ISSN 1066-5307, Mathematical Methods of Statistics, 2020, Vol. 29, No. 2, pp. 87–105. © Allerton Press, Inc., 2020.

Optimal Rates for Nonparametric F-Score Binary Classification via Post-Processing

Evgenii Chzhen[*]

LMO, Université Paris-Saclay, CNRS, Inria

Received October 12, 2020; revised December 21, 2020; accepted February 2, 2021

Abstract—This work studies the problem of binary classification with the F-score as the performance measure. We propose a post-processing algorithm for this problem which fits a threshold for any score base classifier to yield high F-score. The post-processing step involves only unlabeled data and can be performed in logarithmic time. We derive a general finite sample post-processing bound for the proposed procedure and show that the procedure is minimax rate optimal, when the underlying distribution satisfies classical nonparametric assumptions. This result improves upon previously known rates for the F-score classification and bridges the gap between standard classification risk and the F-score. Finally, we discuss the generalization of this approach to the set-valued classification.

DOI: 10.3103/S1066530720020027

Keywords: F-score, universal consistency, minimax rate, margin assumption, nonparametric classification

1. THE PROBLEM

Consider a random couple $(\mathbf{X}, Y)$ taking values in $\mathbb{R}^d \times \{0, 1\}$ with joint distribution $\mathbb{P}$. The vector $\mathbf{X} \in \mathbb{R}^d$ is the feature vector and the binary variable $Y \in \{0, 1\}$ is the label. Denote by $\mathbb{P}_{\mathbf{X}}$ the marginal distribution of the feature vector $\mathbf{X} \in \mathbb{R}^d$ and by $\eta(\mathbf{X}) := \mathbb{E}[Y|\mathbf{X}]$ 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 $\mathcal{G}$.

A common way to measure the risk of a classifier g is via its misclassification error $\mathbb{P}(Y \neq g(\mathbf{X}))$. However, once the relation $\mathbb{P}(Y = 1) \approx \mathbb{P}(Y = 0)$ fails to be satisfied, practitioners often seek for a balance between precision and recall of the constructed classifier [14], which are defined as

$$\mathrm{Pr}(g) = \mathbb{P}(Y = 1|g(\mathbf{X}) = 1), \quad \mathrm{Re}(g) = \mathbb{P}(g(\mathbf{X}) = 1|Y = 1),$$

respectively. Ultimately, a desired classifier maximizes both quantities simultaneously or seeks for some trade-off among them. Aggregates which combine both measures into one are often employed in practical applications [1]. Among others, a popular choice of such an aggregate is the F-score which is defined for a fixed parameter of choice $b > 0$ as

$$\mathrm{F}_b(g) = \left(\frac{1}{1 + b^2}\Big(\mathrm{Pr}(g)\Big)^{-1} + \frac{b^2}{1 + b^2}\Big(\mathrm{Re}(g)\Big)^{-1}\right)^{-1}. \tag{1}$$

In other words, the F-score with the parameter b is the weighted harmonic average of the precision and the recall.[1)] In practice common choices of parameters are $b = 2; 0.5; 1$, which puts more emphasis on recall ($b = 2$), precision ($b = 0.5$), and treats both equally ($b = 1$). Our goal is to build a data-driven post-processing procedure, whose expected F-score is as high as possible. In the paradigm of post-processing we assume that an initial estimator $\hat{\eta}$ of the regression function η is available and is constructed using an i.i.d. sample $(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)$ drawn from $\mathbb{P}$ and independent from $(\mathbf{X}, Y)$. A post-processing procedure is any procedure which receives the initial estimator $\hat{\eta}$ plus an additional

[*] E-mail: evgenii.chzhen@universite-paris-saclay.fr

[1)] The (weighted) harmonic mean of any real number and zero is defined as zero.

sample and outputs a classifier $\hat{g} : \mathbb{R}^d \to \{0,1\}$. In our case, this additional sample is *unlabeled*. It consists of $\mathbf{X}_{n+1}, \ldots, \mathbf{X}_{n+N}$ sampled[2)] i.i.d. from $\mathbb{P}_{\mathbf{X}}$. We define an F_b-score optimal classifier $g^* : \mathbb{R}^d \mapsto \{0,1\}$ as a solution of

$$\max_{g \in \mathcal{G}} \left(\frac{1}{1+b^2} \Big(\mathrm{Pr}(g) \Big)^{-1} + \frac{b^2}{1+b^2} \Big(\mathrm{Re}(g) \Big)^{-1} \right)^{-1} .$$

Consequently, the theoretical performance of any classifier $g : \mathbb{R}^d \to \{0,1\}$ is measured by its excess score $\mathcal{E}$ defined as

$$\mathcal{E}_b(g) := \mathrm{F}_b(g^*) - \mathrm{F}_b(g),$$

which is the difference of the optimal possible value and the F_b-score of g.

Related works and contributions. In practical applications the classification power is often measured by the precision and the recall of a classifier [12, 14, 16]. In particular, practitioners consider more complex measures, which trade-off the two conflicting quantities [see for instance 16, Sections 3.5 and 3.5.4]. A popular choice of such measure is the F_b-score—a weighted harmonic mean of the precision and the recall, which is often used when $\mathbb{P}(Y=1) < \mathbb{P}(Y=0)$ to discover rare positive labels [9]. For example, the F_b-score is reported as one of the success criteria is several applied and algorithmic works [4, 9, 15, 22, 24, 33].

Until very recently, the theoretical study of the binary classification was almost exclusively focused on the classical misclassification risk or its convex surrogates. More recently, statistically grounded post-processing estimation procedures under the F_b-score classification received an increasing amount of attention over the recent years [3, 8, 17, 25, 37, 38]. However, most of the previous works were almost exclusively focused on the asymptotic results, with a notable exception of [38], where they establish finite sample results for their procedure.

In this work we propose and analyze yet another post-processing algorithm (see Section 3), which calibrates any score based classifier to maximize the F_b-score. In particular, in Section 3.1 we derive a general post-processing bound, which holds without any assumptions on the distribution. This result shows that our algorithm is universally consistent, recovering similar guarantees as that of [17, 25]. Meanwhile, unlike the algorithms of [17, 25], the proposed post-processing step is performed using only *unlabeled* data and it requires only *logarithmic* time in the unlabeled data, making it attractive for applications where the cost of labels is high and the computational powers are limited.

In Section 4 we consider the scenario of nonparametric classification and establish an upper bound for the proposed procedure, moreover, our lower bound demonstrates that this rate is minimax optimal. In standard nonparametric classification, a classical result of Audibert and Tsybakov [2] establishes that the optimal rate of convergence of the misclassification risk is $n^{-(1+\alpha)\beta/(2\beta+d)}$, where α is the margin parameter [34], β is the smoothness of the conditional distribution of the label; and d is the dimension of the feature space. Recently, under stronger assumptions, Yan et al. [38] proposed a post-processing algorithm for the F_b-score maximization whose rate is $n^{-(1+\min\{\alpha,1\}\beta)/(2\beta+d)}$, that is, it is slower than that of misclassification risk. Our result shows that the rate of Yan et al. [38] is suboptimal. In particular, under milder assumptions our algorithm achieves $n^{-(1+\alpha)\beta/(2\beta+d)}$ rate of convergence for the excess F_b-score.

Finally, in Section 5 we show that our approach can be generalized to the set-valued classification framework [10, 30] using an extension of the F_b-score to this setup [7, 23].

Notation. We make use of the following notation. For any $x \in \mathbb{R}$ we denote by $(x)_+ = \max\{x, 0\}$ its positive part. For two real numbers a, b we denote by $a \vee b$ (resp., $a \wedge b$) the maximum (resp., the minimum) between a, b. In this work the symbols $\mathbf{P}$ and $\mathbf{E}$ stand for a generic probability and expectation respectively, while $\mathbb{P}$ and $\mathbb{P}_{\mathbf{X}}$ are the distributions of $(\mathbf{X}, Y)$ and $\mathbf{X}$, respectively. For any function $f : \mathbb{R}^d \to \mathbb{R}$ we set $||f||_p = (\int_{\mathbb{R}^d} f(x) d\mathbb{P}_{\mathbf{X}}(x))^{1/p}$. By the abuse of notation for any $x \in \mathbb{R}^d$ we denote by $||\mathbf{x}||_2$ its Euclidean norm. Finally, for any $x \in \mathbb{R}^d$ we denote by $||\mathbf{x}||_\infty$ the sup norm of x.

[2)] It is assumed that $\mathbf{X}_{n+1}, \ldots, \mathbf{X}_{n+N}$ are independent from $(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n), (\mathbf{X}, Y)$.

2. REMINDER ON THE F_b-SCORE

Zhao et al. [40] demonstrated that a maximizer of the F_1-score can be obtained by comparing the regression function $\eta(\mathbf{X})$ with a threshold $\theta^*\in [0, 1]$. An important consequence of their analysis, is the fact that this threshold θ^* depends on the whole joint distribution $\mathbb{P}$. This fact is the crucial distinction between the classical setup with misclassification risk, where the threshold is known beforehand and is equal to $1/2$ [11].

Theorem 1 (Zhao et al. [40]). *Assume that $\mathbb{P}(Y = 1) \neq 0$ and let $\theta^* \in [0, 1]$ be a solution of the following equation*

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

Let $g^* : \mathbb{R}^d \to \{0, 1\}$ be defined for all $x \in \mathbb{R}^d$ as

$$g^*(x) = \mathbb{I}\{\{\{\eta(x) > \theta^*\}\}. \tag{3}$$

Then, it holds that

1. *θ^* exists and unique,*

2. $\mathrm{F}_b(g^*) = \theta^*(1 + b^2)$,

3. *g^* is a F_b-score optimal classifier.*

Zhao et al. [40] derived the form of the F_b-score optimal classifier for $b = 1$ and the extension of their proof to other values of $b > 0$ trivially follows from their analysis. For the sake of completeness we provide a complete and an *alternative* proof of this result. This particular proof strategy plays a crucial role in our consequent statistical analysis.

Proof of Theorem 1. In Lemma 13 of Section 7 it is shown that θ^*, defined by Eq. (2) exists and is unique for any $b > 0$. Let us show that if θ^* satisfies Eq. (2), then it holds that $\mathrm{F}_b(g^*) = \theta^*(1 + b^2)$ for $g^*(x) = \mathbb{I}\{\{\{\eta(x) > \theta^*\}\}\}$. Simple computations imply that the expression for the F_b-score in Eq. (1) can be written as

$$\mathrm{F}_b(g) = \frac{(1 + b^2)\mathbb{P}(Y = 1, g(\mathbf{X}) = 1)}{b^2\mathbb{P}(Y = 1) + \mathbb{P}(g(\mathbf{X}) = 1)}.$$

Using this expression and the definition of g^* we can write

$$\begin{aligned}\mathrm{F}_b(g^*) = (1 + b^2)\frac{\mathbb{P}(Y = 1, g^*(\mathbf{X}) = 1)}{b^2\mathbb{P}(Y = 1) + \mathbb{P}(g^*(\mathbf{X}) = 1)} &= (1 + b^2)\frac{\mathbb{E}[\eta(\mathbf{X})\mathbb{I}\{\{\eta(\mathbf{X}) \geq \theta^*\}]}{b^2\mathbb{P}(Y = 1) + \mathbb{P}(\eta(\mathbf{X}) \geq \theta^*)}\\ &= (1 + b^2)\frac{\mathbb{E}[(\eta(\mathbf{X}) - \theta^*)\mathbb{I}\{\{\eta(\mathbf{X}) \geq \theta^*\}] + \theta^*\mathbb{E}\mathbb{I}\{\eta(\mathbf{X}) \geq \theta^*\}}{b^2\mathbb{P}(Y = 1) + \mathbb{P}(\eta(\mathbf{X}) \geq \theta^*)}\\ &= (1 + b^2)\frac{\mathbb{E}(\eta(\mathbf{X}) - \theta^*)_+ + \theta^*\mathbb{P}(\eta(\mathbf{X}) \geq \theta^*)}{b^2\mathbb{P}(Y = 1) + \mathbb{P}(\eta(\mathbf{X}) \geq \theta^*)}.\end{aligned}$$

Finally, thanks to the definition of θ^* we obtain

$$\mathrm{F}_b(g^*) = (1 + b^2)\frac{\theta^*b^2\mathbb{P}(Y = 1) + \theta^*\mathbb{P}(\eta(\mathbf{X}) \geq \theta^*)}{b^2\mathbb{P}(Y = 1) + \mathbb{P}(\eta(\mathbf{X}) \geq \theta^*)} = (1 + b^2)\theta^*.$$

Hence $\mathrm{F}_b(g^*) = (1 + b^2)\theta^*$, the optimality of g^* follows from Lemma 2 below. □

Lemma 2. *Let $g : \mathbb{R}^d \mapsto \{0, 1\}$ be any classifier and assume that $\mathbb{P}(Y = 1) \neq 0$, then*

$$\mathrm{F}_b(g^*) - \mathrm{F}_b(g) = \frac{(1 + b^2)\mathbb{E}|\eta(\mathbf{X}) - \theta^*|\mathbb{I}\{g^*(\mathbf{X}) \neq g(\mathbf{X})\}}{b^2\mathbb{P}(Y = 1) + \mathbb{P}(g(\mathbf{X}) = 1)},$$

where g^* is defined in Theorem 1.

Proof. Fix any classifier $g : \mathbb{R}^d \mapsto \{0, 1\}$. The excess score of g can be expressed as

$$\frac{\mathcal{E}_b(g)}{1 + b^2} = \frac{\mathbb{P}(Y = 1, g^*(\mathbf{X}) = 1)}{b^2\mathbb{P}(Y = 1) + \mathbb{P}(g^*(\mathbf{X}) = 1)} - \frac{\mathbb{P}(Y = 1, g(\mathbf{X}) = 1)}{b^2\mathbb{P}(Y = 1) + \mathbb{P}(g(\mathbf{X}) = 1)}$$

$$= \frac{\mathbb{E}[\eta(\mathbf{X})\mathbb{I}\{\eta(\mathbf{X}) > \theta^*\}]}{b^2\mathbb{P}(Y=1) + \mathbb{P}(g^*(\mathbf{X})=1)} - \frac{\mathbb{E}[\eta(\mathbf{X})\mathbb{I}\{g(\mathbf{X})=1]\}}{b^2\mathbb{P}(Y=1) + \mathbb{P}(g(\mathbf{X})=1)}.$$

Adding and subtracting $\frac{\mathbb{E}[\eta(\mathbf{X})\mathbb{I}\{g(\mathbf{X})=1]\}}{b^2\mathbb{P}(Y=1)+\mathbb{P}(g^*(\mathbf{X})=1)}$ on the right hand side we derive after rearranging that

$$\frac{\mathcal{E}_b(g)}{1+b^2} = \frac{\mathbb{E}[\eta(\mathbf{X})\mathbb{I}\{\eta(\mathbf{X}) > \theta^*\}] - \mathbb{E}[\eta(\mathbf{X})\mathbb{I}\{g(\mathbf{X})=1]\}}{b^2\mathbb{P}(Y=1) + \mathbb{P}(g^*(\mathbf{X})=1)}$$
$$+ \frac{\mathbb{E}[\eta(\mathbf{X})\mathbb{I}\{g(\mathbf{X})=1]\}}{b^2\mathbb{P}(Y=1) + \mathbb{P}(g(\mathbf{X})=1)}\left(\frac{\mathbb{P}(g(\mathbf{X})=1) - \mathbb{P}(g^*(\mathbf{X})=1)}{b^2\mathbb{P}(Y=1)+\mathbb{P}(g^*(\mathbf{X})=1)}\right).$$

Notice that $\mathbb{E}[\eta(\mathbf{X})\mathbb{I}\{\eta(\mathbf{X}) > \theta^*\}] - \mathbb{E}[\eta(\mathbf{X})\mathbb{I}\{g(\mathbf{X})=1\}] = \mathbb{E}[(\eta(\mathbf{X}) - \theta^*)(\mathbb{I}\{\eta(\mathbf{X}) > \theta^*\} - \mathbb{I}\{g(\mathbf{X})=1\})] + \theta^*(\mathbb{P}(g^*(\mathbf{X})=1) - \mathbb{P}(g(\mathbf{X})=1))$, therefore

$$\frac{\mathcal{E}_b(g)}{1+b^2} = \frac{\mathbb{E}[(\eta(\mathbf{X}) - \theta^*)(\mathbb{I}\{\eta(\mathbf{X}) > \theta^*\} - \mathbb{I}\{g(\mathbf{X})=1\})] + \theta^*(\mathbb{P}(g^*(\mathbf{X})=1) - \mathbb{P}(g(\mathbf{X})=1))}{b^2\mathbb{P}(Y=1)+\mathbb{P}(g^*(\mathbf{X})=1)}$$
$$+ \frac{\mathrm{F}_b(g)}{1+b^2}\left(\frac{\mathbb{P}(g(\mathbf{X})=1) - \mathbb{P}(g^*(\mathbf{X})=1)}{b^2\mathbb{P}(Y=1)+\mathbb{P}(g^*(\mathbf{X})=1)}\right)$$
$$= \frac{\mathbb{E}|\eta(\mathbf{X}) - \theta^*|\mathbb{I}\{g^*(\mathbf{X}) \neq g(\mathbf{X})\}}{b^2\mathbb{P}(Y=1)+\mathbb{P}(g^*(\mathbf{X})=1)} + \left(\theta^* - \frac{\mathrm{F}_b(g)}{1+b^2}\right)\frac{\mathbb{P}(g^*(\mathbf{X})=1) - \mathbb{P}(g(\mathbf{X})=1)}{b^2\mathbb{P}(Y=1)+\mathbb{P}(g^*(\mathbf{X})=1)}.$$

Thanks to Theorem 1 we know that $\theta^* = \frac{\mathrm{F}_b(g^*)}{1+b^2}$ and hence previous equality can be equivalently expressed as

$$\frac{\mathcal{E}_b(g)}{1+b^2} = \frac{\mathbb{E}|\eta(\mathbf{X}) - \theta^*|\mathbb{I}\{g^*(\mathbf{X}) \neq g(\mathbf{X})\}}{b^2\mathbb{P}(Y=1)+\mathbb{P}(g^*(\mathbf{X})=1)} + \frac{\mathcal{E}_b(g)}{1+b^2}\cdot\frac{\mathbb{P}(g^*(\mathbf{X})=1) - \mathbb{P}(g(\mathbf{X})=1)}{b^2\mathbb{P}(Y=1)+\mathbb{P}(g^*(\mathbf{X})=1)}. \tag{4}$$

We conclude by solving the previous equality for $\mathcal{E}_b(g)$. □

Lemma 2 is the main advantage of our proof over the previously available one. It is a cornerstone in our statistical analysis, allowing us to derive both finite-sample and asymptotic results. Following Theorem 1, in this work we call θ^* the *optimal* threshold. Let us point out that Theorem 1 allows to obtain a trivial upper bound on the optimal threshold, narrowing the search region. Indeed, since $\theta^*(1+b^2) = \mathrm{F}_b(g^*)$ and for any classifier $g \in \mathcal{G}$ we have $\mathrm{F}_b(g) \leq 1$ (since the F_b-score is the harmonic average of the precision and the recall), then it holds that $\theta^* \in [0, 1/(1+b^2)]$. Thanks to the expression on the excess score in Lemma 2 we can observe that if the optimal threshold is known a priori, the problem of binary classification with the F-score is identical to the standard setting of binary classification with the cost-sensitive misclassification risk [31]. Indeed, notice that the optimal classifier g^* minimizes

$$\mathcal{R}_{\theta^*}(g) = (1-\theta^*)\mathbb{P}(Y=1, g(\mathbf{X})=0) + \theta^*\mathbb{P}(Y=0, g(\mathbf{X})=1).$$

However, since the threshold depends on the whole distribution $\mathbb{P}$, the empirical version of $\mathcal{R}_{\theta^*}$ is generally not straightforward. This relation of the F_b-score and its cost-sensitive formulation was exploited in [26] to build a practical algorithm for the F_b-score maximization, whose consistency is unfortunately not established.

3. PROPOSED PROCEDURE

In this section we describe the proposed procedure $\hat{g}$ to estimate the F_b-score optimal classifier g^*. We assume that we have access to two datasets: labeled $\mathcal{D}_n^{\mathrm{L}} = \{(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)\}$ consists of $n \in \mathbb{N}$ i.i.d. copies of $(\mathbf{X}, Y) \sim \mathbb{P}$; and unlabeled $\mathcal{D}_N^{\mathrm{U}} = \{\mathbf{X}_{n+1}, \ldots, \mathbf{X}_{n+N}\}$ consists of $N \in \mathbb{N}$ independent copies of $\mathbf{X} \sim \mathbb{P}_{\mathbf{X}}$. W.l.o.g. it is assumed that the size of the unlabeled dataset is not smaller that the size of the labeled dataset, that is, $N \geq n$. The above assumption is w.l.o.g. as long as we are willing to sacrifice constant multiplicative factors in the rate of convergence. Indeed, since we have access to $(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n), \mathbf{X}_{n+1}, \ldots, \mathbf{X}_{n+N}$, we can create another[3] dataset: $(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_{\frac{n}{2}}, Y_{\frac{n}{2}})$,

[3] For this discussion let us assume that n is even.

$\mathbf{X}_{\frac{n}{2}+1}, \ldots, \mathbf{X}_n, \mathbf{X}_{n+1}, \ldots, \mathbf{X}_{n+N}$ by simply pooling out half of the labels. This step artificially ensures that we have at least as many unlabeled data as the labeled ones.

Algorithm 1. Threshold estimation for F_b-score

```
Input: unlabeled data $\mathcal{D}_N^{\mathrm{U}}$; estimator $\hat{\eta}$; parameter $b > 0$; number of iterations $K_{\max}$
Output: threshold estimator $\hat{\theta}$
procedure BISECTION ESTIMATOR
    $\hat{R}(\theta) \leftarrow \theta b^2 \hat{\mathbb{E}}_N[\hat{\eta}(\mathbf{X})] - \hat{\mathbb{E}}_N(\hat{\eta}(\mathbf{X}) - \theta)_+$
    $\theta_{\min} \leftarrow 0$, $\theta_{\max} \leftarrow \frac{1}{1+b^2}$, $K \leftarrow 1$
while $K \leq K_{\max}$:
    if $\hat{R}\left(\frac{\theta_{\min}+\theta_{\max}}{2}\right) = 0$ then return $\frac{\theta_{\min}+\theta_{\max}}{2}$
    if $\hat{R}\left(\frac{\theta_{\min}+\theta_{\max}}{2}\right) < 0$ then $\theta_{\min} \leftarrow \frac{\theta_{\min}+\theta_{\max}}{2}$ else $\theta_{\max} \leftarrow \frac{\theta_{\min}+\theta_{\max}}{2}$
    $K \leftarrow K + 1$
endwhile
    return $\hat{\theta} = \frac{\theta_{\min}+\theta_{\max}}{2}$
```

As already mentioned, the goal is to perform a post-processing of any plug-in estimator. To this end, let $\hat{\eta}$ be any estimator of the regression function η based on the labeled data $\mathcal{D}_n^{\mathrm{L}}$. Notably, any consistent estimator of η can be used [32]. Then, we use the output of Algorithm 1 to estimate θ^*. This algorithm requires unlabeled data $\mathcal{D}_N^{\mathrm{U}}$ and the estimator $\hat{\eta}$ to be adjusted. Finally, the constructed classifier $\hat{g}$ is defined as

$$\hat{g}(\mathbf{x}) = \mathbb{I}\{\hat{\eta}(\mathbf{x}) > \hat{\theta}\}, \tag{5}$$

where $\hat{\theta}$ is computed according to Algorithm 1. Algorithm 1 is a bisection algorithm [6] applied to the function $\hat{R}(\theta)$, defined as

$$\hat{R}(\theta) = \theta b^2 \hat{\mathbb{E}}_N[\hat{\eta}(\mathbf{X})] - \hat{\mathbb{E}}_N(\hat{\eta}(\mathbf{X}) - \theta)_+,$$

where $\hat{\mathbb{E}}_N$ is an expectation taken with respect to the empirical measure $\frac{1}{N}\sum_{i=n+1}^{n+N} \delta_{\mathbf{X}_i}$ evaluated on unlabeled data This function is an empirical version of the condition on the optimal threshold θ^* imposed by Eq. (2). Note that it can be applied on top of any off-the-shelf base estimator $\hat{\eta}$ and it requires only an unlabeled sample to approximate the marginal distribution $\mathbb{P}_{\mathbf{X}}$ of $\mathbf{X}$.

3.1. General Analysis

Before stating an explicit rate on a standard nonparametric class of distributions, we provide a general analysis of the proposed estimator which can be applied in a large variety of statistical models. In particular, in this section we do not pose any assumption on the joint distribution $\mathbb{P}$ of $(\mathbf{X}, Y)$. Of course the performance of $\hat{g}$ highly depends on the quality of the initial estimator $\hat{\eta}$. Hence, our goal is to derive a bound which explicitly involves the term responsible for estimation of η by $\hat{\eta}$.

Proposition 3. *Let $\hat{\eta}$ be any estimator of η such that $\hat{\eta}(\mathbf{x}) \in (0, 1]$ almost surely. Consider $\hat{g}(\cdot) = \mathbb{I}\{\hat{\eta}(\cdot) \geq \hat{\theta}\}$ where $\hat{\theta}$ is the output of Algorithm 1 with some $K_{\max} \in \mathbb{N}$, then it holds that*

$$\mathbf{E}|\mathrm{F}_b(g^*) - (1+b^2)\hat{\theta}| \leq (1+b^2)\left(\mathtt{A}(\mathbb{P}, b)\left(\mathbf{E}||\eta - \hat{\eta}||_1 + \sqrt{\frac{\pi\mathbb{P}(Y=1)}{N}} + \frac{4}{N}\right) + 2^{-K_{\max}}\right),$$

where $\mathtt{A}(\mathbb{P}, b) = 1/(b^2\mathbb{P}(Y=1))$.

The above result states that the output of Algorithm 1 approximates the optimal value of the F-score, provided that the base estimator $\hat{\eta}$ is a good proxy for the regression function η. It also highlights the fact that this algorithm is computationally efficient, as it requires only a logarithmic number of iterations. We assumed that $\hat{\eta}(\mathbf{x}) \in (0, 1]$, which is not restrictive, since we can always project $\hat{\eta}$ on $[\delta_N, 1]$ with $0 < \delta_N \leq N^{-1/2}$ without harming its statistical properties. The derived bound depends on the probability of positive class as $1/\mathbb{P}(Y = 1)$, that is, the bound is most informative in the regime of moderately low $\mathbb{P}(Y = 1)$. At the same time, the dependency on the size of unlabeled data is $\sqrt{1/(N\mathbb{P}(Y = 1))} + 1/(N\mathbb{P}(Y = 1))$, implying that the low values of $\mathbb{P}(Y = 1)$ mostly affect the first stage of the procedure.

Proof of Proposition 3. Using Theorem 1 we known that $\mathrm{F}_b(g^*) = \theta^*(1 + b^2)$, hence it is sufficient to bound $\mathbf{E}|\theta^* - \hat{\theta}|$. Let $\bar{\theta}$ be a unique solution of $\hat{R}(\theta) = 0$ defined in Algorithm 1. Using the triange inequality we can write

$$\mathbf{E}|\theta^* - \hat{\theta}| \leq \mathbf{E}|\bar{\theta} - \hat{\theta}| + \mathbf{E}|\theta^* - \bar{\theta}|. \tag{6}$$

We bound both terms on the r.h.s. of Eq. (6) separately. For the first term in Eq. (6), notice that since $\hat{R}(\theta)$ is continuous on $[0, 1]$ and $\hat{R}(0) < 0$, $\hat{R}(1) > 0$, then classical analysis of the bisection algorithm implies that $|\hat{\theta} - \bar{\theta}| \leq 2^{-K_{\max}}$ almost surely.

We denote by F_η the cumulative distribution of $\eta(\mathbf{X})$ and by $\hat{F}_{\hat{\eta}}$ the empirical cumulative distribution of $\hat{\eta}(\mathbf{X})$ conditional on labeled data $\mathcal{D}^{\mathrm{L}}_n$. Recall the following classical fact: for any random variable $Z \in [0, T]$ a.s. and any $\theta \in [0, T]$, it holds that $\mathbf{E}[(Z - \theta)_+] = \int_\theta^T \mathbf{P}(Z \geq t)dt$. Using this fact, the thresholds $\theta^*, \bar{\theta} \in [0, 1]$ satisfy

$$b^2\theta^* = \frac{\int_{\theta^*}^1 (1 - F_\eta(t))dt}{\int_0^1 (1 - F_\eta(t))dt}, \quad b^2\bar{\theta} = \frac{\int_{\bar{\theta}}^1 (1 - \hat{F}_{\hat{\eta}}(t))dt}{\int_0^1 (1 - \hat{F}_{\hat{\eta}}(t))dt}. \tag{7}$$

Using the fact that $1 - F_\eta(t) \geq 0$, on the event $\{\bar{\theta} \leq \theta^*\}$ Eq. (7) yields

$$b^2(\theta^* - \bar{\theta}) \leq \frac{\int_{\bar{\theta}}^1 (1 - F_\eta(t))dt}{\int_0^1 (1 - F_\eta(t))dt} - \frac{\int_{\bar{\theta}}^1 (1 - \hat{F}_{\hat{\eta}}(t))dt}{\int_0^1 (1 - \hat{F}_{\hat{\eta}}(t))dt}.$$

Adding and subtracting $\int_{\bar{\theta}}^1 (1 - \hat{F}_{\hat{\eta}}(t))dt \Big/ \int_0^1 (1 - F_\eta(t))dt$ on the right hand side of the above inequality we get

$$\begin{aligned} b^2(\theta^* - \bar{\theta}) &\leq \frac{\int_{\bar{\theta}}^1 (\hat{F}_{\hat{\eta}}(t) - F_\eta(t))dt - \bar{\theta}\int\limits_0^1 (\hat{F}_{\hat{\eta}}(t) - F_\eta(t))dt}{\int_0^1 (1 - F_\eta(t))dt} \\ &\leq \frac{1}{\mathbb{P}(Y = 1)} \int\limits_0^1 |F_\eta(t) - \hat{F}_{\hat{\eta}}(t)|dt = \frac{1}{\mathbb{P}(Y = 1)} ||F_\eta - \hat{F}_{\hat{\eta}}||_1. \end{aligned} \tag{8}$$

Similar argument used on the event $\{\bar{\theta} > \theta^*\}$ yields

$$b^2(\hat{\theta} - \theta^*) \leq \frac{1}{\mathbb{P}(Y = 1)} ||F_\eta - \hat{F}_{\hat{\eta}}||_1. \tag{9}$$

The combination of Eqs. (8), (9) allows us to derive

$$b^2\mathbf{E}|\theta^* - \bar{\theta}| \leq \frac{1}{\mathbb{P}(Y = 1)} \mathbf{E}||F_\eta - \hat{F}_{\hat{\eta}}||_1.$$

Let us introduce $\hat{F}_\eta$, which stands for the empirical cumulative distribution of $\eta(\mathbf{X})$ based on $\mathcal{D}^{\mathrm{U}}_N$. Using the triangle inequality we get

$$||F_\eta - \hat{F}_{\hat{\eta}}||_1 \leq ||F_\eta - \hat{F}_\eta||_1 + ||\hat{F}_{\hat{\eta}} - \hat{F}_\eta||_1. \tag{10}$$

Let $p(t) = \mathbb{P}(\eta(\mathbf{X}) \geq t)$, then by Bernstein's inequality we have

$$\mathbf{E}|\mathbb{P}(\eta(\mathbf{X}) \geq t) - \hat{\mathbb{P}}_N(\eta(\mathbf{X}) \geq t)| = \int\limits_0^\infty \mathbf{P}\left(|\mathbb{P}(\eta(\mathbf{X}) \geq t) - \hat{\mathbb{P}}_N(\eta(\mathbf{X}) \geq t)| \geq x\right) dx$$

$$\leq 2\int\limits_0^\infty \exp\left(-\frac{Nx^2}{2(p(t) + \frac{1}{3}x)}\right) dx.$$

Furthermore, we can write for the inner integral

$$\int\limits_0^\infty \exp\left(-\frac{Nx^2}{2(p(t) + \frac{1}{3}x)}\right) dx = \left(\int\limits_0^{3p(t)} + \int\limits_{3p(t)}^\infty\right) \exp\left(-\frac{Nx^2}{2(p(t) + \frac{1}{3}x)}\right) dx$$

$$\leq \int\limits_0^{3p(t)} \exp\left(-\frac{N^2x^2}{4p(t)}\right) dx + \int\limits_{3p(t)}^\infty \exp\left(-\frac{Nx}{4}\right) dx \leq \sqrt{\frac{\pi p(t)}{N}} + \frac{4}{N}.$$

Therefore, using Fubini's theorem, we obtain

$$\mathbf{E}||F_\eta - \hat{F}_\eta||_1 = \int\limits_0^1 \mathbf{E}|\mathbb{P}(\eta(\mathbf{X}) \geq t) - \hat{\mathbb{P}}_N(\eta(\mathbf{X}) \geq t)|dt \leq \sqrt{\frac{\pi}{N}} \int\limits_0^1 \sqrt{\mathbb{P}(\eta(\mathbf{X}) \geq t)}dt + \frac{4}{N}.$$

Applying Cauchy–Schwarz inequality to the first term on the r.h.s. of the above inequality we derive the following bound

$$\mathbf{E}||F_\eta - \hat{F}_\eta||_1 \leq \sqrt{\frac{\pi\mathbb{P}(Y=1)}{N}} + \frac{4}{N}.$$

It remains to bound $||\hat{F}_{\hat{\eta}} - \hat{F}_\eta||_1$. Let $Z_i = \eta(\mathbf{X}_i)$ and $\hat{Z}_i = \hat{\eta}(\mathbf{X}_i)$ for all $i = n+1, \ldots, N$, then for the second term on the r.h.s. of Eq. (10) we can write

$$||\hat{F}_{\hat{\eta}} - \hat{F}_\eta||_1 = \frac{1}{N}\int\limits_0^1 \left|\sum_{i=n+1}^{n+N} \left(\mathbb{I}\{Z_i \leq t\} - \mathbb{I}\{\hat{Z}_i \leq t\}\right)\right| dt.$$

This expression corresponds to the Wasserstein-1 distance between empirical measures of $\{Z_{n+1}, \ldots, Z_{n+N}\}$ and $\{\hat{Z}_{n+1}, \ldots, \hat{Z}_{n+N}\}$. Hence, using its alternative definition we get

$$||\hat{F}_{\hat{\eta}} - \hat{F}_\eta||_1 = \inf_{\omega \in \mathbb{S}_N} \frac{1}{N}\sum_{i=n+1}^{n+N} |Z_i - \hat{Z}_{\omega(i)}| \leq \frac{1}{N}\sum_{i=n+1}^{N} |\eta(\mathbf{X}_i) - \hat{\eta}(\mathbf{X}_i)|,$$

where the infimum is taken over all permutations $\mathbb{S}_N$ of $\{n+1, \ldots, n+N\}$. Finally, since conditionally on the labeled data $\mathcal{D}_n^{\mathrm{L}}$ the random variables $|\eta(\mathbf{X}_i) - \hat{\eta}(\mathbf{X}_i)|$ with $i = n+1, \ldots, n+N$ are i.i.d., then $\mathbf{E}||\hat{F}_{\hat{\eta}} - \hat{F}_\eta||_1 \leq \mathbf{E}||\eta - \hat{\eta}||_1$. □

While Proposition 3 allows to estimate the value of the optimal F-score, it does not guarantee that the proposed post-processing performs well in terms of this measure. Propositions 4 addresses this question.

Proposition 4 (Post-processing bound). *Let $\hat{\eta}$ be any estimator of η satisfying assumptions of Proposition* 3. *Consider $\hat{g}(\cdot) = \mathbb{I}\{\hat{\eta}(\cdot) \geq \hat{\theta}\}$ where $\hat{\theta}$ is the output of Algorithm 1 with some $K_{\max} \in \mathbb{N}$, then it holds that*

$$\mathrm{F}_b(g^*) - \mathbf{E}[\mathrm{F}_b(\hat{g})] \leq (1+b^2)\left(2\mathtt{A}^2(\mathbb{P}, b) \vee \mathtt{A}(\mathbb{P}, b)\mathbf{E}||\eta - \hat{\eta}||_1 + \mathtt{A}^2(\mathbb{P}, b)\left(\sqrt{\frac{\pi\mathbb{P}(Y=1)}{N}} + \frac{4}{N}\right)\right)$$

$$+ (1+b^2)\mathtt{A}(\mathbb{P}, b)2^{-K_{\max}},$$

where $\mathtt{A}(\mathbb{P}, b) = 1/(b^2\mathbb{P}(Y = 1))$.

Proof. Observe that Lemma 2 and Proposition 3 immediately yield

$$\begin{aligned}
\mathrm{F}_b(g^*) - \mathbf{E}[\mathrm{F}_b(\hat{g})] &\leq \mathbf{E}\left[\frac{1+b^2}{b^2\mathbb{P}(Y=1) + \mathbb{P}(\hat{g}(\mathbf{X}) = 1)}\left(||\eta - \hat{\eta}||_1 + |\theta^* - \hat{\theta}|\right)\right] \\
&\leq (1+b^2)\mathbf{E}\left[\mathtt{A}(\mathbb{P}, b)\left(||\eta - \hat{\eta}||_1 + |\theta^* - \hat{\theta}|\right)\right] \\
\leq (1+b^2)&\left(2\mathtt{A}^2(\mathbb{P}, b) \vee \mathtt{A}(\mathbb{P}, b)\mathbf{E}||\eta - \hat{\eta}||_1 + \mathtt{A}^2(\mathbb{P}, b)\left(\sqrt{\frac{\pi\mathbb{P}(Y=1)}{N}} + \frac{4}{N}\right)\right) \\
&\quad + (1+b^2)\mathtt{A}(\mathbb{P}, b)2^{-K_{\max}}.
\end{aligned}$$

□

The interpretation of both bound in Propositions 3 and 4 is straightforward. There are three terms: the first term $\mathbf{E}||\eta - \hat{\eta}||_1$ is the estimation error of the regression function; the second term $N^{-1/2}$ is the price that we pay for the unknown marginal distribution of the features; and the last term $2^{-K_{\max}}$ correspond to the deterministic error of Algorithm 1—we are not solving the equation $\hat{R}(\theta) = 0$ precisely.

Proposition 4 also allows to make the following corollary.

Corollary 1 (Universal consistency). Assume that the estimator $\hat{\eta}$ constructed on labeled data $\mathcal{D}_n^{\mathrm{L}}$ satisfies

$$\lim_{n\to\infty} \mathbf{E}||\eta - \hat{\eta}||_1 = 0, \quad \forall \mathbb{P} \text{ s.t. } \mathbb{P}(Y=1) \neq 0,$$

then, the constructed classifier $\hat{g}$ satisfies

$$\lim_{n,N,K_{\max}\to\infty} \mathbf{E}[\mathrm{F}_b(\hat{g})] = \mathrm{F}_b(g^*), \quad \forall \mathbb{P} \text{ s.t. } \mathbb{P}(Y=1) \neq 0.$$

Following the celebrated result of [32] we can conclude for instance that if $\hat{\eta}$ is the k-Nearest Neighbors estimator with $k \to \infty$ and $k/n \to 0$, then the classifier $\hat{g}$ is universally consistent for all non-degenerate distributions in terms of the F_b-score.

4. NONPARAMETRIC MINIMAX ANALYSIS

Note that the result of Proposition 4 was derived without any assumptions[4] on the joint distribution of $(\mathbf{X}, Y)$. While in nonparametric scenarios this bound is known to be optimal already for the misclassification risk [39], it can be significantly improved under the celebrated margin assumption [1, 21, 27, 38]. The purpose of this section is to understand the minimax rate of convergence of the excess score of the proposed classifier $\hat{g}$ under the standard nonparametric assumptions with the margin condition and establish its rate optimality.

4.1. Minimax Setup

We start by describing the class of joint distributions of $(\mathbf{X}, Y)$ that is considered. The first assumption is made on smoothness of the regression function $\eta : \mathbb{R}^d \mapsto [0, 1]$.

Definition 6 (Hölder smoothness). *Let* $L > 0$ *and* $\beta > 0$. *The class of functions* $\Sigma(\beta, L, \mathbb{R}^d)$ *consists of all functions* $h : \mathbb{R}^d \mapsto [0, 1]$ *such that for all* $x, x' \in \mathbb{R}^d$, *we have*

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

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

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

[4] We only assumed that $\mathbb{P}(Y = 1) > 0$ which is hardly an assumption in this context.

Assumption 1 is usually not sufficient to obtain non-asymptotic rates: extra assumptions are required on the marginal distribution $\mathbb{P}_{\mathbf{X}}$ of the vector $\mathbf{X} \in \mathbb{R}^d$. There are various assumptions that can be imposed on $\mathbb{P}_{\mathbf{X}}$. For instance strong and weak density assumptions [2], tail assumption [13], or assumption on the covering number of the support of $\mathbb{P}_{\mathbf{X}}$ [18]. Since this is not the main concern of this work, we stick to the most basic assumption on the marginal distribution. We refer to [13] and references therein for the profound study of different conditions on the distribution of $\mathbf{X}$. To introduce our density assumption, let us first define the notion of a regular set.

Definition 7. *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.

The strong density assumption is stated below.

Assumption 2. *We say that the marginal distribution $\mathbb{P}_{\mathbf{X}}$ of the vector $\mathbf{X} \in \mathbb{R}^d$ satisfies the strong density assumption if*

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

Note that the bound in Proposition 4 explodes with the growth of $1/\mathbb{P}(Y=1)$, thus to make our minimax bounds meaningful we need to assume that $\mathbb{P}(Y=1)$ is lower bounded by an arbitrary small, but fixed positive constant p.

Assumption 3. *There exists a positive constant $p \leq 1/2$ such that $\mathbb{P}(Y=1) \geq p$.*

The assumption that $p \leq 1/2$ is mostly technical as our lower bound construction is tailored for this scenario. It is possible to relax this assumption to $p \leq 1 - \zeta$ with some fixed positive ζ. However, let us also mention that the F_b-score is mostly used in the situation when $\mathbb{P}(Y=1) < \mathbb{P}(Y=0)$, hence, from practical perspective assumption $p \leq 1/2$ is reasonable.

We study the rates of convergence under the margin assumption [2, 21, 34]. A version of this assumption was used originally in the density level set estimation by [27].

Assumption 4. *We say that the distribution $\mathbb{P}$ of the pair $(\mathbf{X}, Y) \in \mathbb{R}^d \times \{0,1\}$ satisfies the α-margin assumption if there exist constants $C_0 > 0$, and $\alpha \geq 0$ such that for every positive $\delta \leq (n^{-\beta/(2\beta+d)} \ln^2 n) \wedge 1$ we have*

$$\mathbb{P}_{\mathbf{X}}(0 < |\eta(\mathbf{X}) - \theta^*| \leq \delta) \leq C_0 \delta^{\alpha}.$$

This is a natural adaptation of the classical margin assumption [34] to the F_b-score setup. It allows to prove more optimistic minimax rates of convergence compared to the case with no assumption. Intuitively, the margin condition specifies the behavior of the regression function around the decision threshold θ^*. The case $\alpha = 0$ and $C_0 \geq 1$ corresponds to no assumption, and the classification problem gets statistically easier for high values of α. We require the margin Assumption 4 only in a small strip around the optimal threshold θ^* of size $(n^{-\beta/(2\beta+d)} \ln^2 n) \wedge 1$. This modification is convenient for us as it simplifies the lower bound construction.

Finally, we are in position to define the family of joint distribution of $(\mathbf{X}, Y)$.

Definition 8. *Let $\mathcal{P}(\alpha, \beta)$ be a class of distributions on $\mathbb{R}^d \times \{0,1\}$ such that Assumptions 1–4 are satisfied.*

Our goal is to understand the behavior of the minimax risk over $\mathcal{P}$, defined as

$$\inf_{\hat{g}} \sup_{\mathbb{P} \in \mathcal{P}(\alpha,\beta)} \left\{ \mathrm{F}_b(g^*) - \mathbf{E}[\mathrm{F}_b(\hat{g})] \right\},$$

where the infimum is taken over all estimators $\hat{g}$ based on labeled $\mathcal{D}_n^{\mathrm{L}}$ and unlabeled $\mathcal{D}_N^{\mathrm{U}}$, not necessary of the post-processing nature that we discussed above.

4.2. Lower Bound

Theorem 9. *If $\alpha\beta \leq d$, then there exists constant $c > 0$, which depends on $d, C_0, \alpha, p, c_0, r_0$, $\mu_{\min}, \mu_{\max}$ such that for all n satisfying $12 \ln^2 n \leq n^{\beta/(2\beta+d)}$*

$$\inf_{\hat{g}} \sup_{\mathbb{P} \in \mathcal{P}(\alpha,\beta)} \left\{ \mathrm{F}_b(g^*) - \mathbf{E}[\mathrm{F}_b(\hat{g})] \right\} \geq c n^{-\frac{(1+\alpha)\beta}{2\beta+d}}, \tag{11}$$

where the infimum is taken over all estimators $\hat{g}$ based on n labeled and N independent unlabeled samples.

The proof of this lower bound relies on [35, Theorem 2.7] and the construction of the distributions is inspired by both Rigollet et al. [29] and Tsybakov [2]. However, in the classical setups considered by Rigollet et al. [29] and Audibert and Tsybakov [2] the threshold θ^* is known and it is independent from the distribution $\mathbb{P}$, which is the main difficulty in the case of the F_b-score classification.

We stress that this lower bound does not include the size of the unlabeled data. As confirmed by our upper bound in the next section, this is the correct dependency on both n and N.

4.3. Upper Bound

To derive upper bound on the minimax risk of the proposed classifier $\hat{g}$, we need a good estimator $\hat{\eta}$ of η. In the nonparametric setup considered in this work, there are various ways to build such estimator. We require an estimator $\hat{\eta}$ based on $\mathcal{D}_n^{\mathrm{L}}$ which satisfies for all $t > 0$

$$\mathbf{P}(|\hat{\eta}(x) - \eta(x)| \geq t) \leq C_1 \exp(-C_2 a_n t^2) \text{ a.s. } \mathbb{P}_{\mathbf{X}} \tag{12}$$

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

For instance, in the case of β-smooth regression function considered here, a typical nonparametric 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].

Theorem 10. *Assume that $\hat{\eta}$ satisfies Eq.* (12) *and $K_{\max} \geq \lceil (1+\alpha) \log N \rceil$ then there exists a constant $C > 0$ independent of p and b such that*

$$\sup_{\mathbb{P} \in \mathcal{P}(\alpha,\beta)} (\mathrm{F}_b(g^*) - \mathbf{E}[\mathrm{F}_b(\hat{g})]) \leq C(1+b^2) \left(\mathtt{A}(\mathcal{P}, b) \vee \mathtt{A}^{2+\alpha}(\mathcal{P}, b) \right) \left(n^{-\frac{(1+\alpha)\beta}{2\beta+d}} + N^{-\frac{1+\alpha}{2}} \right),$$

where $\hat{g}$ is defined in Eq. (5) *and $\mathtt{A}(\mathcal{P}, b) = 1/(b^2 p)$.*

Strictly speaking the size of unlabeled data N is present in this upper bound. However, for the minimax perspective the associated rate $N^{-(1+\alpha)/2}$ is always faster than $n^{-(1+\alpha)\beta/(2\beta+d)}$ under assumption that $N \geq n$. Actually, even if the assumption $N \geq n$ fails to be satisfied, we can artificially augment unlabeled data by polling out half of the labels from $\mathcal{D}_n^{\mathrm{L}}$ and augmenting the unlabeled dataset by this half. Applying Theorem 10 for the same algorithm $\hat{g}$ that uses $\lceil \frac{n}{2} \rceil$ labeled and $N + \lfloor \frac{n}{2} \rfloor$ unlabeled data we get

$$\begin{aligned}\sup_{\mathbb{P} \in \mathcal{P}(\alpha,\beta)} (\mathrm{F}_b(g^*) - \mathbf{E}[\mathrm{F}_b(\hat{g})]) &\leq C'(1+b^2) \left(\mathtt{A}(\mathcal{P}, b) \vee \mathtt{A}^{2+\alpha}(\mathcal{P}, b) \right) \left(n^{-\frac{(1+\alpha)\beta}{2\beta+d}} + (N+n)^{-\frac{1+\alpha}{2}} \right) \\ &\leq C''(1+b^2) \left(\mathtt{A}(\mathcal{P}, b) \vee \mathtt{A}^{2+\alpha}(\mathcal{P}, b) \right) n^{-\frac{(1+\alpha)\beta}{2\beta+d}},\end{aligned}$$

which matches (up to constant factors) the rate derived in the lower bound.

Theorem 10 indicates that it is statistically more difficult to estimate the regression function η than the unknown threshold θ^*. Besides, this result improves the rate derived in [38], which is of order $n^{-(1+\alpha\wedge 1)\beta/(2\beta+d)}$ and which is derived under the additional assumption that $\eta(\mathbf{X})$ does not have atoms. Finally, there is a direct analogy of the rate in the standard setup derived by [2] and that of Theorem 10.

To prove Theorem 10 let us first recall the following theorem due to [2].

Theorem 11 (Audibert and Tsybakov [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 $\mathbb{P}_{\mathbf{X}}$ satisfies the strong density assumption. Then, there exists an estimator $\hat{\eta}$ of the regression function satisfying*

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

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

Proof of Theorem 10. Set $a_n = n^{-\beta/(2\beta+d)}$. Using Lemma 2 and our assumption that $\mathbb{P}(Y=1) \geq p$ we get

$$\mathrm{F}_b(g^*) - \mathbf{E}[\mathrm{F}_b(\hat{g})] \leq \frac{1+b^2}{b^2 p}\mathbf{E}|\eta(\mathbf{X}) - \theta^*|\mathbb{I}\{g^*(\mathbf{X}) \neq \hat{g}(\mathbf{X})\}.$$

Then, since using the definition of $\hat{g}$ and the form of the F_b-score optimal classifier g^* we can write

$$\mathrm{F}_b(g^*) - \mathbf{E}[\mathrm{F}_b(\hat{g})] \leq \frac{1+b^2}{b^2 p}\Big\{\mathbf{E}|\eta(\mathbf{X}) - \theta^*|\mathbb{I}\{|\eta(\mathbf{X}) - \theta^*| \leq 2|\eta(\mathbf{X}) - \hat{\eta}(\mathbf{X})|\} \tag{13}$$

$$+\,\mathbf{E}|\eta(\mathbf{X}) - \theta^*|\mathbb{I}\{|\eta(\mathbf{X}) - \theta^*| \leq 2|\theta^* - \hat{\theta}|\}\Big\}. \tag{14}$$

For the first term in Eq. (13) we closely follow the peeling argument of Audibert and Tsybakov [2]. Recall that the margin Assumption 4 is required to hold only for $\delta \leq n^{-\beta/(2\beta+d)}\ln^2 n$, thus it requires us to slightly modify the peeling argument to account for this subtlety. For $\delta = a_n^{-1/2}$ and $j_n = (1/2)\log_2 \ln^2(n)$, with C_2 from Eq. (12), define the following sets

$$\mathcal{A} = \left\{x \in \mathbb{R}^d \colon 0 < |\eta(x) - \theta^*| \leq \delta\right\},$$
$$\mathcal{A}_j = \left\{x \in \mathbb{R}^d \colon 2^j < |\eta(x) - \theta^*| \leq 2^{j+1}\delta\right\}, \quad \forall j = 0, \dots j_n,$$
$$\bar{\mathcal{A}} = \left\{x \in \mathbb{R}^d \colon |\eta(x) - \theta^*| \geq 2^{j_n+1}\delta\right\}.$$

Set $T = \mathbf{E}|\eta(\mathbf{X}) - \theta^*|\mathbb{I}\{|\eta(\mathbf{X})\} - \theta^*| \leq 2|\eta(\mathbf{X}) - \hat{\eta}(\mathbf{X})|\}$, then we have thanks to the margin Assumption 4 and Eq. (12) that

$$T \leq C_0\delta^{1+\alpha} + C_0C_1\delta^{1+\alpha}\sum_{j=0}^{j_n} 2^{(j+1)(1+\alpha)}\exp\left(-C_2 2^{2j}\right) + \int \mathbf{P}\left(|\eta(x) - \hat{\eta}(x)| \geq 2^{j_n}\delta\right) d\mathbb{P}_{\mathbf{X}}(x)$$

$$\leq C_0\delta^{1+\alpha} + C_0C_1\delta^{1+\alpha}\sum_{j=0}^{j_n} 2^{(j+1)(1+\alpha)}\exp\left(-C_2 2^{2j}\right) + C_1\exp\left(-C_2 2^{2j_n}\right)$$

$$\leq C_0\delta^{1+\alpha} + C_0C_1\delta^{1+\alpha}\sum_{j=0}^{\infty} 2^{(j+1)(1+\alpha)}\exp\left(-C_2 2^{2j}\right) + C_1 n^{-C_2\ln n} \leq \mathtt{A}\left(a_n^{-\frac{1+\alpha}{2}} + n^{-C_2\ln n}\right),$$

with $\mathtt{A}$ being independent from $\mathbb{P}(Y=1), b$.

We use Eqs. (8), (9) to bound the second term in Eq. (13) by

$$\mathbf{E}|\eta(\mathbf{X}) - \theta^*|\mathbb{I}\{|\eta(\mathbf{X}) - \theta^*| \leq 2b^{-2}(\mathbb{P}(Y=1))^{-1}\left(||F_{\hat{\eta}} - \hat{F}_{\hat{\eta}}||_\infty + ||\eta - \hat{\eta}||_1\right) + 2^{-K_{\max}+1}\}.$$

The application of the margin assumption yields

$$\mathbf{E}|\eta(\mathbf{X}) - \theta^*|\mathbb{I}\{|\eta(\mathbf{X}) - \theta^*| \leq 2|\theta^* - \hat{\theta}|\} \leq \mathtt{A}_1\left(\mathbf{E}||F_{\hat{\eta}} - \hat{F}_{\hat{\eta}}||_\infty^{1+\alpha} + \mathbf{E}||\eta - \hat{\eta}||_1^{1+\alpha}\right) + 2^{-K_{\max}+3},$$

where $\mathtt{A}_1 = C_0(6/(b^2\mathbb{P}(Y=1)))^{1+\alpha}$. Thanks to the Dvoretzky–Kiefer–Wolfowitz inequality we have

$$\mathbf{E}||F_{\hat{\eta}} - \hat{F}_{\hat{\eta}}||_\infty^{1+\alpha} \leq \mathtt{A}_2 N^{-\frac{1+\alpha}{2}}$$

with some unversal $\mathtt{A}_2 > 0$.

Since the estimator $\hat{\eta}$ satisfies Assumption in Eq. (12), it holds that

$$\mathbf{E}||\eta - \hat{\eta}||_1^{1+\alpha} \leq \mathtt{A}_2 a_n^{-\frac{1+\alpha}{2}},$$

with $\mathtt{A}_2$ being independent from $\mathbb{P}(Y=1)$, b and n. Finally, in the family $\mathcal{P}(\alpha, \beta)$ it holds uniformly that $\mathbb{P}(Y=1) \geq p$, which concludes the proof of the result. □

Algorithm 2 Threshold estimation for F$_b$-score (set-valued classification)

```
  Input: unlabeled data $\mathcal{D}_N^{\mathrm{U}}$; estimators $\hat{\eta}_1, \dots, \hat{\eta}_K$; parameter $b > 0$; number of iterations $K_{\max}$
  Output: threshold estimator $\hat{\theta}$
procedure BISECTION ESTIMATOR
      $\hat{R}(\theta) \leftarrow b^2\theta - \sum_{k=1}^{K} \hat{\mathbb{E}}_N(\hat{\eta}_k(\mathbf{X}) - \theta)_+$
      $\theta_{\min} \leftarrow 0$
      $\theta_{\max} \leftarrow \frac{1}{1+b^2}$
      $K \leftarrow 1$
while $K \leq K_{\max}$:
    if $\hat{R}\left(\frac{\theta_{\min}+\theta_{\max}}{2}\right) = 0$ then return $\frac{\theta_{\min}+\theta_{\max}}{2}$
    if $\hat{R}\left(\frac{\theta_{\min}+\theta_{\max}}{2}\right) < 0$ then $\theta_{\min} \leftarrow \frac{\theta_{\min}+\theta_{\max}}{2}$ else $\theta_{\max} \leftarrow \frac{\theta_{\min}+\theta_{\max}}{2}$
      $K \leftarrow K + 1$
endwhile
    return $\frac{\theta_{\min}+\theta_{\max}}{2}$
```

5. GENERALIZATION FOR SET-VALUED CLASSIFICATION

In this section we generalize the proposed procedure to the setup of set-valued classification [10, 23, 30, 36]. Recall, that in the set-valued classification instead of a binary label we observe a multi-class variable $Y \in [K] := \{1, \dots, K\}$. The K conditional distributions of labels are defined $\forall k \in [K]$ as $\eta_k(\mathbf{X}) = \mathbb{P}(Y = k|\mathbf{X})$. The main idea behind the set-valued classification is to replace the single-label prediction $g : \mathbb{R}^d \to [K]$ with a set-valued prediction of the form $\Gamma : \mathbb{R}^d \to 2^{[K]}$. In words, a set-valued classifier Γ outputs a set of possible candidates for a given feature $\mathbf{X} \in \mathbb{R}^d$. Clearly, each single-labeled prediction can be seen as a set-valued one which always outputs a singleton.

This framework received a great deal of attention in recent years, with contributions ranging from more applied to more theoretical [5, 19, 20, 28, 30]. The sudden popularity of this framework is mainly connected with its ability to tackle complex large-scale multi-class classification problems and provide a quantification of uncertainty about the provided prediction. On intuitive level a good set-valued classifier $\Gamma : \mathbb{R}^p \to 2^{[K]}$ strikes for the balance between two concurrent notions: the size defined as $\mathbb{E}|\Gamma(\mathbf{X})|$ and the error rate defined as $\mathbb{P}(Y \notin \Gamma(\mathbf{X}))$. Set-valued classifiers Γ with large size are less informative, while, small size implies that Γ is less likely to contain the true class—the error rate is higher. To address this balance, we follow [7, 23] and define the precision and the recall of Γ as

$$\mathrm{Pr}(\Gamma) = \frac{\mathbb{P}(Y \in \Gamma(\mathbf{X}))}{\mathbb{E}|\Gamma(\mathbf{X})|}, \quad \mathrm{Re}(\Gamma) = \mathbb{P}(Y \in \Gamma(\mathbf{X})),$$

where $|\cdot|$ stands for the cardinality of a finite set. High recall implies that the classifier Γ captures the correct class with high probability and it is trivially maximized by $\Gamma^{\mathrm{all}}(\mathbf{X}) \equiv \{1, \dots, K\}$. However, this classifier yields low values of precision for even moderately large K, since $\mathrm{Pr}(\Gamma^{\mathrm{all}}) = \frac{1}{K}$. We look for a trade-off in terms of the F$_b$-score, defined for a fixed $b > 0$ analogously to the binary case[5] as

$$\mathrm{F}_b(\Gamma) = \left(\frac{1}{1+b^2}\Big(\mathrm{Pr}(\Gamma)\Big)^{-1} + \frac{b^2}{1+b^2}\Big(\mathrm{Re}(\Gamma)\Big)^{-1}\right)^{-1},$$

[5] Again it is assumed that the (weighted) harmonic mean of any real number and zero is zero.

which is again the weighted harmonic mean between the precision and the recall. Consequently, the optimal set-valued classifier Γ^* is defined as a maximizer of the F_b-score, that is,

$$\Gamma^* \in \arg\max_{\Gamma} \mathrm{F}_b(\Gamma),$$

where the maximum is taken over all set-valued classifiers. Our analysis of the binary case can be generalized to this setup using the following result.

Theorem 12. *An optimal set-valued classifier $\Gamma^* : \mathbb{R}^d \to 2^{[K]}$ can be defined point-wise as*

$$\Gamma^*(\mathbf{x}) = \{k \in [K] \colon \eta_k(\mathbf{x}) > \theta^*\},$$

where θ^* is a unique root of

$$\theta \mapsto b^2\theta - \sum_{k=1}^{K} \mathbb{E}(\eta_k(\mathbf{x}) - \theta)_+.$$

Moreover, for any set-valued classifier $\Gamma : \mathbb{R}^p \to [K]$ it holds that

$$\mathrm{F}_b(\Gamma^*) - \mathrm{F}_b(\Gamma) = \frac{1 + b^2}{b^2 + \mathbb{E}|\Gamma(\mathbf{X})|}\mathbb{E}\left[\sum_{k \in \Gamma(\mathbf{X}) \triangle \Gamma^*(\mathbf{X})} |\eta_k(\mathbf{X}) - \theta^*|\right],$$

where $\triangle$ stands for the symmetric difference of two sets.

After this result Algorithm 1 is extended in a straightforward way to the set-valued classification framework. As before, assume that we have two i.i.d. datasets: labeled $\mathcal{D}_n^{\mathrm{L}}$ and unlabeled $\mathcal{D}_N^{\mathrm{U}}$. Given estimators $\hat{\eta}_1, \ldots, \hat{\eta}_K$ of $\eta_1, \ldots, \eta_K$, the only required modification to Algorithm 1 is the definition of the function $\hat{R}(\theta)$. This modification is summarized in Algorithm 2, where we used the empirical version of the condition imposed on θ^* in the context of set-valued classification. Similar guarantees can be derived on this estimator exploiting that form of the excess score provided by Theorem 12 note, however, that the role of $\mathbb{P}(\hat{g}(\mathbf{X}) = 1)$ in this case is played by the expected size $\mathbb{E}|\Gamma(\mathbf{X})|$ of the set-valued classifier.

6. CONCLUSION

A new post-processing algorithm for F_b-score maximization, which is able to leverage the unlabeled data is proposed. This algorithm enjoys a general post-processing finite-sample bound, which leads to a universally consistent approach. Under nonparametric assumptions the algorithm yields minimax optimal rate of convergence, improving upon previously known results. Finally, the extension to the set-valued classification is discussed. An interesting open question is to understand the dependency of the excess score on the marginal probability of the positive class. Besides, it would be valuable to derive classification procedures under parametric assumptions on the joint distribution, such as the logit model.

7. OMITTED PROOFS

Lemma 13. *Assume that $\mathbb{P}(Y = 1) \neq 0$, then there exists $\theta^* \in [0, 1]$ which is a unique solution of*

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

Proof. The mapping $\theta \mapsto b^2\mathbb{P}(Y = 1)\theta$ is continuous and strictly increasing on $[0, 1]$ and the mapping $\theta \mapsto \mathbb{E}(\eta(\mathbf{X}) - \theta)_+$ is non-increasing on $[0, 1]$. Moreover, if $\mathbb{P}(Y = 1) \neq 0$, then for $R^*(\theta) = b^2\mathbb{P}(Y = 1)\theta - \mathbb{E}(\eta(\mathbf{X}) - \theta)_+$ it holds that

$$R^*(0) < 0, \quad R^*(1) > 0.$$

Thus, it is sufficient to demonstrate that R^* is continuous. Let $\theta, \theta' \in [0, 1]$, then, due to the Lipschitz continuity of $(\cdot)_+$ we can write

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

This implies that the mapping $\theta \mapsto \mathbb{E}(\eta(\mathbf{X}) - \theta)_+$ is a contraction and thus is continuous, hence R^* is continuous and the threshold θ^* is well-defined, that is, it exists and is unique. □

Proof of Theorem 9. For simplicity, we provide the proof for $b = 1$ and will write $\mathcal{E}(\hat{g})$ instead of $\mathcal{E}_1(\hat{g})$. The construction is inspired by lower bounds derived in [2, 29]. We define the regular grid on $\mathbb{R}^d$ as

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

and denote by $n_q(x) \in G_q$ as the closest point to 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 $\mathcal{X}'_1, \dots, \mathcal{X}'_{q^d}$. 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 $\mathcal{X}_j := \mathcal{X}'_j$, $\mathcal{X}_{-j} := \mathcal{X}'_{-j}$. For every $\boldsymbol{\omega} = (\omega_1, \dots, \omega_m)^\top \in \{-1, 1\}^m$ we define a regression function $\eta_{\boldsymbol{\omega}}$ as

$$\eta_{\boldsymbol{\omega}}(\mathbf{x}) = \begin{cases} \frac{1}{4} + \omega_j \varphi(x), & \text{if } x \in \mathcal{X}_j \\ \frac{1}{4} - \omega_j \varphi(x), & \text{if } x \in \mathcal{X}_{-j} \\ \frac{1}{4}, \quad \text{if } x \in \mathcal{B}(0, \sqrt{d}) \setminus \left(\cup_{j=-m, j \neq 0}^{m} \mathcal{X}_j \right) \\ \tau, \quad \text{if } x \in \mathbb{R}^d \setminus \mathcal{B}(0, \sqrt{d} + \rho) \\ \xi(x), \quad \text{if } x \in \mathcal{B}(0, \sqrt{d} + \rho) \setminus \mathcal{B}(0, \sqrt{d}), \end{cases}$$

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. Set $\varphi(x) := C_\varphi q^{-\beta} u(q||\mathbf{x} - n_q(x)||_2)$ 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 $|\xi(x) - \xi_x(x')| \leq L||x - x'||_2^\beta$ for any $x, x' \in \mathbb{R}^d$.

For any $\boldsymbol{\omega} \in \{-1, 1\}^m$ we construct a marginal distribution $P_{\mathbf{X}}$ which is independent from $\boldsymbol{\omega}$ 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 $\text{Leb}(A_0) = 1 - mq^{-d}$. The density μ is constructed as

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

To complete the construction it remains 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. To achieve this we would like to set $\theta^* = 1/4$ and we would like to demonstrate that there exists an appropriate choice of τ which ensures that such a choice is valid. First, recall that the optimal threshold θ^* for the classification with the F_b-score must satisfy the equation

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

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

$$\mathbb{E}_\mu \eta_{\boldsymbol{\omega}}(\mathbf{X}) = \int\limits_{\mathbb{R}^d} \eta_{\boldsymbol{\omega}}(x) d\mu(x)$$

$$= \sum_{j=1}^{m} \int\limits_{\mathcal{X}_j} (1/4 + \omega_j \xi(x)) d\mu(x) + \sum_{j=1}^{m} \int\limits_{\mathcal{X}_{-j}} (1/4 - \omega_j \xi(x)) d\mu(x) + \int\limits_{A_0} \tau d\mu(x)$$

$$= \frac{mw}{2} + \tau(1 - 2mw).$$

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

$$\mathbb{E}_\mu(\eta_{\boldsymbol{\omega}}(\mathbf{X}) - 1/4)_+ = \sum_{j=1}^{m} \int\limits_{\mathcal{X}_j} (\omega_j \xi(x))_+ d\mu(x) + \sum_{j=1}^{m} \int\limits_{\mathcal{X}_{-j}} (-\omega_j \xi(x))_+ d\mu(x) + \int\limits_{A_0} (\tau - 1/4) d\mu(x)$$

$$= mwb' + (\tau - 1/4)(1 - 2mw).$$

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

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

Notice that this choice of τ implies that for all $\boldsymbol{\omega} \in \{-1, 1\}^m$ the optimal threshold is given by $\theta^* = 1/4$. Moreover, if $mw \leq 1/2$ we can ensure that the value of $\tau \in [0, 1]$, that is, it is a valid choice for the regression function. Let us demonstrate that (the margin) Assumption 4 holds for an appropriate choice of m and w. Define $x_0 = (1/(2q), \ldots, 1/(2q))^\top$, then for every $\boldsymbol{\omega} \in \{-1, 1\}^m$ we have that $P_{\mathbf{X}}(0 < |\eta_{\boldsymbol{\omega}}(\mathbf{X}) - 1/4| \leq \delta)$ is equal to

$$\frac{2mw}{\mathrm{Leb}(\mathcal{B}(0, (4q)^{-1}))} \int\limits_{\mathcal{B}(\mathbf{x}_0, (4q)^{-1})} \mathbb{I}\{C_\varphi q^{-\beta} u(q||\mathbf{x} - n_q(\mathbf{x})||_2) \leq \delta\} d\mathbf{x}$$

$$+ \frac{1 - 2mw}{\mathrm{Leb}(A_0)} \int\limits_{A_0} \mathbb{I}\left\{\frac{1}{3} + \left(\frac{1}{12} - \frac{2b'}{3}\right)\left(\frac{2mw}{1 - 2mw}\right) - \frac{1}{4} \leq \delta\right\} d\mathbf{x}$$

$$= \frac{1 - 2mw}{\mathrm{Leb}(A_0)} \int\limits_{A_0} \mathbb{I}\left\{\frac{1}{12} + \left(\frac{1}{12} - \frac{2b'}{3}\right)\left(\frac{2mw}{1 - 2mw}\right) \leq \delta\right\} d\mathbf{x}$$

$$+ 2mw\mathbb{I}\{\delta \geq C_\varphi q^{-\beta}\}.$$

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

$$P_{\mathbf{X}}(0 < |\eta_{\boldsymbol{\omega}}(\mathbf{X}) - 1/4| \leq \delta) \leq 2mw\mathbb{I}\{\delta \geq C_\varphi q^{-\beta}\} + \mathbb{I}\left\{\delta \geq \frac{1}{12}\right\}$$

$$\leq 2mw\mathbb{I}\{\delta \geq C_\varphi q^{-\beta}\} + 12^\alpha \delta^\alpha.$$

Therefore, if mw is of order $q^{-\alpha\beta}$ the margin assumption is satisfied as long as $n^{-\beta/2\beta+d} \ln^2 n \leq 1/12$. The strong density assumption can be checked similarly to [2]. To finish the proof, for every $\boldsymbol{\omega} \in \{-1, 1\}^m$ we denote by $P^{\boldsymbol{\omega}}$ the distribution of $(\mathbf{X}, Y)$ with the marginal $P_{\mathbf{X}}$ and the regression function $\eta_{\boldsymbol{\omega}}$. Thus, one can write for any $\hat{g}$

$$\sup_{\mathbb{P} \in \mathcal{P}(\alpha, \beta)} \mathbf{E}[\mathcal{E}(\hat{g})] \geq \sup_{\boldsymbol{\omega} \in \{-1,1\}^m} \frac{1}{2}\mathbf{E}^{\boldsymbol{\omega}} \sum_{i=-m, i \neq 0}^{m} \mathbb{E}_{P_{\mathbf{X}}} |\varphi(\mathbf{X})| \mathbb{I}\{(1 + \mathrm{sgn}(i)\omega_i)/2 \neq \hat{g}(\mathbf{X})\} \mathbb{I}\{\mathbf{X} \in \mathcal{X}_i\}$$

$$\geq \sup_{\boldsymbol{\omega} \in \{-1,1\}^m} \frac{1}{2}\mathbf{E}^{\boldsymbol{\omega}} \sum_{i=1}^{m} \mathbb{E}_{P_{\mathbf{X}}} |\varphi(\mathbf{X})| \mathbb{I}\left\{\frac{1 + \omega_i}{2} \neq \hat{g}(\mathbf{X})\right\} \mathbb{I}\{\mathbf{X} \in \mathcal{X}_i\}, \qquad (15)$$

where $\mathbf{E}^{\boldsymbol{\omega}}$ is the expectation taken w.r.t. to the i.i.d. realizations of $\mathcal{D}_n^{\mathrm{L}}$ and $\mathcal{D}_N^{\mathrm{U}}$ from $P^{\boldsymbol{\omega}}$ and $P_{\mathbf{X}}$ respectively, $\operatorname{sgn}(i) = 1$ if $i > 0$ and $\operatorname{sgn}(i) = -1$ if $i < 0$, and to derive the last inequality we have dropped the summation over negative indices. Define the following pseudo-distance between two classifiers g, g':

$$d(g, g') = \frac{1}{2}\sum_{i=1}^{m} \mathbb{E}_{P_{\mathbf{X}}} |\varphi(\mathbf{X})| \mathbb{I}\{g(\mathbf{X}) \neq g'(\mathbf{X})\}\mathbb{I}\{\mathbf{X} \in \mathcal{X}_i\}.$$

For each $\boldsymbol{\omega} \in \Omega$ define a classifier $g_{\boldsymbol{\omega}}$ as

$$g_{\boldsymbol{\omega}}(\mathbf{x}) = \begin{cases} \frac{1+\omega_i}{2} & \text{if } \exists i = 1, \dots, m \text{ s.t. } x \in \mathcal{X}_i \\ 1 & \text{otherwise.} \end{cases}$$

Hence, Eq. (15) implies that

$$\inf_{\hat{g}} \sup_{\mathbb{P} \in \mathcal{P}(\alpha, \beta)} \mathbf{E}[\mathcal{E}(\hat{g})] \geq \inf_{\hat{g}} \sup_{\boldsymbol{\omega} \in \{-1,1\}^m} \mathbf{E}^{\boldsymbol{\omega}}[d(g_{\boldsymbol{\omega}}, \hat{g})]. \tag{16}$$

Our goal is to apply [35, Theorem 2.7]. First, let Ω be the set provided by Varshamov–Gilbert bound [35, Lemma 2.9] with $|\Omega| \geq 2^{\frac{m}{8}}$ and $\delta(\boldsymbol{\omega}, \boldsymbol{\omega}') \geq \frac{m}{4}$ for all $\boldsymbol{\omega}, \boldsymbol{\omega}' \in \Omega$. We note that for all $\boldsymbol{\omega}, \boldsymbol{\omega}' \in \Omega$ with $\boldsymbol{\omega} \neq \boldsymbol{\omega}'$ we have

$$d(g_{\boldsymbol{\omega}}, g_{\boldsymbol{\omega}'}) \geq 2\frac{C_\varphi q^{-\beta} m}{16} \quad \text{item (i) in [35, Theorem 2.7].}$$

To apply [35, Theorem 2.7] it remains to upper-bound the KL-divergence between any two fixed $\boldsymbol{\omega}, \bar{\boldsymbol{\omega}} \in \Omega$. Since the marginal distribution $\mathbb{P}_{\mathbf{X}}$ is independent from $\boldsymbol{\omega}$, we can write for product measures

$$\mathrm{KL}\left(P_{\mathbf{X}}^{\otimes N} \otimes (P^{\boldsymbol{\omega}})^{\otimes n}, P_{\mathbf{X}}^{\otimes N} \otimes (P^{\bar{\boldsymbol{\omega}}})^{\otimes n}\right) \leq n\mathrm{KL}(P^{\boldsymbol{\omega}}, P^{\bar{\boldsymbol{\omega}}}).$$

Furthermore, for some universal constant $C > 0$

$$\mathrm{KL}(P^{\boldsymbol{\omega}}, P^{\bar{\boldsymbol{\omega}}}) \leq 2\sum_{i=-m, i\neq 0}^{m} \mu\left(\varphi(\mathbf{X}) \log\left(\frac{1/4 + \varphi(\mathbf{X})}{1/4 - \varphi(\mathbf{X})}\right), \mathbf{X} \in \mathcal{X}_i\right) \leq Cq^{-2\beta} wm.$$

Fixing arbitrary $\bar{\boldsymbol{\omega}} \in \Omega$ we arrive for some universal C'' at

$$\frac{1}{|\Omega| - 1}\sum_{\boldsymbol{\omega} \in \Omega, \boldsymbol{\omega} \neq \bar{\boldsymbol{\omega}}} \mathrm{KL}\left(P_{\mathbf{X}}^{\otimes N} \otimes (P^{\boldsymbol{\omega}})^{\otimes n}, P_{\mathbf{X}}^{\otimes N} \otimes (P^{\bar{\boldsymbol{\omega}}})^{\otimes n}\right) \leq Cnq^{-\beta} wm = C'' nq^{-\beta} w \log(|\Omega| - 1).$$

Setting $q = \lfloor \bar{C} n^{\frac{1}{2\beta + d}} \rfloor, \quad w = C' q^{-d}$ for appropriately chosen $\bar{C}, C' > 0$ independent from n, we deduce that

$$\frac{1}{|\Omega| - 1}\sum_{\boldsymbol{\omega} \in \Omega, \boldsymbol{\omega} \neq \bar{\boldsymbol{\omega}}} \mathrm{KL}\left(P_{\mathbf{X}}^{\otimes N} \otimes (P^{\boldsymbol{\omega}})^{\otimes n}, P_{\mathbf{X}}^{\otimes N} \otimes (P^{\bar{\boldsymbol{\omega}}})^{\otimes n}\right) \leq \frac{1}{16}\log(|\Omega| - 1),$$

which corresponds to item (ii) in [35, Theorem 2.7]. Applying Theorem 2.7 from [35] with $s = \frac{C_\varphi q^{-\beta} m}{16}$ we get for some universal $c > 0$

$$\inf_{\hat{g}} \sup_{\boldsymbol{\omega} \in \{-1,1\}^m} \mathbf{E}^{\boldsymbol{\omega}}[d(g_{\boldsymbol{\omega}}, \hat{g})] \geq cC_\varphi q^{-\beta} m.$$

The proof is concluded after setting $m = \lfloor C''' q^{d - \alpha\beta} \rfloor$. Note that this choice of m is valid since we assumed that $\alpha\beta \leq d$. □

Proof of Theorem 12. The proof of this result follows similar scheme as that of the binary case. First of all observe that for all $b > 0$ the function $R(\theta) := b^2\theta - \sum_{k=1}^{K} \mathbb{E}(\eta_k(x) - \theta)_+$ is 1) continuous; 2) $R(0) = -1 < 0$, $R(1) = b^2 > 0$; 3) strictly increasing.[6] Hence, by the mean-value theorem there

[6] Indeed note that $R(\theta) = H_1(\theta) + H_2(\theta)$ with $H_1(\theta) = b^2\theta$ being strictly increasing and $H_2(\theta) = -\sum_{k=1}^{K} \mathbb{E}(\eta_k(x) - \theta)_+$ being non-decreasing.

is a θ^* such that $R(\theta^*) = 0$. Moreover, since R is increasing, such θ^* is unique, implying that Γ^* is well-defined.

Furthermore, for the F_b-score of Γ^* we can write

$$\begin{aligned}
\mathrm{F}_b(\Gamma^*) &= \frac{1+b^2}{b^2+\mathbb{E}|\Gamma^*(\mathbf{X})|}\mathbb{P}(Y \in \Gamma^*(\mathbf{X})) = \frac{1+b^2}{b^2+\mathbb{E}|\Gamma^*(\mathbf{X})|}\sum_{k=1}^{K}\mathbb{E}[\eta_k(\mathbf{X})\mathbb{I}\{k \in \Gamma^*(\mathbf{X})\}] \\
&\overset{(a)}{=} \frac{1+b^2}{b^2+\mathbb{E}|\Gamma^*(\mathbf{X})|}\sum_{k=1}^{K}\mathbb{E}[(\eta_k(\mathbf{X})-\theta^*)\mathbb{I}\{k \in \Gamma^*(\mathbf{X})\}] + \theta^*\mathbb{E}|\Gamma^*(\mathbf{X})|\frac{1+b^2}{b^2+\mathbb{E}|\Gamma^*(\mathbf{X})|} \\
&\overset{(b)}{=} \frac{1+b^2}{b^2+\mathbb{E}|\Gamma^*(\mathbf{X})|}\sum_{k=1}^{K}(\eta_k(\mathbf{X})-\theta^*)_+ + \theta^*\mathbb{E}|\Gamma^*(\mathbf{X})|\frac{1+b^2}{b^2+\mathbb{E}|\Gamma^*(\mathbf{X})|} \\
&\overset{(c)}{=} \frac{1+b^2}{b^2+\mathbb{E}|\Gamma^*(\mathbf{X})|}b^2\theta^* + \theta^*\mathbb{E}|\Gamma^*(\mathbf{X})|\frac{1+b^2}{b^2+\mathbb{E}|\Gamma^*(\mathbf{X})|} = (1+b^2)\theta^*,
\end{aligned}$$

where (a) is a consequence of $\mathbb{E}[\sum_{k=1}^{K}\mathbb{I}\{k \in \Gamma(\mathbf{X})\}] = \mathbb{E}|\Gamma(\mathbf{X})|$, (b) uses the definition of Γ^*, and (c) uses the definition of θ^*.

Finally, for any $\Gamma : \mathbb{R}^d \to 2^{[K]}$ we can write

$$\begin{aligned}
\frac{\mathrm{F}_b(\Gamma^*)-\mathrm{F}_b(\Gamma)}{1+b^2} &= \frac{\sum_{k=1}^{K}\mathbb{E}[\eta_k(\mathbf{X})\mathbb{I}\{k \in \Gamma^*(\mathbf{X})\}]}{b^2+\mathbb{E}|\Gamma^*(\mathbf{X})|} - \frac{\sum_{k=1}^{K}\mathbb{E}[\eta_k(\mathbf{X})\mathbb{I}\{k \in \Gamma(\mathbf{X})\}]}{b^2+\mathbb{E}|\Gamma(\mathbf{X})|} \\
&= \frac{\sum_{k=1}^{K}\mathbb{E}[\eta_k(\mathbf{X})(\mathbb{I}\{k \in \Gamma^*(\mathbf{X})\} - \mathbb{I}\{k \in \Gamma(\mathbf{X})\})]}{b^2+\mathbb{E}|\Gamma^*(\mathbf{X})|} \\
&+ \sum_{k=1}^{K}\mathbb{E}[\eta_k(\mathbf{X})\mathbb{I}\{k \in \Gamma(\mathbf{X})\}]\left(\frac{1}{b^2+\mathbb{E}|\Gamma^*(\mathbf{X})|} - \frac{1}{b^2+\mathbb{E}|\Gamma(\mathbf{X})|}\right) \\
&= \frac{\sum_{k=1}^{K}\mathbb{E}[(\eta_k(\mathbf{X})-\theta^*)_+\mathbb{I}\{k \in \Gamma^*(\mathbf{X})\triangle\Gamma(\mathbf{X})\}]}{b^2+\mathbb{E}|\Gamma^*(\mathbf{X})|} \\
&+ \theta^*\frac{\mathbb{E}|\Gamma^*(\mathbf{X})| - \mathbb{E}|\Gamma(\mathbf{X})|}{b^2+\mathbb{E}|\Gamma^*(\mathbf{X})|} + \frac{\mathrm{F}_b(\Gamma)}{1+b^2}\frac{\mathbb{E}|\Gamma(\mathbf{X})| - \mathbb{E}|\Gamma^*(\mathbf{X})|}{b^2+\mathbb{E}|\Gamma^*(\mathbf{X})|}.
\end{aligned}$$

We have already shown that $\theta^* = \mathrm{F}_b(\Gamma^*)/(1+b^2)$, hence

$$\begin{aligned}
\frac{\mathrm{F}_b(\Gamma^*)-\mathrm{F}_b(\Gamma)}{1+b^2} &= \frac{\sum_{k=1}^{K}\mathbb{E}[(\eta_k(\mathbf{X})-\theta^*)_+\mathbb{I}\{k \in \Gamma^*(\mathbf{X})\triangle\Gamma(\mathbf{X})\}]}{b^2+\mathbb{E}|\Gamma^*(\mathbf{X})|} \\
&+ \frac{\mathrm{F}_b(\Gamma^*)-\mathrm{F}_b(\Gamma)}{1+b^2}\left(\frac{\mathbb{E}|\Gamma^*(\mathbf{X})| - \mathbb{E}|\Gamma(\mathbf{X})|}{b^2+\mathbb{E}|\Gamma^*(\mathbf{X})|}\right).
\end{aligned}$$

Solving previous equation for $(\mathrm{F}_b(\Gamma^*) - \mathrm{F}_b(\Gamma))/(1+b^2)$, we conclude. □

REFERENCES

1. J.-Y. Audibert, *Progressive Mixture Rules are Deviation Suboptimal*, in NIPS, (2007), pp. 41–48.
2. J.-Y. Audibert and A. B. Tsybakov, "Fast learning rates for plug-in classifiers," Ann. Statist. **35** (2), 608–633 (2007).
3. H. Bao and M. Sugiyama, Calibrated surrogate maximization of linear-fractional utility in binary classification (2019), arXiv preprint arXiv:1905.12511.
4. M. Binkhonain and L. Zhao, A review of machine learning algorithms for identification and classification of non-functional requirements. Expert Systems with Applications: X, 1:100001 (2019).
5. E. Chzhen, C. Denis, and M. Hebiri, Minimax semi-supervised confidence sets for multi-class classification. preprint (2019). https://arxiv.org/abs/1904.12527.
6. S. Conte and C. Boor, *Elementary Numerical Analysis: An Algorithmic Approach* (McGraw-Hill Higher Education, 3rd edition, 1980).

7. J. Del Coz, J. Díez, and A. Bahamonde, "Learning nondeterministic classifiers," Journal of Machine Learning Research **10** (10) (2009).
8. K. Dembczynski, A. Jachnik, W. Kotlowski, W. Waegeman, and E. Hüllermeier, "Optimizing the f-measure in multi-label classification: Plugin rule approach versus structured loss minimization," in International conference on machine learning, 1130–1138 (2013).
9. K. Dembczynski, W. Waegeman, W. Cheng, and E. Hüllermeier, "An exact algorithm for f-measure maximization," Advances in neural information processing systems **24**, 1404–1412 (2011).
10. C. Denis and M. Hebiri, "Confidence sets with expected sizes for multiclass classification," JMLR **18** (1), 3571–3598 (2017).
11. L. Devroye, L. Györfi, and G. Lugosi, *A Probabilistic Theory of Pattern Recognition*, vol. 31 of *Applications of Mathematics* (Springer-Verlag, New York, 1996).
12. P. Flach, "Performance evaluation in machine learning: The good, the bad, the ugly, and the way forward," In Proceedings of the AAAI Conference on Artificial Intelligence **33**, 9808–9814 (2019).
13. S. Gadat, T. Klein, and C. Marteau, "Classification in general finite dimensional spaces with the k-nearest neighbor rule," Ann. Statist. **44** (3), 982–1009 (2016).
14. A. Gunawardana and G. Shani, "A survey of accuracy evaluation metrics of recommendation tasks," Journal of Machine Learning Research **10** (12) (2009).
15. M. Jansche, "Maximum expected f-measure training of logistic regression models," in Proceedings of the conference on human language technology and empirical methods in natural language processing, Association for Computational Linguistics, 692–699 (2005).
16. N. Japkowicz and M. Shah, *Evaluating Learning Algorithms: A Classification Perspective* (Cambridge University Press, 2011).
17. O. Koyejo, N. Natarajan, P. Ravikumar, and I. Dhillon, "Consistent binary classification with generalized performance metrics," in NIPS, 2744–2752 (2014).
18. S. Kpotufe and G. Martinet, "Marginal singularity, and the benefits of labels in covariate-shift," in Conference on Learning Theory, 1882–1886 (2018).
19. M. Lapin, M. Hein, and B. Schiele, "Top-k multiclass svm," in Advances in Neural Information Processing Systems, 325–333 (2015).
20. O. Mac Aodha, E. Cole, and P. Perona, "Presence-only geographical priors for fine-grained image classification," in Proceedings of the IEEE International Conference on Computer Vision, 9596–9606 (2019).
21. E. Mammen and A. B. Tsybakov, "Smooth discrimination analysis," Ann. Statist. **27** (6), 1808–1829 (1999).
22. D. R. Martin, C. C. Fowlkes, and J. Malik, "Learning to detect natural image boundaries using local brightness, color, and texture cues," IEEE transactions on pattern analysis and machine intelligence **26** (5), 530–549 (2004).
23. T. Mortier, M. Wydmuch, K. Dembczyński, E. Hüllermeier, and W. Waegeman, Efficient set-valued prediction in multi-class classification (2019), arXiv preprint arXiv:1906.08129.
24. D. R. Musicant, V. Kumar, A. Ozgur, et al. "Optimizing f-measure with support vector machines," in FLAIRS conference, 356–360 (2003).
25. H. Narasimhan, R. Vaish, and S. Agarwal, "On the statistical consistency of plug-in classifiers for non-decomposable performance measures," in NIPS, 1493–1501 (2014).
26. S. P. Parambath, N. Usunier, and Y. Grandvalet, "Optimizing f-measures by cost-sensitive classification," in Advances in Neural Information Processing Systems, 2123–2131 (2014).
27. W. Polonik, "Measuring mass concentrations and estimating density contour clusters-an excess mass approach," Ann. Statist. **23** (3), 855–881 (1995).
28. H. G. Ramaswamy, A. Tewari, S. Agarwal, et al. "Consistent algorithms for multiclass classification with an abstain option," Electronic Journal of Statistics **12** (1), 530–554 (2018).
29. P. Rigollet, R. Vert, et al., "Optimal rates for plug-in estimators of density level sets," Bernoulli **15** (4), 1154–1178 (2009).
30. M. Sadinle, J. Lei, and L. Wasserman, "Least ambiguous set-valued classifiers with bounded error levels," Journal of the American Statistical Association **114** (525), 223–234 (2019).
31. C. Scott, "Calibrated asymmetric surrogate losses," Electronic Journal of Statistics **6**, 958–992 (2012).
32. C. J. Stone, "Consistent nonparametric regression," The annals of statistics, 595–620 (1977).
33. E. F. Tjong Kim Sang and F. De Meulder, "Introduction to the conll-2003 shared task: Language-independent named entity recognition," in *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003*, Vol. 4, pp. 142–147.

34. A. B. Tsybakov, "Optimal aggregation of classifiers in statistical learning," Ann. Statist. **32** (1), 135–166 (2004).

35. A. B. Tsybakov, *Introduction to Nonparametric Estimation. Springer Series in Statistics* (Springer, New York, 2009).

36. V. Vovk, I. Nouretdinov, V. Fedorova, I. Petej, and A. Gammerman, "Criteria of efficiency for set-valued classification," Annals of Mathematics and Artificial Intelligence **81**, 21–47 (2017).

37. W. Waegeman, K. Dembczyński, A. Jachnik, W. Cheng, and E. Hüllermeier, "On the bayes-optimality of f-measure maximizers," Journal of Machine Learning Research **15**, 3333–3388 (2014).

38. B. Yan, S. Koyejo, K. Zhong, and P. Ravikumar, Binary classification with karmic, threshold-quasi-concave metrics, in ICML, vol. 80 (2018).

39. Y. Yang, "Minimax nonparametric classification: Rates of convergence," IEEE Transactions on Information Theory **45** (7), 2271–2284 (1999).

40. M.-J. Zhao, N. Edakunni, A. Pocock, and G. Brown, "Beyond fano's inequality: bounds on the optimal f-score, ber, and cost-sensitive risk and their implications," JMLR **14**(Apr), 1033–1090 (2013).