{1}{p}\mathbb{E}\,|\eta(X) - \theta^*|\,\mathbb{1}_{\{|\eta(X)-\theta^*|\leq 2|\hat{\eta}(X)-\eta(X)|\}} \\
&+ \frac{1}{p}\mathbb{E}\,|\eta(X) - \theta^*|\,\mathbb{1}_{\left\{|\eta(X)-\theta^*|\leq \frac{2}{p}\int_0^1 |\mathbb{P}_X(\eta(X)\leq t) - \mathbb{P}_X(\hat{\eta}(X)\leq t)| dt\right\}} \\
&+ \frac{1}{p}\mathbb{E}\,|\eta(X) - \theta^*|\,\mathbb{1}_{\left\{|\eta(X)-\theta^*|\leq \frac{2}{p}\int_0^1 \left|\frac{1}{N}\sum_{X_i\in\mathcal{D}_N}\mathbb{1}_{\{\hat{\eta}(X_i)\leq t\}} - \mathbb{P}_X(\hat{\eta}(X)\leq t)\right| dt\right\}}\ .
\end{aligned}$$

Notice that $\int_0^1 |\mathbb{P}_X(\eta(X)\leq t) - \mathbb{P}_X(\hat{\eta}(X)\leq t)|\,dt = \left\|F_\eta - F_{\hat{\eta}}\right\|_1$, with $F_\eta, F_{\hat{\eta}}$ being the cumulative distribution functions of $\eta, \hat{\eta}$ respectively, corresponds to the 1-Wasserstein distance, see [4] for an in-depth discussion. Therefore, we have

$$\int_0^1 |\mathbb{P}_X(\eta(X)\leq t) - \mathbb{P}_X(\hat{\eta}(X)\leq t)|\,dt \leq \mathbb{E}_{X\sim\mathbb{P}_X}|\eta(X) - \hat{\eta}(X)| := \|\eta - \hat{\eta}\|_1\ ,$$

and introducing notation $\hat{\mathbb{P}}_X := \frac{1}{N}\sum_{X_i\in\mathcal{D}_N}\delta_{X_i}$ for the empirical measure of the feature vector X we can write

$$\begin{aligned}
\mathcal{E}_1(\hat{g}) \leq& \frac{1}{p}\mathbb{E}\,|\eta(X) - \theta^*|\,\mathbb{1}_{\{|\eta(X)-\theta^*|\leq 2|\hat{\eta}(X)-\eta(X)|\}} \\
&+ \frac{1}{p}\mathbb{E}\,|\eta(X) - \theta^*|\,\mathbb{1}_{\left\{|\eta(X)-\theta^*|\leq \frac{2}{p}\|\eta-\hat{\eta}\|_1\right\}} \\
&+ \frac{1}{p}\mathbb{E}\,|\eta(X) - \theta^*|\,\mathbb{1}_{\left\{|\eta(X)-\theta^*|\leq \frac{2}{p}\sup_{t\in[0,1]}|\hat{\mathbb{P}}_X(\hat{\eta}(X)\leq t) - \mathbb{P}_X(\hat{\eta}(X)\leq t)|\right\}}\ .
\end{aligned}$$

Finally, using the margin Assumption 1 we can write

$$\begin{aligned}
\mathcal{E}_1(\hat{g}) \leq& \frac{1}{p}\mathbb{E}\,|\eta(X) - \theta^*|\,\mathbb{1}_{\{|\eta(X)-\theta^*|\leq 2|\hat{\eta}(X)-\eta(X)|\}} \\
&+ \frac{1}{p}\mathbb{E}\,|\eta(X) - \theta^*|\,\mathbb{1}_{\left\{|\eta(X)-\theta^*|\leq \frac{2}{p}\|\eta-\hat{\eta}\|_1\right\}} \\
&+ \frac{1}{p}\mathbb{E}\,|\eta(X) - \theta^*|\,\mathbb{1}_{\left\{|\eta(X)-\theta^*|\leq \frac{2}{p}\sup_{t\in[0,1]}|\hat{\mathbb{P}}_X(\hat{\eta}(X)\leq t) - \mathbb{P}_X(\hat{\eta}(X)\leq t)|\right\}} \\
\leq& \frac{1}{p}\mathbb{E}\,|\eta(X) - \theta^*|\,\mathbb{1}_{\{|\eta(X)-\theta^*|\leq 2|\hat{\eta}(X)-\eta(X)|\}} \\
&+ \frac{2}{p^2}\|\eta - \hat{\eta}\|_1\,\mathbb{P}\left(|\eta(X) - \theta^*| \leq \frac{2}{p}\|\eta - \hat{\eta}\|_1\right) \\
&+ \frac{1}{p}\mathbb{E}\,|\eta(X) - \theta^*|\,\mathbb{1}_{\left\{|\eta(X)-\theta^*|\leq \frac{2}{p}\sup_{t\in[0,1]}|\hat{\mathbb{P}}_X(\hat{\eta}(X)\leq t) - \mathbb{P}_X(\hat{\eta}(X)\leq t)|\right\}} \\
\leq& \frac{1}{p}\mathbb{E}\,|\eta(X) - \theta^*|\,\mathbb{1}_{\{|\eta(X)-\theta^*|\leq 2|\hat{\eta}(X)-\eta(X)|\}} + \frac{2^{\alpha+1}c_0}{p^{2+\alpha}}\|\eta - \hat{\eta}\|_1^{1+\alpha} \\
&+ \frac{1}{p}\mathbb{E}\,|\eta(X) - \theta^*|\,\mathbb{1}_{\left\{|\eta(X)-\theta^*|\leq \frac{2}{p}\sup_{t\in[0,1]}|\hat{\mathbb{P}}_X(\hat{\eta}(X)\leq t) - \mathbb{P}_X(\hat{\eta}(X)\leq t)|\right\}}\ .
\end{aligned}$$

Taking expectation from the both sides with respect to the distribution of $(\mathcal{D}_n, \mathcal{D}_N)$ we follow [2, Lemma 3.1] to bound the first term on the right hand side. This peeling argument became classical in the literature and thus is omitted here. Moreover, using Assumption 2 the second term can be bounded with the same rate as the first term. These arguments would imply that there exists $C \geq 0$ such that for all $n, N \geq 1$ it holds that

$$\begin{aligned}\mathbb{E}_{(\mathcal{D}_n,\mathcal{D}_N)}\mathcal{E}_1(\hat{g}) &\leq C a_n^{-\frac{1+\alpha}{2}} \\ &\quad + \frac{1}{p}\mathbb{E}_{(\mathcal{D}_n,\mathcal{D}_N)}\mathbb{E}\,|\eta(X)-\theta^*|\,\mathbb{1}_{\left\{|\eta(X)-\theta^*|\leq \frac{2}{p}\sup_{t\in[0,1]}\left|\hat{\mathbb{P}}_X(\hat{\eta}(X)\leq t)-\mathbb{P}_X(\hat{\eta}(X)\leq t)\right|\right\}} \\ &\leq C a_n^{-\frac{1+\alpha}{2}} + \frac{2^{\alpha+1}c_0}{p^{2+\alpha}}\mathbb{E}_{(\mathcal{D}_n,\mathcal{D}_N)}\left(\sup_{t\in[0,1]}\left|\hat{\mathbb{P}}_X(\hat{\eta}(X)\leq t)-\mathbb{P}_X(\hat{\eta}(X)\leq t)\right|\right)^{1+\alpha}\end{aligned}$$

It remains to upper bound the second term in the bound above, to this end we recall the classical Dvoretzky-Kiefer-Wolfowitz inequality [12]

Lemma 3 (Dvoretzky-Kiefer-Wolfowitz inequality) *Given $N \geq 0$, let $Z_1, \ldots, Z_N$ be* i.i.d. *real-valued random variables with cumulative distribution function F_Z, denote by $\hat{F}_Z$ the cumulative distribution function with respect to the empirical measure, that is, with respect to $\frac{1}{N}\sum_{i=1}^N \delta_{Z_i}$, then for every $t > 0$ we have*

$$\mathbb{P}\left(\sup_{z\in\mathbb{R}}\left|\hat{F}_Z(z) - F_Z(z)\right| \geq t\right) \leq 2\exp\left(-2Nt^2\right) \quad .$$

Let us apply this lemma to $Z_i := \hat{\eta}(X_i)$, conditionally on $\mathcal{D}_n$ these random variables are *i.i.d.* real-valued, thus for all $t > 0$

$$\mathbb{P}\left(\sup_{t\in[0,1]}\left|\hat{\mathbb{P}}_X(\hat{\eta}(X)\leq t) - \mathbb{P}_X(\hat{\eta}(X)\leq t\right| \geq t\middle|\mathcal{D}_n\right) \leq 2\exp\left(-2Nt^2\right), \quad \text{a.s. } \mathcal{D}_n \quad .$$

Finally, to conclude the upper bound we apply this exponential concentration to upper bound the expectation as

$$\begin{aligned}\mathbb{E}_{\mathcal{D}_n}\mathbb{E}_{\mathcal{D}_N}\left[\left(\Delta_{(\mathcal{D}_N,\mathcal{D}_n)}\right)^{1+\alpha}\middle|\mathcal{D}_n\right] &= \mathbb{E}_{\mathcal{D}_n}\int_0^\infty \mathbb{P}\left(\Delta_{(\mathcal{D}_N,\mathcal{D}_n)} \geq t^{\frac{1}{1+\alpha}}\middle|\mathcal{D}_n\right)dt \\ &\leq \int_0^\infty 2\exp\left(-2Nt^{\frac{2}{1+\alpha}}\right)dt \\ &= N^{-\frac{1+\alpha}{2}}2\int_0^\infty \exp\left(-2t^{\frac{2}{1+\alpha}}\right)dt \\ &\leq CN^{-\frac{1+\alpha}{2}} \quad ,\end{aligned}$$

where we used the shortcut $\Delta_{(\mathcal{D}_N,\mathcal{D}_n)}$ for the desired empirical process. Combining all the bounds we conclude.

D Proof of the lower bound

Proof The proof is similar to the one used in [2] and in [17] and is based on Assouad lemma. Similarly, we define the regular grid on $\mathbb{R}^d$ as

$$G_q := \left\{\left(\frac{2k_1+1}{2q}, \ldots, \frac{2k_d+1}{2q}\right)^\top : k_i \in \{0, \ldots, q-1\}, i = 1, \ldots, d\right\} \quad ,$$

and denote by $n_q(x) \in G_q$ as the closest point to of the grid G_q to the point $x \in \mathbb{R}^d$. Such a grid defines a partition of the unit cube $[0,1]^d \subset \mathbb{R}^d$ denoted by $\mathcal{X}'_1, \ldots, \mathcal{X}'_{q^d}$. Besides, denote by $\mathcal{X}'_{-j} := \{x \in \mathbb{R}^d :$

$-x \in \mathcal{X}'_j\}$ for all $j = 1, \dots, q^d$. For a fixed integer $m \leq q^d$ and for any $j \in \{1, \dots, m\}$ define $\mathcal{X}_i \coloneqq \mathcal{X}'_i$, $\mathcal{X}_{-i} \coloneqq \mathcal{X}'_{-i}$. For every $\sigma \in \{-1, 1\}^m$ we define a regression function η_σ as

$$\eta_\sigma(x) = \begin{cases} \frac{1}{4} + \sigma_j \varphi(x), & \text{if } x \in \mathcal{X}_i \\ \frac{1}{4} - \sigma_j \varphi(x), & \text{if } x \in \mathcal{X}_{-i} \\ \frac{1}{4}, & \text{if } x \in \mathcal{B}(0, \sqrt{d}) \setminus \left(\cup_{i=-m, i\neq 0}^m \mathcal{X}_i \right) \\ \tau, & \text{if } x \in \mathbb{R}^d \setminus \mathcal{B}(0, \sqrt{d} + \rho) \\ \xi(x), & \text{if } x \in \mathcal{B}(0, \sqrt{d} + \rho) \setminus \mathcal{B}(0, \sqrt{d}) \end{cases} ,$$

where ρ, φ, ξ, τ are to be specified and $\mathcal{B}(0, \sqrt{d}+\rho), \mathcal{B}(0, \sqrt{d})$ are Euclidean balls of radius $\sqrt{d}+\rho$ and $\sqrt{d}$ respectively. The definition of the function φ is exactly the same as in [2]. That is, $\varphi \coloneqq C_\varphi q^{-\beta} u(q \left\| x - n_q(x) \right\|_2)$ with some non-increasing infinitely differentiable function such that $u(x) = 1$ for $x \in [0, 1/4]$ and $u(x) = 0$ for $x \geq 1/2$. The function ξ is defined as $\xi(x) = (\tau - 1/4) v([\|x\|_2 - \sqrt{d}]/\rho) + 1/4$, where v is non-decreasing infinitely differentiable function such that $v(x) = 0$ for $x \leq 0$ and $v(x) = 1$ for $x \geq 1$. The constant ρ is chosen big enough to ensure that $|\xi(x) - \xi_x(x')| \leq L \left\| x - x' \right\|_2^\beta$ for any $x, x' \in \mathbb{R}^d$.

For any $\sigma \in \{-1, 1\}^m$ we construct a marginal distribution P_X which is independent of σ and has a density μ *w.r.t.* to the Lebesgue measure on $\mathbb{R}^d$. Fix some $0 < w \leq m^{-1}$ and set A_0 a Euclidean ball in $\mathbb{R}^d$ that has an empty intersection with $\mathcal{B}(0, \sqrt{d} + \rho)$ and whose Lebesgue measure is $\lambda(A_0) = 1 - mq^{-d}$. The density μ is constructed as

- $\mu(x) = \frac{w}{\lambda(\mathcal{B}(0,(4q)^{-1}))}$ for every $z \in G_q$ and every $x \in \mathcal{B}(z, (4q)^{-1}))$ or $x \in \mathcal{B}(-z, (4q)^{-1}))$,
- $\mu(x) = \frac{1-2mw}{\lambda(A_0)}$ for every $x \in A_0$,
- $\mu(x) = 0$ for every other $x \in \mathbb{R}^d$.

To complete the construction it remain to specify the value of $\tau \in [0, 1]$. The idea here is to force the optimal threshold θ^* to be equal to some predefined constant using the additional degree of freedom provided by the parameter τ. Importantly, this optimal threshold should not depend on the binary vector $\sigma \in \{-1, 1\}^m$. To achieve this recall that we set $\theta^* = 1/4$ and show that there exists an appropriate choice of τ. First, recall that the optimal threshold θ^* satisfies

$$\theta^* \mathbb{E}\eta(X) = \mathbb{E}(\eta(X) - \theta^*) \ .$$

Define $b' = \int_{\mathcal{X}_1} \varphi(x)\mu(x)dx / \int_{\mathcal{X}_1} \mu(x)dx$ and put $\theta^* = 1/4$, notice that the left hand side of the last equality for every $\sigma \in \{-1, 1\}^m$ is given by

$$\begin{aligned} \mathbb{E}_\mu \eta_\sigma(X) &= \int_{\mathbb{R}^d} \eta(x) d\mu(x) \\ &= \sum_{j=1}^m \int_{\mathcal{X}_j} (1/4 + \sigma_j \xi(x)) d\mu(x) + \sum_{j=1}^m \int_{\mathcal{X}_{-j}} (1/4 - \sigma_j \xi(x)) d\mu(x) + \int_{A_0} \tau d\mu(x) \\ &= \frac{mw}{2} + \tau(1 - 2mw) \ . \end{aligned}$$

For the right hand side $\mathbb{E}_\mu(\eta_\sigma(X) - 1/4)_+$, there are two cases $\tau > 1/4$ and $0 < \tau \leq 1/4$, one can easily show that as long as $b' \leq 1/8$ there are no values of τ which allow to fix $\theta^* = 1/4$. Therefore, $\tau > 1/4$ and we can write for every $\sigma \in \{-1, 1\}$

$$\begin{aligned} \mathbb{E}_\mu(\eta_\sigma(X) - 1/4)_+ &= \sum_{j=1}^m \int_{\mathcal{X}_j} (\sigma_j \xi(x))_+ d\mu(x) + \sum_{j=1}^m \int_{\mathcal{X}_{-j}} (-\sigma_j \xi(x))_+ d\mu(x) + \int_{A_0} (\tau - 1/4) d\mu(x) \\ &= mwb' + (\tau - 1/4)(1 - 2mw) \ . \end{aligned}$$

Finally, the parameter τ must satisfy the following equality

$$\frac{1}{4}\left(\frac{mw}{2} + \tau(1 - 2mw) \right) = mwb' + (\tau - 1/4)(1 - 2mw) \ ,$$

solving for τ we get

$$\tau = \frac{1}{3} + \left(\frac{1}{12} - \frac{2b'}{3}\right)\left(\frac{2mw}{1-2mw}\right) \ .$$

If $mw \leq 1/2$ we can ensure that the value of $\tau \leq 1$, that is, it is a valid choice for the regression function. Let us demonstrate that the margin assumption 1 holds for an appropriate choice of m and w. Define $x_0 = (1/2q, \dots, 1/2q)^\top$, then for every $\sigma \in \{-1, 1\}$ we have

$$\begin{aligned} P_X(0 < |\eta_\sigma(X) - 1/4| \leq \delta) &= \frac{2mw}{\lambda(\mathcal{B}(0,(4q)^{-1}))} \int_{\mathcal{B}(x_0,(4q)^{-1})} \mathbb{1}_{\left\{C_\varphi q^{-\beta} u(q\|x-n_q(x)\|_2) \leq \delta\right\}} dx \\ &\quad + \frac{1-2mw}{\lambda(A_0)} \int_{A_0} \mathbb{1}_{\left\{\frac{1}{3} + \left(\frac{1}{12} - \frac{2b'}{3}\right)\left(\frac{2mw}{1-2mw}\right) - \frac{1}{4} \leq \delta\right\}} dx \\ &= 2mw \mathbb{1}_{\{\delta \geq C_\varphi q^{-\beta}\}} + \frac{1-2mw}{\lambda(A_0)} \int_{A_0} \mathbb{1}_{\left\{\frac{1}{12} + \left(\frac{1}{12} - \frac{2b'}{3}\right)\left(\frac{2mw}{1-2mw}\right) \leq \delta\right\}} dx \ , \end{aligned}$$

as long as $b' \leq 3/24$ we can continue as

$$\begin{aligned} P_X(0 < |\eta_\sigma(X) - 1/4| \leq \delta) &\leq 2mw \mathbb{1}_{\{\delta \geq C_\varphi q^{-\beta}\}} + \mathbb{1}_{\{\delta \geq \frac{1}{12}\}} \\ &\leq 2mw \mathbb{1}_{\{\delta \geq C_\varphi q^{-\beta}\}} + 12^\alpha \delta^\alpha \ . \end{aligned}$$

Therefore, if mw is of order $q^{-\alpha\beta}$ the margin assumption is satisfied with $\delta_0 = 1/12$. The strong density assumption can be checked similarly to [2]. To finish the prove, for every $\sigma \in \{-1, 1\}^m$ we denote by P^σ the distribution of (X, Y) with the marginal P_X and the regression function η^σ. Thus, one can write for any $\hat{g}$

$$\sup_{\mathbb{P} \in \mathcal{P}_\Sigma} \mathbb{E}_{(\mathcal{D}_n, \mathcal{D}_N)} \mathcal{E}(\hat{g}) \geq \sup_{\sigma \in \{-1,1\}^m} \frac{1}{2} \mathbb{E}^\sigma_{(\mathcal{D}_n, \mathcal{D}_N)} \sum_{i=-m, i\neq 0}^{m} \mathbb{E}_{P_X} |\varphi(X)| \, \mathbb{1}_{\{(1+\mathrm{sign}(i)\sigma_i)/2 \neq \hat{g}(X)\}} \mathbb{1}_{\{X \in \mathcal{X}_i\}} \ ,$$

where $\mathbb{E}^\sigma_{(\mathcal{D}_n, \mathcal{D}_N)}$ is the expectation taken *w.r.t.* to the *i.i.d.* realizations of $\mathcal{D}_n$ and $\mathcal{D}_N$ from P^σ and P_X respectively, and $\mathrm{sign}(i) = 1$ if $i > 0$ and $\mathrm{sign}(i) = -1$ if $i < 0$. The rest of the proof is obtained following the proof of [1, Lemma 5.1.] and in particular the chain of inequalities in [1, Eq. (6.26)]. That is, we get for some $C > 0$ independent from N, n

$$\sup_{\mathbb{P} \in \mathcal{P}_\Sigma} \mathbb{E}_{(\mathcal{D}_n, \mathcal{D}_N)} \mathcal{E}(\hat{g}) \geq Cmwq^{-\beta}(1 - C_\varphi q^{-\beta}\sqrt{nw})$$

Finally, we conclude by setting the parameters m, w, q as

$$q = \lfloor \bar{C} n^{\frac{1}{2\beta+d}} \rfloor, \quad w = C' q^{-d}, \quad m = \lfloor C'' q^{d-\alpha\beta} \rfloor \ .$$

Note that thanks to the condition $\alpha\beta \leq d$ such a choice is is always valid for appropriately chosen constants $\bar{C}, C', C''$.

References

1. Audibert, J.Y.: Aggregated estimators and empirical complexity for least square regression. Ann. Inst. H. Poincaré Probab. Statist. **40**(6), 685–736 (2004)
2. Audibert, J.Y., Tsybakov, A.B.: Fast learning rates for plug-in classifiers. Ann. Statist. **35**(2), 608–633 (2007)
3. Bartlett, P.L., Mendelson, S.: Rademacher and Gaussian complexities: risk bounds and structural results. J. Mach. Learn. Res. **3**(Spec. Issue Comput. Learn. Theory), 463–482 (2002)
4. Bobkov, S., Ledoux, M.: One-dimensional empirical measures, order statistics and Kantorovich transport distances (2016). To appear in the Memoirs of the Amer. Math. Soc.
5. Chzhen, E., Denis, C., Hebiri, M.: Minimax semi-supervised confidence sets for multi-class classification (2019). Preprint, https://arxiv.org/abs/1904.12527

6. Conte, S., Boor, C.: Elementary Numerical Analysis: An Algorithmic Approach, 3rd edn. McGraw-Hill Higher Education (1980)
7. Dembczynski, K., Kotłowski, W., Koyejo, O., Natarajan, N.: Consistency analysis for binary classification revisited. In: ICML, pp. 961–969. JMLR. org (2017)
8. Dvoretzky, A., Kiefer, J., Wolfowitz, J.: Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator. Ann. Math. Statist. **27**(3), 642–669 (1956)
9. Keerthi, S., Sindhwani, V., Chapelle, O.: An efficient method for gradient-based adaptation of hyperparameters in svm models. In: NIPS, pp. 673–680 (2007)
10. Koyejo, O., Natarajan, N., Ravikumar, P., Dhillon, I.: Consistent binary classification with generalized performance metrics. In: NIPS, pp. 2744–2752 (2014)
11. Lewis, D.: Evaluating and optimizing autonomous text classification systems. In: ACM, pp. 246–254. ACM Press (1995)
12. Massart, P.: The tight constant in the dvoretzky-kiefer-wolfowitz inequality. Ann. Probab. **18**(3), 1269–1283 (1990)
13. Massart, P., Nédélec, É.: Risk bounds for statistical learning. Ann. Statist. **34**(5), 2326–2366 (2006)
14. Menon, A., Narasimhan, H., Agarwal, S., Chawla, S.: On the statistical consistency of algorithms for binary classification under class imbalance. In: ICML, vol. 28, pp. 603–611. PMLR (2013)
15. Narasimhan, H., Vaish, R., Agarwal, S.: On the statistical consistency of plug-in classifiers for non-decomposable performance measures. In: NIPS, pp. 1493–1501 (2014)
16. Rigollet, P.: Generalization error bounds in semi-supervised classification under the cluster assumption. Journal of Machine Learning Research **8**(Jul), 1369–1392 (2007)
17. Rigollet, P., Vert, R.: Optimal rates for plug-in estimators of density level sets. Bernoulli (2009)
18. Singh, A., Nowak, R., Zhu, J.: Unlabeled data: Now it helps, now it doesn't. In: NIPS, pp. 1513–1520 (2009)
19. Tsybakov, A.B.: Introduction to nonparametric estimation. Springer Series in Statistics. Springer, New York (2009)
20. Vallender, S.: Calculation of the wasserstein distance between probability distributions on the line. Theory of Probability & Its Applications **18**(4), 784–786 (1974)
21. van Rijsbergen, C.: Foundation of evaluation. Journal of documentation **30**(4), 365–373 (1974)
22. Vapnik, V.N.: Statistical learning theory. Wiley (1998)
23. Yan, B., Koyejo, S., Zhong, K., Ravikumar, P.: Binary classification with karmic, threshold-quasi-concave metrics. In: ICML, vol. 80. PMLR (2018)
24. Yang, Y.: Minimax nonparametric classification: Rates of convergence. IEEE Transactions on Information Theory **45**(7), 2271–2284 (1999)
25. Ye, N., Chai, K., Lee, W., Chieu, H.: Optimizing f-measures: A tale of two approaches. In: ICML (2012)
26. Zhao, M.J., Edakunni, N., Pocock, A., Brown, G.: Beyond fano's inequality: bounds on the optimal f-score, ber, and cost-sensitive risk and their implications. JMLR **14**(Apr), 1033–1090 (2013)