



**HAL**  
open science

# Consistency of plug-in confidence sets for classification in semi-supervised learning

Christophe Denis, Mohamed Hebiri

► **To cite this version:**

Christophe Denis, Mohamed Hebiri. Consistency of plug-in confidence sets for classification in semi-supervised learning. 2015. hal-01180363

**HAL Id: hal-01180363**

**<https://hal.science/hal-01180363v1>**

Preprint submitted on 26 Jul 2015

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# Consistency of plug-in confidence sets for classification in semi-supervised learning

Christophe Denis\* and Mohamed Hebiri†

LAMA, UMR-CNRS 8050,  
Université Paris Est – Marne-la-Vallée

## Abstract

Confident prediction is highly relevant in machine learning; for example, in applications such as medical diagnoses, wrong prediction can be fatal. For classification, there already exist procedures that allow to not classify data when the confidence in their prediction is weak. This approach is known as classification with reject option. In the present paper, we provide new methodology for this approach. Predicting a new instance via a confidence set, we ensure an exact control of the probability of classification. Moreover, we show that this methodology is easily implementable and entails attractive theoretical and numerical properties.

*Keywords* : Classification, classification with reject option, conformal predictors, confidence sets, plug-in confidence sets.

## 1 Introduction

Binary classification aims at assigning a label  $Y \in \{0, 1\}$  to a given example  $X \in \mathcal{X}$ . The goal is then to build a classification rule  $s : \mathcal{X} \rightarrow \{0, 1\}$  so that  $s(X)$ , the predicted label for the observed example  $X$  is as close as possible to the label  $Y$ . In this framework the question of confident prediction, which results in wondering how accurate is the prediction  $s(X)$ , becomes a central question. Doubts about the confidence of the predicted label  $s(X)$  may arise in these situations: if the conditional probability  $\eta^*(x) = \mathbb{P}(Y = 1|X = x)$  is close to  $1/2$  so that the feature  $x$  might be hard to classify whatever the classification rule is; or, if the classification rule is inefficient. In such cases, it is worth considering procedures that allow to not classify an observation when the doubt is too important. We talk about *classification with reject option*. This setting is particularly relevant in some applications where wrong classification may lead to big issues: it is hence better to not assign a label rather than to assign a non confident one. Procedures for classification with reject option has been studied by several authors [Cho70, NZH10, GRKC09, VGS99, VGS05, HW06, BW08, WY11, Lei14] and references therein. In this context, two questions arise: how to determine whether we should classify an example or not; and how to take into account the reject option? The works on classification with reject option can be separated according to two approaches.

i) The works which rely on the *conformal predictors* algorithm [VGS99, VGS05]. The general idea of conformal prediction is to build for a given feature  $X$  a set  $\Gamma(X)$ , which takes its value

---

\*Christophe.Denis@u-pem.fr

†Mohamed.Hebiri@u-pem.fr

in  $\{\emptyset, \{0\}, \{1\}, \{0, 1\}\}$ , and contains the true label with high probability. The feature  $X$  is not classified if  $\text{card}(\Gamma(X)) \neq 1$ . One of the most important ideas behind the construction of conformal predictors is the notion of conformity. More precisely, the value of  $\Gamma(X)$  depends on the similarity between the example  $X$  and an already collected labeled dataset. Then, the procedure uses local arguments and can be seen as a transductive method [GKKW02, Vap98]. In terms of performance, the set  $\Gamma(X)$  is built in order to control the overall misclassification risk: for a given significance level  $\varepsilon \in (0, 1)$ , the set  $\Gamma(X)$  satisfies<sup>1</sup>  $\mathbb{P}(Y \notin \Gamma(X)) \leq \varepsilon$ . The major drawback of the conformal prediction approach is that it does not take into account the reject option in the risk. Moreover, if the significance level  $\varepsilon$  is too small, the resulting set  $\Gamma(X)$  belongs to  $\{\emptyset, \{0, 1\}\}$  for all  $X$ . Hence, the use of reject option is irrelevant.

ii) The other works rely on the setting provided in [Cho70, HW06, BW08, WY11]. In this case, a *classification rule with reject option*  $s_R$  takes its values in  $\{0, 1, R_e\}$ , where  $s_R(X) = R_e$  means reject: no label is affected to the instance  $X$ . In the above mentioned works, rejecting is viewed as an error and for a fixed value of some parameter  $\alpha \in [1/2, 1]$ , the cost of the rejection is  $1 - \alpha$ . Therefore, the risk function associated with a classifier with reject option is given by  $L_\alpha(s_R) = \mathbb{P}(\{s_R(X) \neq Y\} \text{ and } \{X \text{ is classified}\}) + (1 - \alpha)\mathbb{P}(\{X \text{ is rejected}\})$ . The results provided in [Cho70] illustrate that the optimal reject procedure for  $L_\alpha$  is given by

$$s_{R_\alpha}^*(X) = \begin{cases} 0 & \text{if } \eta^*(X) \leq 1 - \alpha, \\ 1 & \text{if } \eta^*(X) \geq \alpha, \\ R_e & \text{otherwise.} \end{cases}$$

Herbei and Wegkamp [HW06] study the asymptotic optimality of procedures based on plug-in rules or on empirical risk minimization. We address some limits of this approach. First, the choice of the parameter  $\alpha$  is fundamental for the procedure and fixing it is tricky. As an immediate consequence, if the value of parameter  $\alpha$  is either too small or too large, the use of reject option can be irrelevant. Moreover, this approach does not allow to control any of the two parts of the risk function, in particular the rejection probability. Hence, comparing two classifiers with reject option in terms of the risk function  $L_\alpha$  remains difficult to interpret: they do not have necessarily the same rejection probabilities.

Both approaches previously presented bring into play the reject option through rather a set (conformal predictor) or a classifier with reject option. However, none provides a control on the probability of classifying a feature. In the present paper we consider a new way to tackle the problem of classification with reject option. We aim at controlling the rejection probability and at bounding the misclassification risk restricted to the set of label examples. Both considerations are new. For a given classifier  $s$  and a feature  $X$ , our methodology involves a statistical procedure which provides a set  $\Gamma_s(X) \in \{\{0\}, \{1\}, \{0, 1\}\}$ , namely a *confidence set*. We introduce in the present work oracle confidence sets, says  $\Gamma^*$ , which relies on a score function deduced from  $\eta^*$  and also on its cumulative distribution function. The main characteristic of oracle confidence sets is that they are able to control exactly the rejection probability  $\mathbb{P}(\Gamma^*(X) = \{0, 1\})$ : under mild assumption, we get level- $\varepsilon$ -confidence sets. These sets are called  $\varepsilon$ -confidence sets. This aspect makes our procedure prevent irrelevant use of reject option. Hence we do not view the reject as an error but simply as a parameter that we are able to control; moreover, we evaluate the quality of a confidence set through the misclassification risk conditionally on the set of classified examples. That is, for a given classification rule  $s$ , we focus on the control of the risk function  $\mathcal{R}(\Gamma_s) = \mathbb{P}(Y \in \Gamma_s(X) \mid \{X \text{ is classified}\})$ . To the best of our knowledge, none of the earlier

<sup>1</sup>In the terminology of conformal predictors, both of the outputs  $\emptyset$  and  $\{0, 1\}$  mean that no label is assigned. Both are important to be able to guarantee the exact control of the overall misclassification risk regardless of the classification rule used to build the set  $\Gamma(X)$ .

works provides a control of this risk neither of the rejection probability. According to the risk function  $\mathcal{R}$ , for  $\varepsilon \in ]0, 1]$ , the  $\varepsilon$ -confidence sets are shown to be optimal over the set of all confidence sets with rejection probability equal to  $1 - \varepsilon$ . Another contribution of the paper is to provide an algorithm which involves a consistent estimator of  $\eta^*$  and yields a confidence set. For a given level  $\varepsilon \in ]0, 1]$ , we do not build only a single algorithm of constructing asymptotically level- $\varepsilon$ -confidence sets, but a general device that takes as input a consistent estimator of the regression function and a unlabeled sample, and produces as output a confidence set which is provably asymptotically of level  $\varepsilon$  and consistent (i.e., the excess risk tends to zero). The resulting confidence set is referred as plug-in  $\varepsilon$ -confidence set. Furthermore, we establish rates of convergence under the Tsybakov noise assumption on the data generating distribution. Moreover, these confidence sets have the advantage of being easily implementable.

The rest of the paper is organized as follows. The definition and the important properties of the  $\varepsilon$ -confidence sets are provided in Section 2. We also apply the  $\varepsilon$ -confidence sets in the Gaussian mixture model. Section 3 is devoted to the introduction of the plug-in  $\varepsilon$ -confidence sets and their asymptotic behavior. We present a numerical illustration of our results in Section 4. We finally draw some conclusions and present perspectives of our work in Section 5. Proofs of our results are postponed to the Appendix.

*Notation:* First, we state general notation. Let  $(X, Y)$  be the generic data-structure taking its values in  $\mathcal{X} \times \{0, 1\}$  with distribution  $\mathbb{P}$ . Let  $(X_\bullet, Y_\bullet)$  be a random variable independent of  $(X, Y)$  and with the same law as  $(X, Y)$ . The goal in classification is to predict the label  $Y_\bullet$  given an observation of  $X_\bullet$ . This is performed based on a classifier (or classification rule)  $s$  which is a function mapping  $\mathcal{X}$  onto  $\{0, 1\}$ . Let  $\mathcal{S}$  be the set of all classifiers. The misclassification risk  $R$  associated with  $s \in \mathcal{S}$  is defined as

$$R(s) = \mathbb{P}(s(X) \neq Y).$$

Moreover, the minimizer of  $R$  over  $\mathcal{S}$  is the Bayes classifier, denoted by  $s^*$ , and is characterized by

$$s^*(\cdot) = \mathbf{1}\{\eta^*(\cdot) \geq 1/2\},$$

where  $\eta^*(x) = \mathbb{P}(Y = 1|X = x)$  for  $x \in \mathcal{X}$ . One of the most important quantities in our methodology is the function  $f^*$  defined by  $f^*(\cdot) = \max\{\eta^*(\cdot), 1 - \eta^*(\cdot)\}$ . It will play the role of a score function.

Let us now consider more specific notation related to the classification with reject option setting. Let  $s \in \mathcal{S}$  be a classifier. A confidence set  $\Gamma_s$  associated with the classifier  $s$  is defined as a measurable function that maps  $\mathcal{X}$  onto  $\{\{0\}, \{1\}, \{0, 1\}\}$ , such that for an example  $X_\bullet$ , the set  $\Gamma_s(X_\bullet)$  can be either  $\{s(X_\bullet)\}$  or  $\{0, 1\}$ . We decide to classify the example  $X_\bullet$ , according to the label  $s(X_\bullet)$ , if  $\text{card}(\Gamma_s(X_\bullet)) = 1$ . In the case where  $\Gamma_s(X_\bullet) = \{0, 1\}$ , we decide to not classify (reject) the feature  $X_\bullet$ . Let  $\Gamma_s$  be a confidence set. The probability of classifying a feature is denoted by

$$\mathcal{P}(\Gamma_s) := \mathbb{P}(\text{card}(\Gamma_s(X_\bullet)) = 1). \quad (1)$$

In our approach, the probability of classifying  $\mathcal{P}(\Gamma_s)$  is not viewed as a success or an error but simply as a parameter that we have to control. Hence, the definition of a confidence set makes natural the following definition of the risk associated with  $\Gamma_s$ :

$$\begin{aligned} \mathcal{R}(\Gamma_s) &= \mathbb{P}(Y_\bullet \notin \Gamma_s(X_\bullet) | \text{card}(\Gamma_s(X_\bullet)) = 1) \\ &= \mathbb{P}(s(X_\bullet) \neq Y_\bullet | \text{card}(\Gamma_s(X_\bullet)) = 1). \end{aligned} \quad (2)$$

The risk  $\mathcal{R}(\Gamma_s)$  is the misclassification error risk of  $s$  conditional to the event that  $X_\bullet$  is classified. Moreover, for some  $\varepsilon \in ]0, 1]$ , we say that, for two confidence sets  $\Gamma_s$  and  $\Gamma_{s'}$  such that  $\mathcal{P}(\Gamma_s) = \mathcal{P}(\Gamma_{s'}) = \varepsilon$ , the confidence set  $\Gamma_s$  is “better” than  $\Gamma_{s'}$  if  $\mathcal{R}(\Gamma_s) \leq \mathcal{R}(\Gamma_{s'})$ .

## 2 $\varepsilon$ -confidence sets

In the present section, we define a class of confidence sets referred as  $\varepsilon$ -confidence sets which are optimal according to the definition of risk (2). We always keep in mind that the classification probability (1) will be under control. In Section 2.1, we define and state the important properties of the class of  $\varepsilon$ -confidence sets. We then apply the  $\varepsilon$ -confidence sets to the Gaussian mixture case in Section 2.2. We end up this section with a comparison to classifiers with reject option in Section 2.3.

### 2.1 Definition and properties

The definition of  $\varepsilon$ -confidence sets relies on the Bayes classifier  $s^*$  and the cumulative distribution function of  $f^*(X)$ .

**Definition 1.** Let  $\varepsilon \in ]0, 1[$ , the  $\varepsilon$ -confidence set is defined as follows

$$\Gamma_\varepsilon^\bullet(X_\bullet) = \begin{cases} \{s^*(X_\bullet)\} & \text{if } F_f^*(f^*(X_\bullet)) \geq 1 - \varepsilon \\ \{0, 1\} & \text{otherwise,} \end{cases}$$

where  $F_f^*$  is the cumulative distribution function of  $f^*(X)$  and  $f^*(\cdot) = \max\{\eta^*(\cdot), 1 - \eta^*(\cdot)\}$ .

According to this definition, the construction of the  $\varepsilon$ -confidence sets relies on two important features. First, if a label is assigned to a new feature  $X_\bullet$  by the  $\varepsilon$ -confidence set, it is the one provided by the Bayes classifier  $s^*$ . Second, we assign a label to a new data  $X_\bullet$  if the corresponding score  $f^*(X_\bullet)$  is large enough regarding the distribution of  $f^*(X)$ . This is one of the key ideas behind conformal predictors introduced in [VGS05].

The following assumption is fundamental to establish theoretical guarantees.

**(A1)** The cumulative distribution function  $F_f^*$  of  $f^*(X)$  is continuous.

One of the main motivations of the introduction of the  $\varepsilon$ -confidence set is that, if Assumption (A1) holds, the procedure ensures an exact control of the probability (1) of assigning a label

$$\mathcal{P}(\Gamma_\varepsilon^\bullet) = \mathbb{P}(F_f^*(f^*(X_\bullet)) \geq 1 - \varepsilon) = \varepsilon. \quad (3)$$

This happens since  $F_f^*(f^*(X_\bullet))$  is uniformly distributed under Assumption (A1). Moreover, under this assumption as well, one can rewrite the  $\varepsilon$ -confidence sets in a different way. Indeed, for  $\varepsilon \in ]0, 1[$ , we have

$$F_f^*(f^*(X_\bullet)) \geq 1 - \varepsilon \iff f^*(X_\bullet) \geq (F_f^*)^{-1}(1 - \varepsilon),$$

where  $(F_f^*)^{-1}$  denotes the generalized inverse of the cumulative distribution function  $F_f^*$  (see [vdV98]). Therefore, if we set  $\alpha_\varepsilon = (F_f^*)^{-1}(1 - \varepsilon)$  for  $\varepsilon \in ]0, 1[$  and  $\alpha_1 = 1/2$ , Definition 1 is equivalent to

$$\Gamma_\varepsilon^\bullet(X_\bullet) = \begin{cases} \{s^*(X_\bullet)\} & \text{if } f^*(X_\bullet) \geq \alpha_\varepsilon \\ \{0, 1\} & \text{otherwise.} \end{cases} \quad (4)$$

Next, we provide the most important property of the  $\varepsilon$ -confidence sets:

**Proposition 1.** Denote by  $\Gamma_\varepsilon$  the set  $\Gamma_\varepsilon = \{\Gamma_s; \mathcal{P}(\Gamma_s) = \varepsilon\}$ . Let Assumption (A1) be satisfied.

1. For any  $\varepsilon \in ]0, 1[$ , the  $\varepsilon$ -confidence set satisfies the following property:

$$\mathcal{R}(\Gamma_\varepsilon^\bullet) = \min_{\Gamma_s \in \Gamma_\varepsilon} \mathcal{R}(\Gamma_s).$$

2. For  $\varepsilon \in ]0, 1]$  and for any  $\Gamma_s \in \mathbf{\Gamma}_\varepsilon$ , the following holds

$$0 \leq \mathcal{R}(\Gamma_s) - \mathcal{R}(\Gamma_\varepsilon^\bullet) = \frac{1}{\varepsilon} \left\{ \mathbb{E} [ |2\eta^*(X_\bullet) - 1| \mathbf{1}_C ] + \mathbb{E} [ |\eta^*(X_\bullet) - \alpha_\varepsilon| \mathbf{1}_{\mathcal{A}_0 \cup \mathcal{B}_0} ] + \mathbb{E} [ |1 - \eta^*(X_\bullet) - \alpha_\varepsilon| \mathbf{1}_{\mathcal{A}_1 \cup \mathcal{B}_1} ] \right\}, \quad (5)$$

where  $\alpha_\varepsilon = (F_f^*)^{-1}(1 - \varepsilon)$  and

$$\begin{aligned} \mathcal{A}_y &= \{f^*(X_\bullet) \geq \alpha_\varepsilon, \text{card}(\Gamma_s(X_\bullet)) \neq 1, s^*(X_\bullet) \neq y\}, \quad y = 0, 1, \\ \mathcal{B}_y &= \{f^*(X_\bullet) < \alpha_\varepsilon, \text{card}(\Gamma_s(X_\bullet)) = 1, s(X_\bullet) \neq y\}, \quad y = 0, 1, \\ \mathcal{C} &= \{f^*(X_\bullet) \geq \alpha_\varepsilon, \text{card}(\Gamma_s(X_\bullet)) = 1, s^*(X_\bullet) \neq s(X_\bullet)\}. \end{aligned}$$

Several remarks can be made from Proposition 1. First, for  $\varepsilon \in ]0, 1]$ , the  $\varepsilon$ -confidence set is optimal in the sense that its risk is minimal over  $\mathbf{\Gamma}_\varepsilon$ , the class of all confidence sets that assign a label with probability  $\varepsilon$ . Second, the excess risk of a confidence set is directly linked to the behavior of the function  $f^*$  around  $\alpha_\varepsilon$ . This observation will play a major role in our main result related to rates of convergence in the next section. Third, note that if we apply (5) with  $\varepsilon = 1$ , which implies  $\alpha_\varepsilon = 1/2$ , we obtain the classical result in classification

$$R(s) - R(s^*) = \mathbb{E} [ |2\eta^*(X_\bullet) - 1| \mathbf{1}_{\{s^*(X_\bullet) \neq s(X_\bullet)\}} ].$$

Let us conclude this section by stating a result that specifies the behavior of the risk associated with the  $\varepsilon$ -confidence set w.r.t. the parameter  $\varepsilon$ :

**Proposition 2.** *The function  $\varepsilon \mapsto \mathcal{R}(\Gamma_\varepsilon^\bullet)$  is non decreasing on  $]0, 1]$ .*

This result shows an expected fact: the larger the rejecting probability, the smaller the error. In particular

$$\mathcal{R}(\Gamma_\varepsilon^\bullet) \leq R(s^*) \quad \forall \varepsilon \in ]0, 1].$$

## 2.2 $\varepsilon$ -confidence sets for Gaussian mixture

In this section, we apply the  $\varepsilon$ -confidence set introduced in Definition 1 to the particular case of Gaussian mixture model. We set  $\mathcal{X} = \mathbb{R}^d$  with  $d \in \mathbb{N} \setminus \{0\}$ . Let us assume that the conditional distribution of  $X$  given  $Y$  is Gaussian and that, for simplicity, the marginal distribution of  $Y$  is Bernoulli with parameter  $1/2$ . To fix notation, we set

$$X|Y = 0 \sim \mathcal{N}(\mu_0, \Sigma) \quad \text{and} \quad X|Y = 1 \sim \mathcal{N}(\mu_1, \Sigma),$$

where  $\mu_0$  and  $\mu_1$  are vectors in  $\mathbb{R}^d$  and  $\Sigma$  is the common covariance matrix. We assume that  $\Sigma$  is invertible and denote by  $\|\cdot\|_{\Sigma^{-1}}$  the norm under  $\Sigma^{-1}$ : for any  $\mu \in \mathbb{R}^d$  we have  $\|\mu\|_{\Sigma^{-1}}^2 = \mu^\top \Sigma^{-1} \mu$  where  $\mu^\top$  stands for the transpose of  $\mu$ . The following theorem establishes the classification error of the  $\varepsilon$ -confidence set  $\Gamma_\varepsilon^\bullet$  in this framework.

**Proposition 3.** *For all  $\varepsilon \in ]0, 1]$ , we have*

$$\mathcal{R}(\Gamma_\varepsilon^\bullet) = \frac{\mathbb{P}(\Phi(Z) + \Phi(Z + \|\mu_1 - \mu_0\|_{\Sigma^{-1}}) \leq \varepsilon)}{\varepsilon},$$

where  $Z$  is a standard normal random variable and  $\Phi$  is the standard normal cumulative distribution function.

The proof of this proposition is postponed to the Appendix. Interestingly, in the Gaussian mixture case, we get a close formula for the risk of the  $\varepsilon$ -confidence set. Moreover, this risk depends on  $\|\mu_1 - \mu_0\|_{\Sigma^{-1}}$  as in the binary classification framework which corresponds to the particular case  $\varepsilon = 1$ , where we do not use the reject option and where we get

$$\mathcal{R}(\Gamma_1^\bullet) = R(s^*) = 1 - \Phi\left(\frac{\|\mu_1 - \mu_0\|_{\Sigma^{-1}}}{2}\right).$$

### 2.3 Relation with classifiers with reject option

The problem of classification with reject option has already been introduced in [Cho70]. More recently the terminology of *classifiers with reject option* has been defined in [HW06]: a classifier with reject option is a measurable function which maps  $\mathcal{X}$  onto  $\{0, 1, R_e\}$  where the output  $R_e$  means reject. For a parameter  $\alpha \in [1/2, 1]$  and for  $s_R$  a classifier with reject option, the risk function considered in [HW06] is

$$L_\alpha(s_R) = \mathbb{P}(s_R(X) \neq Y, s_R(X) \neq R_e) + (1 - \alpha)\mathbb{P}(s_R(X) = R_e). \quad (6)$$

This risk has been studied in the context of classification with reject option in the papers [HW06, WY11] and references therein. We notice that the above risk looks at rejecting as a part of the error in the same way as wrong classification. The parameter  $1 - \alpha$  controls the trade-off between these two "errors". In other words, the parameter  $1 - \alpha$  is the cost of using the reject option. This is a major difference with our point of view. Indeed, we recall that the probability of rejection is a parameter in our setting, and then we do not include it in the risk (2). As a consequence, if we bound the risk (2) while keeping under control the probability of classifying (1), we are able to bound the risk (6). The reverse is not true. That is, controlling (6) does not provide any control on the probability of rejection and one cannot avoid irrelevant use of the reject option in some situations. This difference can be significant in some practical situations where the knowledge of  $\mathcal{P}(\Gamma_s)$  is a relevant information. Indeed, when dealing with several examples to label, controlling this probability ensures the amount of the data we wish label. Hence our methodology prevents from irrelevant use of the reject option. For the same reason, a second important feature that differs between both methodologies is that the comparison between two confidence sets is easier than the comparison between two classifiers with reject option. Indeed, for some  $\varepsilon \in ]0, 1]$ , we say that, for two confidence sets  $\Gamma_s$  and  $\Gamma_{s'}$  such that  $\mathcal{P}(\Gamma_s) = \mathcal{P}(\Gamma_{s'}) = \varepsilon$ , the confidence set  $\Gamma_s$  is "better" than  $\Gamma_{s'}$  if  $\mathcal{R}(\Gamma_s) \leq \mathcal{R}(\Gamma_{s'})$ . As the study of the risk function  $L_\alpha$  viewed in Equation (6) does not provide any control on the probability of classifying, it is much more difficult to compare the performance of two classifiers with reject option on the set of labeled data. This point will be made clear in Section 4 with the numerical experiment.

Let us consider the optimality by now: the paper [HW06] also provides the optimal rule for the risk (6). For each  $\alpha \in [1/2, 1]$ , the *Bayes rule with reject option*  $s_{R_\alpha}^*$  is defined such that

$$L_\alpha(s_{R_\alpha}^*) = \min_{s_R} L_\alpha(s_R),$$

where the minimum is taken over all classifiers with reject option. It is characterized by

$$s_{R_\alpha}^*(X_\bullet) = \begin{cases} s^*(X_\bullet) & \text{if } f^*(X_\bullet) \geq \alpha, \\ R_e & \text{otherwise,} \end{cases} \quad (7)$$

where  $f^*(\cdot) = \max\{\eta^*(\cdot), 1 - \eta^*(\cdot)\}$  as in our setting. Obviously, the *Bayes rule with reject option*  $s_{R_\alpha}^*$  can be written in term of confidence sets. This leads to an  $\varepsilon$ -confidence set defined in the

same way as in Equation (4). However, there is an important difference: the main contribution of the present paper is to provide a methodology to pick the parameter  $\alpha_\varepsilon$  in (4) such that the probability of classifying an example (1) is exactly  $\varepsilon$ . The key to be able to do so is the use of the cumulative distribution function of  $f^*(X)$ . In Section 3.1, we will see that the data-driven counterpart of the  $\varepsilon$ -confidence defined in Definition 1 also controls the probability of classifying. Notably, this is possible in a semi-supervised way, that is, only using a set of unlabeled data.

### 3 Plug-in $\varepsilon$ -confidence sets

This section is devoted to the study the data driven counterpart of the  $\varepsilon$ -confidence sets provided by plug-in rule. We provide the construction of the plug-in methods in Section 3.1. Their asymptotic consistency as well as rates of convergence are given in Section 3.2.

#### 3.1 Definition of the plug-in $\varepsilon$ -confidence sets

For  $\varepsilon \in ]0, 1]$ , the construction of our plug-in  $\varepsilon$ -confidence set relies on a previous estimator of the regression function  $\eta^*$ . To this end, we introduce a first dataset,  $\mathcal{D}_n$ , which consists of  $n$  independent copies of  $(X, Y)$ . The dataset  $\mathcal{D}_n$  is used to estimate the function  $\eta$  and therefore the functions  $f^*$  and  $s^*$  as well. Let us denote by  $\hat{\eta}$ ,  $\hat{f}(\cdot) = \max(\hat{\eta}(\cdot), 1 - \hat{\eta}(\cdot))$  and  $\hat{s} = \mathbf{1}_{\{\hat{\eta}(\cdot) \geq 1/2\}}$  the estimators of  $\eta^*$ ,  $f^*$  and  $s^*$  respectively. Thanks to these estimations, a data driven approximation of the  $\varepsilon$ -confidence set given in Definition 1 can be

$$\tilde{\Gamma}_\varepsilon^\bullet(X_\bullet) = \begin{cases} \{\hat{s}(X_\bullet)\} & \text{if } F_{\hat{f}}(\hat{f}(X_\bullet)) \geq 1 - \varepsilon \\ \{0, 1\} & \text{otherwise,} \end{cases} \quad (8)$$

where  $F_{\hat{f}}$  is the cumulative distribution function of  $\hat{f}(X)$ . Hence,  $\tilde{\Gamma}_\varepsilon^\bullet(X_\bullet)$  invokes the cumulative distribution function  $F_{\hat{f}}$ , which is unknown and therefore needs to be estimated. We then consider a second dataset, independent of  $\mathcal{D}_n$ , denoted by  $\mathcal{D}_N = \{X_i, i = 1, \dots, N\}$  where  $X_1, \dots, X_N$  are independent copies of  $X$ . Based on  $\mathcal{D}_N$ , we estimate the cumulative function  $F_{\hat{f}}$  by the empirical cumulative distribution function of  $\hat{f}(X)$  denoted by  $\hat{F}_{\hat{f}}$ . Now, we can define the plug-in  $\varepsilon$ -confidence set:

**Definition 2.** *Let  $\varepsilon \in ]0, 1]$  and  $\hat{\eta}$  be any estimator of  $\eta^*$ , the plug-in  $\varepsilon$ -confidence set is defined as follows:*

$$\hat{\Gamma}_\varepsilon^\bullet(X_\bullet) = \begin{cases} \{\hat{s}(X_\bullet)\} & \text{if } \hat{F}_{\hat{f}}(\hat{f}(X_\bullet)) \geq 1 - \varepsilon \\ \{0, 1\} & \text{otherwise,} \end{cases}$$

where  $\hat{f}(\cdot) = \max\{\hat{\eta}(\cdot), 1 - \hat{\eta}(\cdot)\}$  and  $\hat{F}_{\hat{f}}(\hat{f}(X_\bullet)) = \frac{1}{N} \sum_{i=1}^N \mathbf{1}_{\{\hat{f}(X_i) \leq \hat{f}(X_\bullet)\}}$ .

**Remark 1.** *The samples  $\mathcal{D}_n$  and  $\mathcal{D}_N$  play completely different roles. The sample  $\mathcal{D}_n$  is used to estimate  $\eta^*$  and then must consist of labeled observations. The second dataset  $\mathcal{D}_N$ , used to estimate the cumulative function  $F_{\hat{f}}$ , requires only a set of unlabeled observations. Hence, the construction of the plug-in  $\varepsilon$ -confidence sets does not require more labeled examples than in the classical classification setting. This is particularly interesting in some practical situations where the number of labeled examples is small while a large number of unlabeled observations is available.*



### 3.2 Theoretical performance

This section is devoted to assessing the asymptotic performances of the plug-in  $\varepsilon$ -confidence set. The symbols  $\mathbf{P}$  and  $\mathbf{E}$  stand for generic probability and expectation, respectively. Let  $\varepsilon \in ]0, 1]$ , and  $\widehat{\Gamma}_\varepsilon^\bullet$  be a plug-in  $\varepsilon$ -confidence set. We define the risk of  $\widehat{\Gamma}_\varepsilon^\bullet$  by the natural quantity

$$\mathbf{R} \left( \widehat{\Gamma}_\varepsilon^\bullet \right) = \mathbf{P} \left( \hat{s}(X_\bullet) \neq Y_\bullet | \widehat{F}_{\hat{f}}(\hat{f}(X_\bullet)) \geq 1 - \varepsilon \right). \quad (9)$$

Note that we use here the notation  $\mathbf{R}$  rather than the previous one  $\mathcal{R}$  given by (2) to stress that the probability  $\mathbf{P}$  is taken under the law of  $(\mathcal{D}_n, \mathcal{D}_N, (X_\bullet, Y_\bullet))$  instead of just  $(X_\bullet, Y_\bullet)$ . Throughout this section we assume the following condition on the cumulative distribution function  $F_{\hat{f}}$ , which is analogous to Assumption (A1). However, this assumption relies on the estimator  $\hat{f}$  and then is not restrictive since it can be chosen by the statistician.

**(A2)** *The cumulative distribution function  $F_{\hat{f}}$  of  $\hat{f}(X)$  is continuous.*

We also define the risk of the oracle counterpart  $\widetilde{\Gamma}_\varepsilon^\bullet$  of  $\widehat{\Gamma}_\varepsilon^\bullet$

$$\mathbf{R} \left( \widetilde{\Gamma}_\varepsilon^\bullet \right) = \mathbf{P} \left( \hat{s}(X_\bullet) \neq Y_\bullet | F_{\hat{f}}(\hat{f}(X_\bullet)) \geq 1 - \varepsilon \right).$$

The objective of this section is to prove both that

$$\begin{aligned} \mathbf{P} \left( \widehat{F}_{\hat{f}}(\hat{f}(X_\bullet)) \geq 1 - \varepsilon \right) &\rightarrow \varepsilon, \quad \text{and} \\ \mathbf{R} \left( \widehat{\Gamma}_\varepsilon^\bullet \right) &\rightarrow \mathcal{R}(\Gamma_\varepsilon^\bullet), \quad n, N \rightarrow +\infty, \end{aligned} \quad (10)$$

and to derive rates for these convergences. Since  $\mathcal{D}_n$  is dedicated to the estimation of  $\eta^*$  and  $\mathcal{D}_N$  to the estimation of  $F_{\hat{f}}$ , we prove that

$$\mathbf{R} \left( \widehat{\Gamma}_\varepsilon^\bullet \right) - \mathbf{R} \left( \widetilde{\Gamma}_\varepsilon^\bullet \right) \rightarrow 0, \quad \text{and} \quad (11)$$

$$\mathbf{R} \left( \widetilde{\Gamma}_\varepsilon^\bullet \right) - \mathcal{R}(\Gamma_\varepsilon^\bullet) \rightarrow 0, \quad (12)$$

when both  $n$  and  $N$  go to infinity. The convergences (10) and (11) relies on the Dvoretzky–Kiefer–Wolfowitz inequality [Mas90] while the convergence (12) relies on the following inequality.

**Proposition 4.** *For all  $\varepsilon \in ]0, 1]$ , the following inequality holds under Assumptions (A1) and (A2)*

$$\begin{aligned} 0 \leq \mathbf{R} \left( \widetilde{\Gamma}_\varepsilon^\bullet \right) - \mathcal{R}(\Gamma_\varepsilon^\bullet) &\leq \frac{1}{\varepsilon} \{ \mathbf{E} [ |\eta^*(X_\bullet) - \alpha_\varepsilon| \mathbf{1}_{\{|\hat{\eta}(X_\bullet) - \eta^*(X_\bullet)| \geq |\eta^*(X_\bullet) - \alpha_\varepsilon|\}}] \\ &\quad + \mathbf{E} [ |1 - \eta^*(X_\bullet) - \alpha_\varepsilon| \mathbf{1}_{\{|\hat{\eta}(X_\bullet) - \eta^*(X_\bullet)| \geq |1 - \eta^*(X_\bullet) - \alpha_\varepsilon|\}}] \\ &\quad + \alpha_\varepsilon |F_{\hat{f}}(\alpha_\varepsilon) - F_{\hat{f}}^*(\alpha_\varepsilon)| \}, \end{aligned}$$

where  $\alpha_\varepsilon = (F_{\hat{f}}^*)^{-1}(1 - \varepsilon)$ .

The proof of Proposition 4 relies on Proposition 1. For  $\varepsilon \in ]0, 1]$ , Proposition 4 evaluates the loss of performance using the confidence set  $\widetilde{\Gamma}_\varepsilon^\bullet$  instead of the  $\varepsilon$ -confidence set  $\Gamma_\varepsilon^\bullet$ . We can distinguish two parts in the upper bound. One part is linked to the classification with reject option setting provided by [HW06] and then depends on the behavior of the function  $f$  around

$\alpha_\varepsilon$ . Note that the same quantity is obtained by [HW06]. The second part  $\alpha_\varepsilon |F_{\hat{f}}(\alpha_\varepsilon) - F_f^*(\alpha_\varepsilon)|$  is related to our proposed confidence set and is due to the approximation of  $F_f^*$  by  $F_{\hat{f}}$ .

Observe that when  $\varepsilon = 1$  one can recover a classical inequality in the classification setting. Indeed, in this case,  $\alpha_\varepsilon = 1/2$  and  $F_{\hat{f}}(1/2) = F_f^*(1/2)$ . Hence, we obtain

$$\mathbf{R}(\tilde{\Gamma}_1^\bullet) - \mathcal{R}(\Gamma_1^\bullet) \leq \mathbf{E} [2\eta^*(X_\bullet) - 1 | \mathbf{1}_{\{|\hat{\eta}(X_\bullet) - \eta^*(X_\bullet)| \geq |\eta^*(X_\bullet) - 1/2\}}] .$$

Finally, we state our main result which describes the asymptotic behavior of our plug-in  $\varepsilon$ -confidence sets:

**Theorem 1.** 1. If  $\hat{\eta}(X_\bullet) \rightarrow \eta^*(X_\bullet)$  in probability when  $n \rightarrow +\infty$ , then for any  $\varepsilon \in ]0, 1]$

$$\mathbf{P}(\hat{F}_{\hat{f}}(\hat{f}(X_\bullet)) \geq 1 - \varepsilon) \rightarrow \varepsilon,$$

and

$$\mathbf{R}(\hat{\Gamma}_\varepsilon^\bullet) - \mathcal{R}(\Gamma_\varepsilon^\bullet) \rightarrow 0,$$

when both  $n$  and  $N$  go to infinity.

2. For any  $\varepsilon \in ]0, 1]$ , assume that there exist  $C_1 < \infty$  and  $\gamma_\varepsilon > 0$  such that

$$\mathbf{P}(|f(X) - \alpha_\varepsilon| \leq t) \leq C_1 t^{\gamma_\varepsilon}, \quad \forall t > 0. \quad (13)$$

Assume also that there exist a sequence of positive numbers  $a_n \rightarrow +\infty$  and some positive constants  $C_2, C_3$  such that

$$\mathbf{P}(|\hat{\eta}(x) - \eta(x)| \geq t) \leq C_2 \exp(-C_3 a_n t^2), \quad \forall t > 0, \forall x \in \mathcal{X}. \quad (14)$$

Then we have

$$\mathbf{P}(\hat{F}_{\hat{f}}(\hat{f}(X_\bullet)) \geq 1 - \varepsilon) = \varepsilon + O(N^{-1/2}), \quad (15)$$

and

$$\mathbf{R}(\hat{\Gamma}_\varepsilon^\bullet) - \mathcal{R}(\Gamma_\varepsilon^\bullet) = O(a_n^{-\gamma_\varepsilon/2}) + O(N^{-1/2}). \quad (16)$$

The proof of this result is postponed to the Appendix. Theorem 1 states that if the estimator of  $\eta^*$  is consistent, then asymptotically the plug-in  $\varepsilon$ -confidence set is level- $\varepsilon$ -confidence set and performs as well as the  $\varepsilon$ -confidence set. Moreover, several observations can be made according to the second point of the theorem. First of all, we mention the rate of convergence (15) does not require any of the assumptions (13) and (14). It only needs the consistency of the estimator of  $\eta^*$ . Second, we mention that the assumption (13) has already been introduced in the classification with reject option setting in [HW06]. It is analogous to Tsybakov's margin condition in [Tsy04, AT07] introduced in the classification framework. We point out here the fact that if  $\eta^*(X_\bullet)$  has a density w.r.t. the Lebesgue measure, the assumption (13) is satisfied with  $\gamma_\varepsilon \geq 1$  for any  $\varepsilon \in ]0, 1]$ . Third, if  $\gamma_\varepsilon \gg 1$ , we can get faster rate of convergence. However, this rate cannot be better than  $O(N^{-1/2})$  which is the term due to the estimation of the cumulative distribution function  $F_{\hat{f}}$ . This term is however not limiting. Indeed, recalling that the sample size  $N$  refers to the dataset  $\mathcal{D}_N$  which can consist only of unlabeled observations, getting large  $N$  is not a big issue. Hence, we can consider the first term  $O(a_n^{-\gamma_\varepsilon/2})$  as the leading term in (16). This term relies on Proposition 4 and on the assumption (14) which is crucial to establish our rate of convergence. Note that various estimators satisfy this condition such as kernel estimators (see [AT07], for more details).

## 4 Numerical results

In this section, we evaluate the plug-in  $\varepsilon$ -confidence sets numerically. Moreover, we indicate the importance of Assumptions (A1) and (A2).

### 4.1 Under Assumptions (A1)-(A2)

In this section both of the cumulative distribution functions  $F_f$  and  $F_{\hat{f}}$  are continuous. We generate  $(X, Y)$  according to the following models.

- Model 1:

1. the feature  $X \stackrel{\mathcal{L}}{=} (U_1, \dots, U_{10})$ , where  $U_i$  are i.i.d from a uniform distribution on  $[0, 1]$ ;
2. conditional on  $X$ , the label  $Y$  is drawn according to a Bernoulli distribution with parameter  $\eta^*(X)$  defined by  $\text{logit}(\eta^*(X)) = X^1 - X^2 - X^3 + X^9$ , where  $X^j$  is the  $j^{\text{th}}$  component of  $X$ .

- Model 2:

1. the feature  $X \stackrel{\mathcal{L}}{=} (\mathcal{N}_1, \mathcal{N}_2, \mathcal{N}_3)$ , where  $\mathcal{N}_i$  are i.i.d from standard Gaussian distribution;
2. conditional on  $X$ , the label  $Y$  is drawn according to a Bernoulli distribution with parameter  $\eta^*(X)$  defined by  $\text{logit}(\eta^*(X)) = (X^1)^2 + \frac{X^2}{2} + \sin(X^1 + X^3) + 3X^3$ .

The first model leads to a classification problem which is quite difficult. Indeed, using a large dataset of features, we evaluate the distribution function of  $\eta^*(X)$ , and then obtain that  $\mathbb{P}(\eta^*(X) \in [0.4, 0.6]) \simeq 0.5$ . On the contrary, estimating  $\eta^*$  is easy since  $\text{logit}(\eta^*(X))$  is a linear function of  $X$ . Model 2 provides a more simple classification problem: the estimation of the distribution function of  $\eta^*(X)$  leads to  $\mathbb{P}(\eta^*(X) \in [0.4, 0.6]) \simeq 0.15$ . On the other side, the estimation of  $\eta^*$  is a little more tricky.

In order to illustrate our convergence result, we first provide estimation of the risk  $\mathcal{R}$  for the  $\varepsilon$ -confidence sets. More precisely, for each model and each  $\varepsilon \in \{\frac{k}{10}, k \in \{1, \dots, 10\}\}$ , we repeat  $B = 100$  times the following steps:

- i*) simulate two data sets  $\mathcal{D}_N$  and  $\mathcal{D}_K$  according to the considered model with  $N = 1000$  and  $K = 1000$ ;
- ii*) based on  $\mathcal{D}_N$ , we compute the empirical cumulative distribution of  $f^*(X)$  (this step requires only the features);
- iii*) finally, we compute, over  $\mathcal{D}_K$ , the empirical counterparts  $\mathcal{R}_K$  of the risk  $\mathcal{R}$  of the  $\varepsilon$ -confidence set using the empirical cumulative distribution of  $f^*(X)$  instead of  $F_f^*$ . We also compute the proportion of classified instances  $\mathcal{P}_K$ .

From these experiments, we compute the mean and standard deviation of  $\mathcal{R}_K$  and  $\mathcal{P}_K$ . The results are reported in Table 1 and illustrated in Figure 1. Next, for each model and each  $\varepsilon \in \{\frac{k}{10}, k \in \{1, \dots, 10\}\}$ , we estimate the risk  $\mathbf{R}$  for the plug-in  $\varepsilon$ -confidence set. We propose to use three popular classification methods for the estimation of  $\eta^*$ : *random forest*, *logistic regression* and *kernel rule* based on the Gaussian kernel and window parameter equal to 1. We perform the following simulation scheme. We repeat independently  $B$  times the following steps:

- i*) simulate three dataset  $\mathcal{D}_n, \mathcal{D}_N, \mathcal{D}_K$  according to the considered model;

| $\varepsilon$ | Model 1         |                 | Model 2         |                 |
|---------------|-----------------|-----------------|-----------------|-----------------|
|               | $\mathcal{R}_K$ | $\mathcal{P}_K$ | $\mathcal{R}_K$ | $\mathcal{P}_K$ |
| 1             | 0.39 (0.01)     | 1.00 (0.00)     | 0.22 (0.01)     | 1.00 (0.00)     |
| 0.9           | 0.38 (0.02)     | 0.90 (0.01)     | 0.19 (0.01)     | 0.90 (0.01)     |
| 0.8           | 0.37 (0.02)     | 0.80 (0.02)     | 0.16 (0.01)     | 0.80 (0.02)     |
| 0.7           | 0.35 (0.02)     | 0.70 (0.02)     | 0.14 (0.01)     | 0.70 (0.02)     |
| 0.6           | 0.34 (0.02)     | 0.60 (0.02)     | 0.12 (0.01)     | 0.60 (0.02)     |
| 0.5           | 0.33 (0.02)     | 0.50 (0.02)     | 0.09 (0.01)     | 0.50 (0.02)     |
| 0.4           | 0.31 (0.02)     | 0.40 (0.02)     | 0.07 (0.01)     | 0.40 (0.02)     |
| 0.3           | 0.29 (0.03)     | 0.30 (0.02)     | 0.05 (0.01)     | 0.30 (0.02)     |
| 0.2           | 0.27 (0.03)     | 0.20 (0.02)     | 0.03 (0.01)     | 0.20 (0.02)     |
| 0.1           | 0.24 (0.03)     | 0.10 (0.01)     | 0.02 (0.01)     | 0.10 (0.01)     |

Table 1: For each of the  $B = 100$  repetitions and each model, we derive the estimates  $\mathcal{R}_K$  of the risk and the estimated proportions of classified instances  $\mathcal{P}_K$  of the  $\varepsilon$ -confidence sets w.r.t.  $\varepsilon$ . We compute the means and standard deviations (between parentheses) over the  $B = 100$  repetitions. Left: the data are generated according to Model 1 – Right: the data are generated according to Model 2.

- ii*) based on  $\mathcal{D}_n$ , we compute an estimate, denoted by  $\hat{f}$ , of  $f^*$  with the random forest, the logistic regression or kernel rule procedure;
- iii*) based on  $\mathcal{D}_N$ , we compute the empirical cumulative distribution of  $\hat{f}(X)$  (we recall that this step requires a dataset which contains only the features);
- iv*) finally, over  $\mathcal{D}_K$ , we compute the empirical counterpart  $\mathbf{R}_K$  of the risk  $\mathcal{R}$  and the proportion  $\mathcal{P}_K$  of the data which are not rejected.

From these results, we compute the means and standard deviations of both empirical risks and proportions of classified instances for  $n \in \{100, 1000\}$ . We fix  $N = 100$  and  $K = 1000$ . The results are illustrated in Figure 1 and provided in Table 2 and 3.

From our numerical study, we make several observations. First, as expected, the risk of the  $\varepsilon$ -confidence sets is decreasing with  $\varepsilon$  as observed in Table 1. In both models, the reject option contributes to improve the overall misclassification risk. As an example, we see that in Model 2 the estimated value of the misclassification risk, that is when  $\varepsilon = 1$ , equals 0.22 whereas if  $\varepsilon = 0.1$  the estimated value of risk is 0.02 which is a significant improvement. Note that in Model 1, the classification problem is quite difficult and then the decrease of the risk seems to be slower and a bit less significant. On the other hand, we also observe in Table 1 that the proportions of classified data match with the theoretical values. Regarding Tables 2-3, the same comments can be made in both models and whatever the used classification procedure. Moreover, some features of Table 2 are worth commenting on. For fixed  $\varepsilon$  and for each scenario, the estimated risk of the all plug-in  $\varepsilon$ -confidence sets decreases with  $n$  which is the size of the sample used to estimate the regression function  $\eta^*$ . Furthermore, for  $n = 1000$  and viewing Table 1, we observe that the estimated risks of the plug-in  $\varepsilon$ -confidence sets are close to the oracle ones. This illustrates the convergence result provided in Theorem 1. However, we can see that the random forest procedure are outperformed by the other procedures (especially when  $\varepsilon$  is small). Indeed, the construction of plug-in  $\varepsilon$ -confidence sets relies on the estimator of  $\eta^*$ : better estimators lead to better confidence sets. Figure 1 summarizes many aspects of our previous discussion.

| Model 1       |             |              |             |             |              |             |
|---------------|-------------|--------------|-------------|-------------|--------------|-------------|
| $n = 100$     |             |              |             | $n = 1000$  |              |             |
| $\varepsilon$ | rforest     | logistic reg | kernel      | rforest     | logistic reg | kernel      |
| 1             | 0.45 (0.02) | 0.43 (0.03)  | 0.47 (0.03) | 0.42 (0.02) | 0.39 (0.02)  | 0.42 (0.03) |
| 0.9           | 0.45 (0.02) | 0.42 (0.03)  | 0.45 (0.03) | 0.41 (0.02) | 0.38 (0.02)  | 0.41 (0.03) |
| 0.8           | 0.44 (0.02) | 0.42 (0.03)  | 0.45 (0.03) | 0.40 (0.02) | 0.37 (0.02)  | 0.39 (0.03) |
| 0.7           | 0.44 (0.03) | 0.41 (0.03)  | 0.44 (0.03) | 0.39 (0.02) | 0.36 (0.02)  | 0.38 (0.03) |
| 0.6           | 0.43 (0.03) | 0.40 (0.03)  | 0.43 (0.03) | 0.38 (0.02) | 0.35 (0.02)  | 0.37 (0.03) |
| 0.5           | 0.42 (0.03) | 0.39 (0.03)  | 0.42 (0.03) | 0.37 (0.02) | 0.34 (0.02)  | 0.36 (0.03) |
| 0.4           | 0.41 (0.03) | 0.38 (0.04)  | 0.40 (0.04) | 0.36 (0.03) | 0.32 (0.03)  | 0.34 (0.02) |
| 0.3           | 0.41 (0.04) | 0.37 (0.04)  | 0.39 (0.04) | 0.35 (0.03) | 0.30 (0.03)  | 0.33 (0.04) |
| 0.2           | 0.40 (0.04) | 0.35 (0.05)  | 0.37 (0.05) | 0.34 (0.03) | 0.28 (0.04)  | 0.30 (0.04) |
| 0.1           | 0.38 (0.06) | 0.33 (0.06)  | 0.35 (0.06) | 0.32 (0.05) | 0.25 (0.05)  | 0.27 (0.05) |

| Model 2       |             |              |             |             |              |             |
|---------------|-------------|--------------|-------------|-------------|--------------|-------------|
| $n = 100$     |             |              |             | $n = 1000$  |              |             |
| $\varepsilon$ | rforest     | logistic reg | kernel      | rforest     | logistic reg | kernel      |
| 1             | 0.26 (0.02) | 0.24 (0.01)  | 0.27 (0.05) | 0.24 (0.01) | 0.22 (0.01)  | 0.23 (0.02) |
| 0.9           | 0.24 (0.02) | 0.21 (0.02)  | 0.25 (0.05) | 0.22 (0.01) | 0.20 (0.01)  | 0.20 (0.01) |
| 0.8           | 0.21 (0.02) | 0.18 (0.02)  | 0.22 (0.04) | 0.19 (0.02) | 0.17 (0.02)  | 0.18 (0.02) |
| 0.7           | 0.19 (0.02) | 0.16 (0.02)  | 0.19 (0.04) | 0.16 (0.02) | 0.14 (0.02)  | 0.15 (0.02) |
| 0.6           | 0.18 (0.02) | 0.13 (0.02)  | 0.16 (0.04) | 0.15 (0.02) | 0.12 (0.02)  | 0.13 (0.02) |
| 0.5           | 0.16 (0.03) | 0.11 (0.02)  | 0.14 (0.04) | 0.12 (0.02) | 0.10 (0.02)  | 0.11 (0.02) |
| 0.4           | 0.15 (0.03) | 0.09 (0.02)  | 0.11 (0.03) | 0.11 (0.02) | 0.08 (0.02)  | 0.08 (0.02) |
| 0.3           | 0.13 (0.03) | 0.07 (0.02)  | 0.08 (0.03) | 0.09 (0.02) | 0.06 (0.02)  | 0.06 (0.02) |
| 0.2           | 0.12 (0.03) | 0.05 (0.02)  | 0.06 (0.02) | 0.08 (0.02) | 0.04 (0.01)  | 0.04 (0.02) |
| 0.1           | 0.10 (0.04) | 0.03 (0.02)  | 0.04 (0.02) | 0.06 (0.03) | 0.02 (0.01)  | 0.02 (0.02) |

Table 2: For each of the  $B = 100$  repetitions and each model, we derive the estimated risks  $\mathbf{R}_K$  of three different plug-in  $\varepsilon$ -confidence sets w.r.t.  $\varepsilon$  and to the sample size  $n$ . We compute the means and standard deviations (between parentheses) over the  $B = 100$  repetitions. For each  $\varepsilon$  and each  $n$ , the plug-in  $\varepsilon$ -confidence sets are based on, from left to right, **rforest**, **logistic reg** and **kernel**, which are respectively the random forest, the logistic regression and the kernel rule procedures. Top: the data are generated according to Model 1 – Bottom: the data are generated according to Model 2.

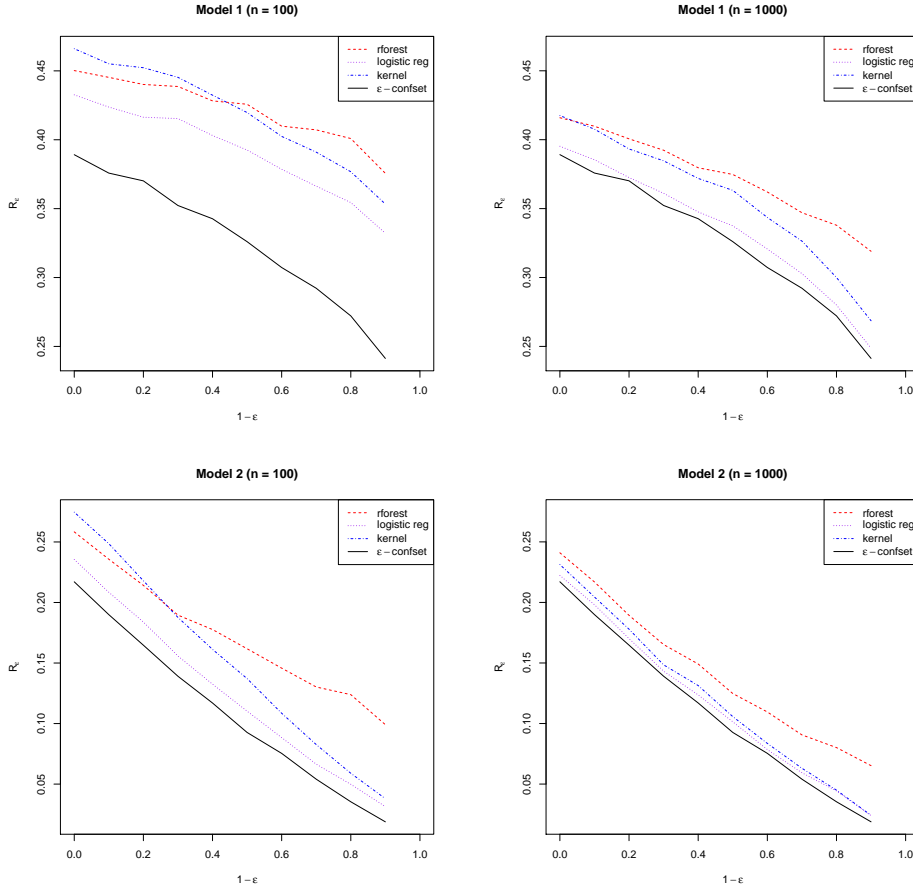


Figure 1: Visual description of the results reported in Table 1 and 2. For each model and each  $n$ , we plot, as a function of  $1 - \epsilon$ , the mean  $R_\epsilon$  over the  $B = 100$  repetitions of the estimated risks  $\mathcal{R}_K$  of the  $\epsilon$ -confidence sets (solid line) and  $\mathbf{R}_K$  of the plug-in  $\epsilon$ -confidence sets based on random forest (dashed line), logistic regression (dotted line) and kernel rule (dotted dashed line). Top: the data are generated according to Model 1 (left:  $n = 100$ ; right:  $n = 1000$ ) – Bottom: the data are generated according to Model 2 (left:  $n = 100$ ; right:  $n = 1000$ ).

| Model 1       |             |              |             |             |              |             |             |
|---------------|-------------|--------------|-------------|-------------|--------------|-------------|-------------|
| $n = 100$     |             |              |             | $n = 1000$  |              |             |             |
| $\varepsilon$ | rforest     | logistic reg | kernel      | rforest     | logistic reg | kernel      |             |
| 1             | 1.00 (0.00) | 1.00 (0.00)  | 1.00 (0.00) | 1.00 (0.00) | 1.00 (0.00)  | 1.00 (0.00) | 1.00 (0.00) |
| 0.9           | 0.90 (0.03) | 0.90 (0.03)  | 0.90 (0.04) | 0.91 (0.03) | 0.90 (0.03)  | 0.90 (0.03) | 0.90 (0.03) |
| 0.8           | 0.80 (0.04) | 0.79 (0.04)  | 0.80 (0.04) | 0.81 (0.04) | 0.80 (0.04)  | 0.80 (0.04) | 0.80 (0.04) |
| 0.7           | 0.70 (0.05) | 0.69 (0.04)  | 0.70 (0.04) | 0.69 (0.05) | 0.69 (0.05)  | 0.69 (0.04) | 0.69 (0.04) |
| 0.6           | 0.61 (0.05) | 0.60 (0.05)  | 0.61 (0.05) | 0.60 (0.05) | 0.60 (0.05)  | 0.60 (0.06) | 0.60 (0.06) |
| 0.5           | 0.51 (0.05) | 0.49 (0.06)  | 0.50 (0.06) | 0.51 (0.05) | 0.50 (0.05)  | 0.51 (0.06) | 0.51 (0.06) |
| 0.4           | 0.40 (0.05) | 0.40 (0.05)  | 0.40 (0.05) | 0.40 (0.05) | 0.40 (0.05)  | 0.39 (0.05) | 0.39 (0.05) |
| 0.3           | 0.30 (0.05) | 0.30 (0.05)  | 0.30 (0.04) | 0.30 (0.05) | 0.30 (0.05)  | 0.29 (0.05) | 0.29 (0.05) |
| 0.2           | 0.20 (0.04) | 0.21 (0.05)  | 0.21 (0.05) | 0.21 (0.04) | 0.21 (0.04)  | 0.20 (0.04) | 0.20 (0.04) |
| 0.1           | 0.10 (0.03) | 0.10 (0.03)  | 0.11 (0.03) | 0.11 (0.03) | 0.10 (0.03)  | 0.11 (0.03) | 0.11 (0.03) |

| Model 2       |             |              |             |             |              |             |             |
|---------------|-------------|--------------|-------------|-------------|--------------|-------------|-------------|
| $n = 100$     |             |              |             | $n = 1000$  |              |             |             |
| $\varepsilon$ | rforest     | logistic reg | kernel      | rforest     | logistic reg | kernel      |             |
| 1             | 1.00 (0.00) | 1.00 (0.00)  | 1.00 (0.00) | 1.00 (0.00) | 1.00 (0.00)  | 1.00 (0.00) | 1.00 (0.00) |
| 0.9           | 0.90 (0.04) | 0.90 (0.03)  | 0.90 (0.04) | 0.90 (0.03) | 0.90 (0.03)  | 0.90 (0.03) | 0.90 (0.03) |
| 0.8           | 0.81 (0.04) | 0.81 (0.03)  | 0.81 (0.04) | 0.80 (0.04) | 0.80 (0.04)  | 0.80 (0.05) | 0.80 (0.05) |
| 0.7           | 0.70 (0.05) | 0.70 (0.04)  | 0.70 (0.04) | 0.69 (0.05) | 0.69 (0.05)  | 0.69 (0.05) | 0.69 (0.05) |
| 0.6           | 0.61 (0.05) | 0.61 (0.05)  | 0.60 (0.05) | 0.61 (0.05) | 0.60 (0.05)  | 0.60 (0.05) | 0.60 (0.05) |
| 0.5           | 0.51 (0.05) | 0.50 (0.04)  | 0.50 (0.04) | 0.50 (0.05) | 0.51 (0.05)  | 0.51 (0.05) | 0.51 (0.05) |
| 0.4           | 0.39 (0.05) | 0.40 (0.05)  | 0.39 (0.05) | 0.40 (0.05) | 0.40 (0.05)  | 0.40 (0.05) | 0.40 (0.05) |
| 0.3           | 0.29 (0.05) | 0.30 (0.04)  | 0.30 (0.05) | 0.30 (0.04) | 0.29 (0.05)  | 0.29 (0.04) | 0.29 (0.04) |
| 0.2           | 0.21 (0.04) | 0.20 (0.04)  | 0.20 (0.04) | 0.20 (0.04) | 0.21 (0.04)  | 0.21 (0.04) | 0.21 (0.04) |
| 0.1           | 0.11 (0.03) | 0.11 (0.03)  | 0.11 (0.03) | 0.12 (0.03) | 0.11 (0.03)  | 0.11 (0.03) | 0.11 (0.03) |

Table 3: For each of the  $B = 100$  repetitions and each model, we derive the estimated proportion of classified instances  $\mathcal{P}_K$  of three different plug-in  $\varepsilon$ -confidence sets w.r.t.  $\varepsilon$  and to the sample size  $n$ . We compute the means and standard deviations (between parentheses) over the  $B = 100$  repetitions. For each  $\varepsilon$  and each  $n$ , the plug-in  $\varepsilon$ -confidence sets are based on, from left to right, **rforest**, **logistic reg** and **kernel**, which are respectively the random forest, the logistic regression and the kernel rule procedures. Top: the data are generated according to Model 1 – Bottom: the data are generated according to Model 2.

| $\varepsilon$ | (A2) fails (CART) |                | (A1) fails (kernel) |                |
|---------------|-------------------|----------------|---------------------|----------------|
|               | $\mathcal{P}_K$   | $\mathbf{R}_K$ | $\mathcal{P}_K$     | $\mathbf{R}_K$ |
| 1             | 1.00 (0.00)       | 0.27 (0.03)    | 1.00 (0.00)         | 0.32 (0.03)    |
| 0.9           | 0.98 (0.04)       | 0.27 (0.04)    | 0.90 (0.03)         | 0.31 (0.03)    |
| 0.8           | 0.90 (0.07)       | 0.24 (0.03)    | 0.80 (0.04)         | 0.29 (0.03)    |
| 0.7           | 0.84 (0.10)       | 0.22 (0.03)    | 0.70 (0.05)         | 0.27 (0.04)    |
| 0.6           | 0.79 (0.13)       | 0.21 (0.04)    | 0.61 (0.05)         | 0.26 (0.04)    |
| 0.5           | 0.75 (0.16)       | 0.21 (0.04)    | 0.50 (0.05)         | 0.24 (0.04)    |
| 0.4           | 0.60 (0.14)       | 0.18 (0.05)    | 0.40 (0.05)         | 0.23 (0.03)    |
| 0.3           | 0.48 (0.13)       | 0.18 (0.06)    | 0.30 (0.04)         | 0.22 (0.03)    |
| 0.2           | 0.39 (0.13)       | 0.18 (0.06)    | 0.20 (0.04)         | 0.20 (0.03)    |
| 0.1           | 0.31 (0.12)       | 0.16 (0.06)    | 0.11 (0.03)         | 0.20 (0.04)    |

Table 4: For each of the  $B = 100$  repetitions, we derive the estimated proportions of classified instances  $\mathcal{P}_K$  and the estimated risks  $\mathbf{R}_K$  of the two plug-in  $\varepsilon$ -confidence sets w.r.t.  $\varepsilon$ . We compute the means and standard deviations (between parentheses) over the  $B = 100$  repetitions. Left: the data are generated according to Model 2, then Assumption (A1) holds; the procedure used to build the plug-in  $\varepsilon$ -confidence set is based on CART method, then **Assumption (A2) fails** – Right: the data are generated according to Model 3, then **Assumption (A1) fails**; the procedure used to build the plug-in  $\varepsilon$ -confidence set is based on kernels, then Assumption (A2) holds.

## 4.2 Importance of Assumptions (A1) and (A2)

In this section, we shed some light on the importance of Assumptions (A1) and (A2). More precisely, we study the behavior of plug-in  $\varepsilon$ -confidence sets when one of these two assumptions is not satisfied.

We first consider a case where the cumulative distribution  $F_f$  is continuous but not  $F_{\hat{f}}$ . We consider the simulation scheme described in Section 4.1 with Model 2 and parameters  $n = 100$ ,  $N = 100$  and  $K = 1000$ . But this time, the plug-in  $\varepsilon$ -confidence set relies on the CART procedure which involves that the Assumption (A2) does not hold. The obtained results are reported in Table 4–Left. Two observations can be made. First, judging by the estimated proportions of classified instances and by the associated standard deviations, we are not able to control these proportions. Therefore, one of the important feature of our procedure fails. Second, although the risk of misclassification is decreasing with  $\varepsilon$ , this decrease is quite slow and cannot be as important as observed with plug-in confidence sets studied in Section 4.1 (see Table 2). Indeed, for CART method and more generally if Assumption (A2) does not hold, the proportion of rejected data is usually not large enough.

Next, we study the reverse case where the cumulative distribution function  $F_{\hat{f}}$  is continuous but not  $F_f$ . We consider the following model.

- Model 3:

1. the feature  $X \stackrel{\mathcal{L}}{=} U$ , where  $U$  follows a uniform distribution on  $[0, 1]$ ;
2. conditional on  $X$ , the label  $Y$  is drawn according a Bernoulli distribution with parameter

$$\eta^*(X) = \frac{1}{5} \mathbf{1}_{\{X \leq 1/4\}} + \frac{2}{5} \mathbf{1}_{\{1/4 < X \leq 1/2\}} + \frac{3}{5} \mathbf{1}_{\{1/2 < X \leq 3/4\}} + \frac{4}{5} \mathbf{1}_{\{3/4 < X\}}.$$

Then, for this model,  $F_f$  is not continuous. Moreover, we have that, for  $\varepsilon \in [0.5, 1]$ ,  $\mathcal{P}(\Gamma_\varepsilon^\bullet) = 1$  and  $\mathcal{R}(\Gamma_\varepsilon^\bullet) = 3/10$ , and for  $\varepsilon \in ]0, 0.5[$ ,  $\mathcal{P}(\Gamma_\varepsilon^\bullet) = 1/2$  and  $\mathcal{R}(\Gamma_\varepsilon^\bullet) = 1/5$ . For this model, we use



the simulation scheme described in Section 4.1 with the plug-in  $\varepsilon$  confidence sets which relies on the kernel rule and the samples sizes  $n = 1000$ ,  $N = 100$  and  $K = 1000$ . The results are provided in Table 4–Right. As a remark, we first note that since Assumption (A2) is satisfied the proportions of classified instances match with the theoretical values. Second, the estimated risk of misclassification of the plug-in  $\varepsilon$ -confidence set decreases with  $\varepsilon$ . However, from our point of view, it is irrelevant to compare the performances of the  $\varepsilon$ -confidence sets and those of the plug-in  $\varepsilon$ -confidence sets. Indeed, except for  $\varepsilon = 1$ , the proportions of classified data differ. As an example, if  $\varepsilon = 0.7$ , the estimated risk of the plug-in  $\varepsilon$ -confidence set is equal to 0.27 which seems better than the risk of the  $\varepsilon$ -confidence set. But, for  $\varepsilon = 0.7$  the proportion of the classified instances is larger for the  $\varepsilon$ -confidence set and equals 1.

## 5 Conclusion

In the classification with reject option framework, we introduce a new procedure that allows us to control exactly the rejection probability. The construction of the  $\varepsilon$ -confidence sets and their plug-in approximations relies on the cumulative distribution function of the score functions  $f^*$  and  $\hat{f}$ . Theoretical guarantees, especially rates of convergence, involve the continuity of these cumulative distribution functions. Numerical experiments emphasize the importance of the continuity assumption. As viewed in Section 3, the plug-in  $\varepsilon$ -confidence set is defined as a two steps algorithm whose second step consists in the estimation of the cumulative distribution function  $F_{\hat{f}}$ . Interestingly, this step does not require a set of labeled data that is suitable for semi-supervised learning. In a future work, we intent to generalize our procedure to the multiclass case and study procedures based on empirical risk minimization.

## 6 Appendix

This section gathers the proofs of our results.

### 6.1 Proof of Proposition 1

We first define the following events

$$\begin{aligned} \mathcal{A}_y &= \{f^*(X_\bullet) \geq \alpha_\varepsilon, \text{card}(\Gamma_s(X_\bullet)) \neq 1, s^*(X_\bullet) \neq y\}, \quad y = 0, 1, \\ \mathcal{B}_y &= \{f^*(X_\bullet) < \alpha_\varepsilon, \text{card}(\Gamma_s(X_\bullet)) = 1, s(X_\bullet) \neq y\}, \quad y = 0, 1, \\ \mathcal{C} &= \{f^*(X_\bullet) \geq \alpha_\varepsilon, \text{card}(\Gamma_s(X_\bullet)) = 1, s^*(X_\bullet) \neq s(X_\bullet)\}, \\ \mathcal{D} &= \mathcal{A}_0 \cup \mathcal{A}_1 \cup \mathcal{B}_0 \cup \mathcal{B}_1, \end{aligned}$$

and the random variable

$$U = \mathbf{1}_{\{s(X_\bullet) \neq Y_\bullet, \text{card}(\Gamma_s(X_\bullet))=1\}} - \mathbf{1}_{\{s^*(X_\bullet) \neq Y_\bullet, \text{card}(\Gamma_s^*(X_\bullet))=1\}}.$$

Since  $\mathcal{P}(\Gamma_s) = \mathcal{P}(\Gamma_s^*) = \varepsilon$ , the proof of the proposition relies on the decomposition of the conditional expectation of  $U$  given  $X_\bullet$  over the sets  $\mathcal{C}$  and  $\mathcal{D}$ . We have

$$\begin{aligned} \mathbb{E}[U \mathbf{1}_{\mathcal{C}} | X_\bullet] &= \mathbf{1}_{\mathcal{C}} \{\eta^*(X_\bullet) \mathbf{1}_{\{s^*(X_\bullet)=1\}} - \eta^*(X_\bullet) \mathbf{1}_{\{s^*(X_\bullet)=0\}} + \\ &\quad (\eta^*(X_\bullet) - 1) \mathbf{1}_{\{s^*(X_\bullet)=1\}} + (1 - \eta^*(X_\bullet)) \mathbf{1}_{\{s^*(X_\bullet)=0\}}\}. \end{aligned}$$

Since,  $s^*(X_\bullet) = 1$  and  $s^*(X_\bullet) = 0$  imply respectively that  $\eta^*(X_\bullet) \geq 1/2$  and  $\eta^*(X_\bullet) \leq 1/2$ , we obtain from the above decomposition

$$\mathbb{E}[U\mathbf{1}_C] = \mathbb{E}[|2\eta^*(X_\bullet) - 1|\mathbf{1}_C]. \quad (17)$$

Next,

$$\mathbb{E}[U\mathbf{1}_D|X_\bullet] = \eta^*(X_\bullet)\mathbf{1}_{B_1} + (1 - \eta^*(X_\bullet))\mathbf{1}_{B_0} - \eta^*(X_\bullet)\mathbf{1}_{A_1} - (1 - \eta^*(X_\bullet))\mathbf{1}_{A_0} \quad (18)$$

Since,  $\mathcal{P}(\Gamma_s) = \mathcal{P}(\Gamma_\varepsilon^\bullet) = \mathbb{P}(f^*(X_\bullet) \geq \alpha_\varepsilon)$ , we deduce

$$\mathbb{P}(\text{card}(\Gamma_s(X_\bullet)) = 1, f^*(X_\bullet) < \alpha_\varepsilon) = \mathbb{P}(\text{card}(\Gamma_s(X_\bullet)) \neq 1, f^*(X_\bullet) \geq \alpha_\varepsilon),$$

which implies

$$(1 - \alpha_\varepsilon)\mathbb{E}[\mathbf{1}_{B_0 \cup B_1}] - (1 - \alpha_\varepsilon)\mathbb{E}[\mathbf{1}_{A_0 \cup A_1}] = 0.$$

Therefore, adding this null term to (18), we obtain

$$\begin{aligned} \mathbb{E}[U\mathbf{1}_D] &= \mathbb{E}[(\alpha_\varepsilon - (1 - \eta^*(X_\bullet))\mathbf{1}_{B_1} + (\alpha_\varepsilon - \eta^*(X_\bullet))\mathbf{1}_{B_0}) \\ &\quad + \mathbb{E}[(\eta^*(X_\bullet) - \alpha_\varepsilon)\mathbf{1}_{A_0} + ((1 - \eta^*(X_\bullet)) - \alpha_\varepsilon)\mathbf{1}_{A_1}]]. \end{aligned} \quad (19)$$

Note that,

$$\begin{aligned} f^*(X_\bullet) < \alpha_\varepsilon &\quad \Rightarrow \quad (\alpha_\varepsilon - (1 - \eta^*(X_\bullet)) \geq 0 \text{ and } (\alpha_\varepsilon - \eta^*(X_\bullet)) \geq 0 \\ f^*(X_\bullet) \geq \alpha_\varepsilon \text{ and } s^*(X_\bullet) \neq 1 &\quad \Rightarrow \quad (1 - \eta^*(X_\bullet) - \alpha_\varepsilon) \geq 0 \\ f^*(X_\bullet) \geq \alpha_\varepsilon \text{ and } s^*(X_\bullet) \neq 0 &\quad \Rightarrow \quad (\eta^*(X_\bullet) - \alpha_\varepsilon) \geq 0. \end{aligned}$$

Hence, from (19), we can write

$$\mathbb{E}[U\mathbf{1}_D] = \mathbb{E}[|\eta^*(X_\bullet) - \alpha_\varepsilon|\mathbf{1}_{A_0 \cup B_0}] + \mathbb{E}[|1 - \eta^*(X_\bullet) - \alpha_\varepsilon|\mathbf{1}_{A_1 \cup B_1}].$$

Combining this result with (17) shows that  $\mathbb{E}[U\mathbf{1}_{C \cup D}] \geq 0$ , and provides in the same time the desired result.

## 6.2 Proof of Proposition 2

We first prove the following inequality for  $\alpha, \tilde{\alpha} \in [0, 1/2[$ ,  $\alpha \leq \tilde{\alpha}$

$$\mathbb{P}(s^*(X_\bullet) \neq Y_\bullet | f^*(X_\bullet) \geq \tilde{\alpha}) \leq \mathbb{P}(s^*(X_\bullet) \neq Y_\bullet | f^*(X_\bullet) \geq \alpha), \quad (20)$$

Since for  $\varepsilon, \varepsilon' \in ]0, 1]$  one has  $\varepsilon \leq \varepsilon' \Leftrightarrow \alpha_\varepsilon \geq \alpha_{\varepsilon'}$ , a direct application of (20) yields the proposition. In order to prove (20), recall that

$$\frac{x}{y} - \frac{z}{t} = \frac{1}{2yt} ((x - z)(t + y) + (x + z)(t - y)), \quad \forall x, z \in \mathbb{R}, \forall y, t \in \mathbb{R} \setminus \{0\}. \quad (21)$$

Thus, if we define

$$\begin{aligned} C_1 &= \mathbb{P}(s^*(X_\bullet) \neq Y_\bullet, f^*(X_\bullet) \geq \alpha) - \mathbb{P}(s^*(X_\bullet) \neq Y_\bullet, f^*(X_\bullet) \geq \tilde{\alpha}) \\ C_2 &= \mathbb{P}(f^*(X_\bullet) \geq \alpha) + \mathbb{P}(f^*(X_\bullet) \geq \tilde{\alpha}) \\ C_3 &= \mathbb{P}(s^*(X_\bullet) \neq Y_\bullet, f^*(X_\bullet) \geq \alpha) + \mathbb{P}(s^*(X_\bullet) \neq Y_\bullet, f^*(X_\bullet) \geq \tilde{\alpha}) \\ C_4 &= \mathbb{P}(f^*(X_\bullet) \geq \tilde{\alpha}) - \mathbb{P}(f^*(X_\bullet) \geq \alpha), \end{aligned}$$

from (21), we have

$$\mathbb{P}(s^*(X_\bullet) \neq Y_\bullet | f^*(X_\bullet) \geq \tilde{\alpha}) \leq \mathbb{P}(s^*(X_\bullet) \neq Y_\bullet | f^*(X_\bullet) \geq \alpha) \Leftrightarrow C_1 C_2 + C_3 C_4 \geq 0.$$

Since  $\mathbb{P}(s^*(X_\bullet) \neq Y_\bullet | X_\bullet) = 1 - f^*(X_\bullet)$ , we deduce that

$$\begin{aligned} C_1 C_2 + C_3 C_4 &= \mathbb{E} \left[ (1 - f^*(X_\bullet)) \mathbf{1}_{\{\alpha \leq f^*(X_\bullet) \leq \tilde{\alpha}\}} \right] \mathbb{E} \left[ \mathbf{1}_{\{f^*(X_\bullet) \geq \alpha\}} + \mathbf{1}_{\{f^*(X_\bullet) \geq \tilde{\alpha}\}} \right] \\ &\quad - \mathbb{E} \left[ (1 - f^*(X_\bullet)) \left( \mathbf{1}_{\{f^*(X_\bullet) \geq \alpha\}} + \mathbf{1}_{\{f^*(X_\bullet) \geq \tilde{\alpha}\}} \right) \right] \mathbb{E} \left[ \mathbf{1}_{\{\alpha \leq f^*(X_\bullet) \leq \tilde{\alpha}\}} \right]. \end{aligned}$$

Note that  $\mathbb{E} \left[ \mathbf{1}_{\{f^*(X_\bullet) \geq \alpha\}} \right] = \mathbb{E} \left[ \mathbf{1}_{\{\alpha \leq f^*(X_\bullet) \leq \tilde{\alpha}\}} \right] + \mathbb{E} \left[ \mathbf{1}_{\{f^*(X_\bullet) \geq \tilde{\alpha}\}} \right]$ .

Hence, from the above decomposition, we obtain

$$\begin{aligned} C_1 C_2 + C_3 C_4 &= 2 \mathbb{E} \left[ \mathbf{1}_{\{\alpha \leq f^*(X_\bullet) \leq \tilde{\alpha}\}} \right] \mathbb{E} \left[ f^*(X_\bullet) \mathbf{1}_{\{f^*(X_\bullet) \geq \tilde{\alpha}\}} \right] \\ &\quad - 2 \mathbb{E} \left[ \mathbf{1}_{\{f^*(X_\bullet) \geq \tilde{\alpha}\}} \right] \mathbb{E} \left[ f^*(X_\bullet) \mathbf{1}_{\{\alpha \leq f^*(X_\bullet) \leq \tilde{\alpha}\}} \right]. \end{aligned}$$

Since,

$$\begin{aligned} \mathbb{E} \left[ \mathbf{1}_{\{\alpha \leq f^*(X_\bullet) \leq \tilde{\alpha}\}} \right] \mathbb{E} \left[ f^*(X_\bullet) \mathbf{1}_{\{f^*(X_\bullet) \geq \tilde{\alpha}\}} \right] &\geq \tilde{\alpha} \mathbb{P}(\alpha \leq f^*(X_\bullet) \leq \tilde{\alpha}) \mathbb{P}(f^*(X_\bullet) \geq \tilde{\alpha}) \text{ and} \\ \mathbb{E} \left[ \mathbf{1}_{\{f^*(X_\bullet) \geq \tilde{\alpha}\}} \right] \mathbb{E} \left[ f^*(X_\bullet) \mathbf{1}_{\{\alpha \leq f^*(X_\bullet) \leq \tilde{\alpha}\}} \right] &\leq \tilde{\alpha} \mathbb{P}(\alpha \leq f^*(X_\bullet) \leq \tilde{\alpha}) \mathbb{P}(f^*(X_\bullet) \geq \tilde{\alpha}), \end{aligned}$$

we deduce Inequality (20).

### 6.3 Proof of Proposition 3

This section is devoted to the proof of the result related to the Gaussian mixture model. Before starting, let us state a few properties that will be often used.

Let us write for short  $f_1$  and  $f_0$  instead of  $\eta^*$  and  $1 - \eta^*$  respectively, so that  $f = \max\{f_0, f_1\}$ . Hence, we can write:

$$\eta^*(x) = f_1(x) = \mathbb{P}(Y = 1 | X = x) = \frac{\mathbb{P}(X = x | Y = 1)}{\mathbb{P}(X = x | Y = 1) + \mathbb{P}(X = x | Y = 0)} := \frac{p_1(x)}{p_1(x) + p_0(x)},$$

for any  $x \in \mathcal{X}$ . Then, for  $y = 0, 1$ , using the fact that given  $Y_\bullet = y$ , the random variable  $X_\bullet^\top \Sigma^{-1}(\mu_0 - \mu_1) \sim \mathcal{N}(\mu_y^\top \Sigma^{-1}(\mu_0 - \mu_1), \|\mu_1 - \mu_0\|_{\Sigma^{-1}}^2)$  we get, on the event  $\{Y_\bullet = y\}$

$$\begin{aligned} s^*(X_\bullet) \neq Y_\bullet &\Leftrightarrow f^*(X_\bullet) = f_{1-y}(X_\bullet) \Leftrightarrow f_y(X_\bullet) \leq \frac{1}{2} \Leftrightarrow \log \left( \frac{p_{1-y}(X_\bullet)}{p_y(X_\bullet)} \right) \geq 0 \\ &\Leftrightarrow X_\bullet^\top \Sigma^{-1}(\mu_{1-y} - \mu_y) - \frac{1}{2} \mu_{1-y}^\top \Sigma^{-1} \mu_{1-y} + \frac{1}{2} \mu_y^\top \Sigma^{-1} \mu_y \geq 0 \\ &\Leftrightarrow (X_\bullet - \mu_y)^\top \Sigma^{-1}(\mu_{1-y} - \mu_y) - \frac{1}{2} \|\mu_1 - \mu_0\|_{\Sigma^{-1}}^2 \geq 0 \\ &\Leftrightarrow \frac{(X_\bullet - \mu_y)^\top \Sigma^{-1}(\mu_{1-y} - \mu_y)}{\|\mu_1 - \mu_0\|_{\Sigma^{-1}}} - \frac{1}{2} \|\mu_1 - \mu_0\|_{\Sigma^{-1}} \geq 0. \end{aligned} \tag{22}$$

where  $\|\cdot\|_{\Sigma^{-1}}$  denotes the norm under  $\Sigma^{-1}$ :  $\|\mu\|_{\Sigma^{-1}}^2 = \mu^\top \Sigma^{-1} \mu$ , for any  $\mu \in \mathcal{X}$ .

### 6.3.1 Intermediate results

The proof of Proposition 3 relies on two intermediate results. Then we state them first and give their proofs. They bring into play the cumulative distribution  $F_f^*$ .

**Proposition 5.** *Let  $y \in \{0, 1\}$ . Conditional on the event  $Y_\bullet = y$  we have*

$$F_f^*(f_{1-y}(X_\bullet)) = \Phi \left( \frac{(X_\bullet - \mu_y)^\top \Sigma^{-1}(\mu_{1-y} - \mu_y)}{\|\mu_1 - \mu_0\|_{\Sigma^{-1}}} \right) + \Phi \left( \frac{(X_\bullet - \mu_y)^\top \Sigma^{-1}(\mu_{1-y} - \mu_y)}{\|\mu_1 - \mu_0\|_{\Sigma^{-1}}} - \|\mu_1 - \mu_0\|_{\Sigma^{-1}} \right) - 1,$$

where  $\Phi$  is the standard normal cumulative distribution function.

*Proof.* To prove this result, we need to investigate the function  $F_f^*(\cdot) = \mathbb{P}(f^*(X) \leq \cdot)$ . Let  $\alpha \in [1/2, 1]$ . We have

$$\begin{aligned} \mathbb{P}(f^*(X) \leq \alpha) &= \mathbb{P}(f^*(X) \leq \alpha, f_1(X) \geq f_0(X)) + \mathbb{P}(f^*(X) \leq \alpha, f_1(X) \leq f_0(X)) = \\ &= \frac{1}{2} \mathbb{P}(f_1(X) \leq \alpha, f_1(X) \geq f_0(X) | Y = 1) + \frac{1}{2} \mathbb{P}(f_1(X) \leq \alpha, f_1(X) \geq f_0(X) | Y = 0) + \\ &= \frac{1}{2} \mathbb{P}(f_0(X) \leq \alpha, f_1(X) \leq f_0(X) | Y = 1) + \frac{1}{2} \mathbb{P}(f_0(X) \leq \alpha, f_1(X) \leq f_0(X) | Y = 0), \end{aligned} \quad (23)$$

where we used in the last equality the fact that  $Y$  is a Bernoulli random variable with parameter  $1/2$ . As already seen, we have for  $y \in \{0, 1\}$ ,

$$f_y(x) = \mathbb{P}(Y = y | X = x) = \frac{p_y(x)}{p_1(x) + p_0(x)}.$$

Hence, denoting by  $u$  the function from  $[1/2, 1)$  into  $[1, +\infty)$  defined by  $u(\alpha) = \frac{\alpha}{1-\alpha}$  and fixing this notation in the above relation (23), we get

$$\begin{aligned} \mathbb{P}(f^*(X) \leq \alpha) &= \frac{1}{2} \left[ \mathbb{P} \left( \frac{p_1(X)}{p_0(X)} \in [1, u(\alpha)] \mid Y = 1 \right) + \mathbb{P} \left( \frac{p_1(X)}{p_0(X)} \in [1, u(\alpha)] \mid Y = 0 \right) \right. \\ &\quad \left. + \mathbb{P} \left( \frac{p_0(X)}{p_1(X)} \in [1, u(\alpha)] \mid Y = 1 \right) + \mathbb{P} \left( \frac{p_0(X)}{p_1(X)} \in [1, u(\alpha)] \mid Y = 0 \right) \right] \\ &:= \frac{1}{2} (A_1 + A_2 + A_3 + A_4). \end{aligned} \quad (24)$$

All of the terms  $A_1, A_2, A_3, A_4$  will be treated in the same way. Then let us consider  $A_1$  for instance: using very close reasoning as in (22) with  $y = 1$ , we have

$$\begin{aligned} A_1 &= \mathbb{P} \left( 0 \leq \log \left( \frac{p_1(X)}{p_0(X)} \right) \leq \log(u(\alpha)) \mid Y = 1 \right) \\ &= \mathbb{P} \left( 0 \leq -(X - \mu_1)^\top \Sigma^{-1}(\mu_0 - \mu_1) + \frac{1}{2} \|\mu_1 - \mu_0\|_{\Sigma^{-1}}^2 \leq \log(u(\alpha)) \mid Y = 1 \right) \\ &= \mathbb{P} \left( 0 \leq Z + \frac{1}{2} \|\mu_1 - \mu_0\|_{\Sigma^{-1}} \leq \frac{\log(u(\alpha))}{\|\mu_1 - \mu_0\|_{\Sigma^{-1}}} \right) \\ &= \mathbb{P} \left( -\frac{1}{2} \|\mu_1 - \mu_0\|_{\Sigma^{-1}} \leq Z \leq \frac{\log(u(\alpha))}{\|\mu_1 - \mu_0\|_{\Sigma^{-1}}} - \frac{1}{2} \|\mu_1 - \mu_0\|_{\Sigma^{-1}} \right), \end{aligned}$$

where  $Z$  is normally distributed. In the same way, we get

$$\begin{aligned} A_1 = A_4 &= \mathbb{P}\left(-\frac{\|\mu_1 - \mu_0\|_{\Sigma^{-1}}}{2} \leq Z \leq \frac{\log(u(\alpha))}{\|\mu_1 - \mu_0\|_{\Sigma^{-1}}} - \frac{\|\mu_1 - \mu_0\|_{\Sigma^{-1}}}{2}\right) \\ A_2 = A_3 &= \mathbb{P}\left(\frac{\|\mu_1 - \mu_0\|_{\Sigma^{-1}}}{2} \leq Z \leq \frac{\log(u(\alpha))}{\|\mu_1 - \mu_0\|_{\Sigma^{-1}}} + \frac{\|\mu_1 - \mu_0\|_{\Sigma^{-1}}}{2}\right). \end{aligned}$$

Coming back to (24) and using twice the following relation  $\Phi(x) + \Phi(-x) = 1$  for any  $x \in \mathbb{R}$  which is valid for the normal distribution, we easily get

$$\begin{aligned} F_f^*(\alpha) &= \mathbb{P}(f^*(X) \leq \alpha) \\ &= \mathbb{P}\left(-\frac{\|\mu_1 - \mu_0\|_{\Sigma^{-1}}}{2} \leq Z \leq \frac{\log(u(\alpha))}{\|\mu_1 - \mu_0\|_{\Sigma^{-1}}} - \frac{\|\mu_1 - \mu_0\|_{\Sigma^{-1}}}{2}\right) \\ &\quad + \mathbb{P}\left(\frac{\|\mu_1 - \mu_0\|_{\Sigma^{-1}}}{2} \leq Z \leq \frac{\log(u(\alpha))}{\|\mu_1 - \mu_0\|_{\Sigma^{-1}}} + \frac{\|\mu_1 - \mu_0\|_{\Sigma^{-1}}}{2}\right) \\ &= \Phi\left(\frac{\log(u(\alpha))}{\|\mu_1 - \mu_0\|_{\Sigma^{-1}}} - \frac{\|\mu_1 - \mu_0\|_{\Sigma^{-1}}}{2}\right) + \Phi\left(\frac{\log(u(\alpha))}{\|\mu_1 - \mu_0\|_{\Sigma^{-1}}} + \frac{\|\mu_1 - \mu_0\|_{\Sigma^{-1}}}{2}\right) \\ &\quad - \left(\Phi\left(-\frac{\|\mu_1 - \mu_0\|_{\Sigma^{-1}}}{2}\right) + \Phi\left(\frac{\|\mu_1 - \mu_0\|_{\Sigma^{-1}}}{2}\right)\right) \\ &= \Phi\left(\frac{\|\mu_1 - \mu_0\|_{\Sigma^{-1}}}{2} + \frac{\log(u(\alpha))}{\|\mu_1 - \mu_0\|_{\Sigma^{-1}}}\right) - \Phi\left(\frac{\|\mu_1 - \mu_0\|_{\Sigma^{-1}}}{2} - \frac{\log(u(\alpha))}{\|\mu_1 - \mu_0\|_{\Sigma^{-1}}}\right) \\ &= \mathbb{P}\left(Z \in \left[\frac{\|\mu_1 - \mu_0\|_{\Sigma^{-1}}}{2} - \frac{\log(u(\alpha))}{\|\mu_1 - \mu_0\|_{\Sigma^{-1}}}, \frac{\|\mu_1 - \mu_0\|_{\Sigma^{-1}}}{2} + \frac{\log(u(\alpha))}{\|\mu_1 - \mu_0\|_{\Sigma^{-1}}}\right]\right). \quad (25) \end{aligned}$$

At this point, we are ready to evaluate the quantity  $F_f^*(f_{1-y}(X_\bullet))$  on the event  $\{Y_\bullet = y\}$  with  $y \in \{0, 1\}$ . Indeed, according to (25), we only need to evaluate  $\frac{\log(u(\alpha))}{\|\mu_1 - \mu_0\|_{\Sigma^{-1}}}$  for  $\alpha = f_{1-y}(X_\bullet)$ . Thanks to (22) we can write when  $Y_\bullet = y$

$$\frac{\log(u(f_{1-y}(X_\bullet)))}{\|\mu_1 - \mu_0\|_{\Sigma^{-1}}} = \frac{\log\left(\frac{p_{1-y}(X_\bullet)}{p_y(X_\bullet)}\right)}{\|\mu_1 - \mu_0\|_{\Sigma^{-1}}} = \frac{(X_\bullet - \mu_y)^\top \Sigma^{-1}(\mu_{1-y} - \mu_y)}{\|\mu_1 - \mu_0\|_{\Sigma^{-1}}} - \frac{1}{2}\|\mu_1 - \mu_0\|_{\Sigma^{-1}}.$$

Finally, using (25), we get when  $Y_\bullet = y$ ,

$$\begin{aligned} F_f^*(f_{1-y}(X_\bullet)) &= \\ &\mathbb{P}\left(Z \in \left[\|\mu_1 - \mu_0\|_{\Sigma^{-1}} - \frac{(X_\bullet - \mu_y)^\top \Sigma^{-1}(\mu_{1-y} - \mu_y)}{\|\mu_1 - \mu_0\|_{\Sigma^{-1}}}, \frac{(X_\bullet - \mu_y)^\top \Sigma^{-1}(\mu_{1-y} - \mu_y)}{\|\mu_1 - \mu_0\|_{\Sigma^{-1}}}\right]\right) = \\ &\Phi\left(\frac{(X_\bullet - \mu_y)^\top \Sigma^{-1}(\mu_{1-y} - \mu_y)}{\|\mu_1 - \mu_0\|_{\Sigma^{-1}}}\right) - \Phi\left(\|\mu_1 - \mu_0\|_{\Sigma^{-1}} - \frac{(X_\bullet - \mu_y)^\top \Sigma^{-1}(\mu_{1-y} - \mu_y)}{\|\mu_1 - \mu_0\|_{\Sigma^{-1}}}\right) = \\ &\Phi\left(\frac{(X_\bullet - \mu_y)^\top \Sigma^{-1}(\mu_{1-y} - \mu_y)}{\|\mu_1 - \mu_0\|_{\Sigma^{-1}}}\right) + \Phi\left(\frac{(X_\bullet - \mu_y)^\top \Sigma^{-1}(\mu_{1-y} - \mu_y)}{\|\mu_1 - \mu_0\|_{\Sigma^{-1}}} - \|\mu_1 - \mu_0\|_{\Sigma^{-1}}\right) - 1, \end{aligned}$$

where we have also used the relation  $\Phi(x) + \Phi(-x) = 1$ , for  $x \in \mathbb{R}$  in the last line, since  $\Phi$  is the normal cumulative distribution function. This ends the proof.  $\square$

The next result is the key tool in the proof of Proposition 3.

**Proposition 6.** Let  $\varepsilon \in ]0, 1]$ . For  $y \in \{0, 1\}$ , we have

$$\begin{aligned} & \mathbb{P}(s^*(X_\bullet) \neq Y_\bullet, F_f^*(f^*(X_\bullet)) \geq 1 - \varepsilon \mid Y_\bullet = y) \\ &= \mathbb{P}\left(\{\Phi(Z) + \Phi(Z - \|\mu_1 - \mu_0\|_{\Sigma^{-1}})\} \geq 2 - \varepsilon, Z \geq \frac{\|\mu_1 - \mu_0\|_{\Sigma^{-1}}}{2}\right), \end{aligned}$$

where  $Z \sim \mathcal{N}(0, 1)$ .

*Proof.* Let  $\varepsilon \in ]0, 1]$ . For  $y \in \{0, 1\}$ , according to the first equivalence stated in (22), we observe that

$$\begin{aligned} & \mathbb{P}(F_f^*(f^*(X_\bullet)) \geq 1 - \varepsilon, s^*(X_\bullet) \neq Y_\bullet \mid Y_\bullet = y) = \\ & \mathbb{P}(F_f^*(f_{1-y}(X_\bullet)) \geq 1 - \varepsilon, f^*(X_\bullet) = f_{1-y}(X_\bullet) \mid Y_\bullet = y). \end{aligned} \quad (26)$$

Moreover, using the last equivalence in (22), we have, if  $Y_\bullet = y$

$$f^*(X_\bullet) = f_{1-y}(X_\bullet) \Leftrightarrow \frac{(X_\bullet - \mu_y)^\top \Sigma^{-1}(\mu_{1-y} - \mu_y)}{\|\mu_1 - \mu_0\|_{\Sigma^{-1}}} - \frac{1}{2}\|\mu_1 - \mu_0\|_{\Sigma^{-1}} \geq 0. \quad (27)$$

Then we just need to rewrite the event  $\{F_f^*(f_{1-y}(X_\bullet)) \geq 1 - \varepsilon\}$ , when conditioned on the event  $\{Y_\bullet = y\}$ , in a convenient way. Using Proposition 5, we can write that when  $Y_\bullet = y$

$$\begin{aligned} F_f^*(f_{1-y}(X_\bullet)) &= \Phi\left(\frac{(X_\bullet - \mu_y)^\top \Sigma^{-1}(\mu_{1-y} - \mu_y)}{\|\mu_1 - \mu_0\|_{\Sigma^{-1}}}\right) + \\ & \Phi\left(\frac{(X_\bullet - \mu_y)^\top \Sigma^{-1}(\mu_{1-y} - \mu_y)}{\|\mu_1 - \mu_0\|_{\Sigma^{-1}}} - \|\mu_1 - \mu_0\|_{\Sigma^{-1}}\right) - 1. \end{aligned} \quad (28)$$

Plugging (27) and (28) into (26), we finally then get

$$\begin{aligned} & \mathbb{P}(F_f^*(f(X_\bullet)) \geq 1 - \varepsilon, s^*(X_\bullet) \neq Y_\bullet \mid Y_\bullet = y) \\ &= \mathbb{P}\left(\{\Phi(Z_\bullet) + \Phi(Z_\bullet - \|\mu_1 - \mu_0\|_{\Sigma^{-1}})\} \geq 2 - \varepsilon, Z_\bullet \geq \frac{\|\mu_1 - \mu_0\|_{\Sigma^{-1}}}{2}\right), \end{aligned}$$

where  $Z_\bullet \sim \mathcal{N}(0, 1)$ . The last equality is due to the fact that given  $Y_\bullet = 1$ , the random variable  $\frac{(X_\bullet - \mu_y)^\top \Sigma^{-1}(\mu_{1-y} - \mu_y)}{\|\mu_1 - \mu_0\|_{\Sigma^{-1}}}$  is normally distributed. We then get the desired result and the proof of the proposition is completed.  $\square$

### 6.3.2 Proposition 3

Let  $\varepsilon \in ]0, 1]$ . Since  $\mathbb{P}(Y_\bullet = 1) = \mathbb{P}(Y_\bullet = 0) = 1/2$ , we have

$$\begin{aligned} \mathbb{P}(s^*(X_\bullet) \neq Y_\bullet, F_f^*(f^*(X_\bullet)) \geq 1 - \varepsilon) &= \frac{1}{2} \{\mathbb{P}(s^*(X_\bullet) \neq Y_\bullet, F_f^*(f^*(X_\bullet)) \geq 1 - \varepsilon \mid Y_\bullet = 1) \\ &+ \mathbb{P}(s^*(X_\bullet) \neq Y_\bullet, F_f^*(f^*(X_\bullet)) \geq 1 - \varepsilon \mid Y_\bullet = 0)\}. \end{aligned}$$

Next, using Proposition 6, we get

$$\begin{aligned} & \mathbb{P}(s^*(X_\bullet) \neq Y_\bullet, F_f^*(f^*(X_\bullet)) \geq 1 - \varepsilon) \\ &= \mathbb{P}\left(\{\Phi(Z) + \Phi(Z - \|\mu_1 - \mu_0\|_{\Sigma^{-1}})\} \geq 2 - \varepsilon, Z \geq \frac{\|\mu_1 - \mu_0\|_{\Sigma^{-1}}}{2}\right) \\ &= \mathbb{P}(\{\Phi(Z) + \Phi(Z - \|\mu_1 - \mu_0\|_{\Sigma^{-1}})\} \geq 2 - \varepsilon). \end{aligned}$$

The last equality is due to the following property:

$$Z < \frac{\|\mu_1 - \mu_0\|_{\Sigma^{-1}}}{2} \Rightarrow \Phi(Z - \|\mu_1 - \mu_0\|_{\Sigma^{-1}}) < \Phi\left(-\frac{\|\mu_1 - \mu_0\|_{\Sigma^{-1}}}{2}\right) = 1 - \Phi\left(\frac{\|\mu_1 - \mu_0\|_{\Sigma^{-1}}}{2}\right),$$

which implies that

$$Z < \frac{\|\mu_1 - \mu_0\|_{\Sigma^{-1}}}{2} \Rightarrow \Phi(Z) + \Phi(Z - \|\mu_1 - \mu_0\|_{\Sigma^{-1}}) < 1 \leq 2 - \varepsilon.$$

The end of the proof is straightforward and follows from the relation  $\Phi(x) + \Phi(-x) = 1, \forall x \in \mathbb{R}$ . Indeed, we have

$$\begin{aligned} \mathbb{P}(\Phi(Z) + \Phi(Z - \|\mu_1 - \mu_0\|_{\Sigma^{-1}}) \geq 2 - \varepsilon) &= \mathbb{P}(\Phi(-Z) + \Phi(-Z + \|\mu_1 - \mu_0\|_{\Sigma^{-1}}) \leq \varepsilon) \\ &= \mathbb{P}(\Phi(Z) + \Phi(Z + \|\mu_1 - \mu_0\|_{\Sigma^{-1}}) \leq \varepsilon), \end{aligned}$$

since  $Z$  and  $-Z$  equal in law. This ends the proof.

## 6.4 Proof of Proposition 4

We first define the following events

$$\begin{aligned} \mathcal{A}_y &= \{f^*(X_\bullet) \geq \alpha_\varepsilon, \hat{f}(X_\bullet) < \hat{\alpha}_\varepsilon, s^*(X_\bullet) \neq y\}, \quad y = 0, 1 \\ \mathcal{B}_y &= \{f^*(X_\bullet) < \alpha_\varepsilon, \hat{f}(X_\bullet) \geq \hat{\alpha}_\varepsilon, \hat{s}(X_\bullet) \neq y\}, \quad y = 0, 1. \\ \mathcal{C}_y &= \{f^*(X_\bullet) \geq \alpha_\varepsilon, \hat{f}(X_\bullet) \geq \hat{\alpha}_\varepsilon, s^*(X_\bullet) \neq \hat{s}(X_\bullet), s^*(X_\bullet) \neq y\}, \quad y = 0, 1. \end{aligned}$$

Since  $\mathcal{P}(\tilde{\Gamma}_\varepsilon^\bullet) = \varepsilon$ , we can apply Proposition 2 and then, as

$$|2\eta^*(X_\bullet) - 1| \leq |\eta^*(X_\bullet) - \alpha_\varepsilon| + |1 - \eta^*(X_\bullet) - \alpha_\varepsilon|,$$

we deduce that

$$\begin{aligned} \mathbf{R}(\tilde{\Gamma}_\varepsilon^\bullet) - \mathcal{R}(\Gamma_\varepsilon^\bullet) &\leq \\ &\frac{1}{\varepsilon} \{ \mathbf{E}[|\eta^*(X_\bullet) - \alpha_\varepsilon| \mathbf{1}_{\mathcal{A}_0 \cup \mathcal{B}_0 \cup \mathcal{C}_0 \cup \mathcal{C}_1}] + \mathbf{E}[|1 - \eta^*(X_\bullet) - \alpha_\varepsilon| \mathbf{1}_{\mathcal{A}_1 \cup \mathcal{B}_1 \cup \mathcal{C}_0 \cup \mathcal{C}_1}] \}. \quad (29) \end{aligned}$$

Now,

1. on  $\mathcal{A}_0$ ,  $f^* = \eta^*$ ,  $\eta^*(X_\bullet) \geq \alpha_\varepsilon$  and  $\hat{f}(X_\bullet) < \hat{\alpha}_\varepsilon$ ,  
hence, we have  $|\eta^*(X_\bullet) - \alpha_\varepsilon| \leq |\hat{\eta}(X_\bullet) - \eta^*(X_\bullet)|$  except if  $\alpha_\varepsilon \leq \hat{\alpha}_\varepsilon$  and  $\hat{f}(X_\bullet) \in (\alpha_\varepsilon, \hat{\alpha}_\varepsilon)$ ;
2. on  $\mathcal{B}_0$ ,  $\hat{f} = \hat{\eta}$ ,  $\hat{\eta}(X_\bullet) \geq \hat{\alpha}_\varepsilon$  and  $f^*(X_\bullet) < \alpha_\varepsilon$ ,  
hence, we have  $|\eta^*(X_\bullet) - \alpha_\varepsilon| \leq |\hat{\eta}(X_\bullet) - \eta^*(X_\bullet)|$  except if  $\hat{\alpha}_\varepsilon \leq \alpha_\varepsilon$  and  $\hat{f}(X_\bullet) \in (\hat{\alpha}_\varepsilon, \alpha_\varepsilon)$ ;
3. on  $\mathcal{C}_0$ ,  $f^* = \eta^*$ ,  $\hat{f} = 1 - \hat{\eta}$ ,  $\eta^*(X_\bullet) \geq \alpha_\varepsilon$  and  $\hat{\eta}(X_\bullet) \leq 1/2$ ,  
hence, we always have  $|\eta^*(X_\bullet) - \alpha_\varepsilon| \leq |\hat{\eta}(X_\bullet) - \eta^*(X_\bullet)|$ ;
4. on  $\mathcal{C}_1$ ,  $f^* = 1 - \eta^*$ ,  $\hat{f} = \hat{\eta}$  and  $\hat{\eta}(X_\bullet) \geq \hat{\alpha}_\varepsilon$ ,  
hence, we have  $|\eta^*(X_\bullet) - \alpha_\varepsilon| \leq |\hat{\eta}(X_\bullet) - \eta^*(X_\bullet)|$  except if  $\hat{\alpha}_\varepsilon \leq \alpha_\varepsilon$  and  $\hat{f}(X_\bullet) \in (\hat{\alpha}_\varepsilon, \alpha_\varepsilon)$ .

Since  $\mathcal{A}_0, \mathcal{B}_0, \mathcal{C}_0$  and  $\mathcal{C}_1$  are mutually exclusive events, we deduce

$$\begin{aligned} \mathbf{E} [|\eta^*(X_\bullet) - \alpha_\varepsilon| \mathbf{1}_{\mathcal{A}_0 \cup \mathcal{B}_0 \cup \mathcal{C}_0 \cup \mathcal{C}_1}] &\leq \mathbf{E} [|\eta^*(X_\bullet) - \alpha_\varepsilon| \mathbf{1}_{\{|\hat{\eta}(X_\bullet) - \eta^*(X_\bullet)| \geq |\eta^*(X_\bullet) - \alpha_\varepsilon|\}}] + \\ &\mathbf{E} [|\eta^*(X_\bullet) - \alpha_\varepsilon| \left( \mathbf{1}_{\{\mathcal{A}_0, \alpha_\varepsilon \leq \hat{\alpha}_\varepsilon, \hat{f}(X_\bullet) \in (\alpha_\varepsilon, \hat{\alpha}_\varepsilon)\}} + \mathbf{1}_{\{\mathcal{B}_0 \cup \mathcal{C}_1, \hat{\alpha}_\varepsilon \leq \alpha_\varepsilon, \hat{f}(X_\bullet) \in (\hat{\alpha}_\varepsilon, \alpha_\varepsilon)\}} \right)]. \end{aligned} \quad (30)$$

In the same way, we obtain the following decomposition

$$\begin{aligned} \mathbf{E} [1 - \eta^*(X_\bullet) - \alpha_\varepsilon | \mathbf{1}_{\mathcal{A}_1 \cup \mathcal{B}_1 \cup \mathcal{C}_0 \cup \mathcal{C}_1}] &\leq \mathbf{E} [1 - \eta^*(X_\bullet) - \alpha_\varepsilon | \mathbf{1}_{\{|\hat{\eta}(X_\bullet) - \eta^*(X_\bullet)| \geq |1 - \eta^*(X_\bullet) - \alpha_\varepsilon|\}}] + \\ &\mathbf{E} [1 - \eta^*(X_\bullet) - \alpha_\varepsilon | \left( \mathbf{1}_{\{\mathcal{A}_1, \alpha_\varepsilon \leq \hat{\alpha}_\varepsilon, \hat{f}(X_\bullet) \in (\alpha_\varepsilon, \hat{\alpha}_\varepsilon)\}} + \mathbf{1}_{\{\mathcal{B}_1 \cup \mathcal{C}_0, \hat{\alpha}_\varepsilon \leq \alpha_\varepsilon, \hat{f}(X_\bullet) \in (\hat{\alpha}_\varepsilon, \alpha_\varepsilon)\}} \right)]. \end{aligned} \quad (31)$$

Since  $(\mathcal{A}_y, \mathcal{B}_y, \mathcal{C}_y), y = 0, 1$  are mutually exclusive events, and that  $|\eta^*(X_\bullet) - \alpha_\varepsilon| \leq \alpha_\varepsilon$  and  $|1 - \eta^*(X_\bullet) - \alpha_\varepsilon| \leq \alpha_\varepsilon$ , it derives from Inequalities (29), (30) and (31) that

$$\begin{aligned} \mathbf{R} \left( \tilde{\Gamma}_\varepsilon^\bullet \right) - \mathcal{R} \left( \Gamma_\varepsilon^\bullet \right) &\leq \frac{1}{\varepsilon} \\ &\quad \left\{ \mathbf{E} [|\eta^*(X_\bullet) - \alpha_\varepsilon| \mathbf{1}_{\{|\hat{\eta}(X_\bullet) - \eta^*(X_\bullet)| \geq |\eta^*(X_\bullet) - \alpha_\varepsilon|\}}] + \right. \\ &\quad \left. \mathbf{E} [1 - \eta^*(X_\bullet) - \alpha_\varepsilon | \mathbf{1}_{\{|\hat{\eta}(X_\bullet) - \eta^*(X_\bullet)| \geq |1 - \eta^*(X_\bullet) - \alpha_\varepsilon|\}}] + \right. \\ &\quad \left. \alpha_\varepsilon |F_{\hat{f}}(\hat{\alpha}_\varepsilon) - F_{\hat{f}}(\alpha_\varepsilon)| \right\}. \end{aligned}$$

To conclude the proof, it remains to note that  $1 - \varepsilon = F_{\hat{f}}(\hat{\alpha}_\varepsilon) = F_{\hat{f}}(\alpha_\varepsilon)$ , for all  $\varepsilon \in ]0, 1[$ .

## 6.5 Proof of Theorem 1

We first set a Lemma that will be used in the proof.

### 6.5.1 Tool lemma

The following lemma is inspired by Lemma 3.1 in [AT07].

**Lemma 1.** *Let  $X$  be a real random variable,  $(X_n)_{n \geq 1}$  a be sequence of real random variables and  $t_0 \in \mathbb{R}$ . Assume that there exist  $C_1 < \infty$  and  $\gamma_0 > 0$  such that*

$$\mathbb{P}_X (|X - t_0| \leq \delta) \leq C_1 \delta^{\gamma_0}, \quad \forall \delta > 0,$$

and a sequence of positive numbers  $a_n \rightarrow +\infty$ ,  $C_2, C_3$  some positive constants such that

$$\mathbb{P}_{X_n} (|X_n - X| \geq \delta | X) \leq C_2 \exp(-C_3 a_n \delta^2), \quad \forall \delta > 0, \forall n \in \mathbb{N}.$$

Then, there exists  $C > 0$  depending only on  $C_1, C_2$  and  $C_3$ , such that

$$\begin{aligned} \left| \mathbf{E} [\mathbf{1}_{\{X_n \geq t_0\}} - \mathbf{1}_{\{X \geq t_0\}}] \right| &\leq \mathbf{E} [|\mathbf{1}_{\{X_n \geq t_0\}} - \mathbf{1}_{\{X \geq \alpha_\varepsilon\}}|] \\ &\leq \mathbf{P} (|X_n - X| \geq |X - t_0|) \\ &\leq C a_n^{-\gamma_0/2}. \end{aligned}$$

*Proof.* The following inequality holds

$$|\mathbf{1}_{\{X_n \geq t_0\}} - \mathbf{1}_{\{X \geq t_0\}}| \leq \mathbf{1}_{\{|X_n - X| \geq |X - t_0|\}}.$$

Hence, it remains to prove

$$\mathbf{P} (|X_n - X| \geq |X - t_0|) \leq C a_n^{-\gamma_0/2}.$$



We define, for  $\delta > 0$ ,

$$\begin{aligned}\mathbf{A}_0 &= \{|X - t_0| \leq \delta\} \\ \mathbf{A}_j &= \{2^{j-1}\delta < |X - t_0| \leq 2^j\delta\}, \quad j \geq 1.\end{aligned}$$

Since the events  $(\mathbf{A}_j)_{j \geq 0}$  are mutually exclusive, we deduce

$$\begin{aligned}\mathbf{P}(|X_n - X| \geq |X - t_0|) &= \sum_{j \geq 0} \mathbf{E} [\mathbf{1}_{\{|X_n - X| \geq |X - t_0|\}}] \mathbf{1}_{\mathbf{A}_j} \\ &\leq \mathbb{P}_X(|X - t_0| \leq \delta) + \sum_{j \geq 1} \mathbf{E} [\mathbf{1}_{\{|X_n - X| \geq 2^{j-1}\delta\}}] \mathbf{1}_{\mathbf{A}_j} \\ &\leq C_1 \delta^{\gamma_0} + \sum_{j \geq 1} \mathbb{E}_X [\mathbb{P}_{X_n}(|X_n - X| \geq 2^{j-1}\delta | X) \mathbf{1}_{\mathbf{A}_j}] \\ &\leq C_1 \delta^{\gamma_0} + C_1 C_2 \delta^{\gamma_0} \sum_{j \geq 1} 2^{j\gamma_0} \exp(-C_2 a_n 2^{2j-2} \delta^2),\end{aligned}$$

since  $\mathbb{P}_X(\mathbf{A}_j) \leq \mathbb{P}_X(|X - t_0| \leq 2^j\delta) \leq (2^j\delta)^{\gamma_0}$ . Therefore, choosing  $\delta = a_n^{-1/2}$ , we obtain from the above inequality,

$$\begin{aligned}\mathbf{P}(|X_n - X| \geq |X - t|) &\leq C_1 a_n^{-\gamma_0/2} + 2C_1 C_3 a_n^{-\gamma_0/2} \sum_{j \geq 1} 2^{j\gamma_0} \exp(-C_2 2^{2j-2}) \\ &\leq C a_n^{-\gamma_0/2},\end{aligned}$$

for a constant  $C > 0$ . □

### 6.5.2 Theorem 1

Let  $\varepsilon \in ]0, 1[$ . We first prove that for  $N$  large enough

$$\left| \mathbf{P} \left( \hat{F}_{\hat{f}}(\hat{f}(X_{\bullet})) \geq 1 - \varepsilon \right) - \mathbf{P} \left( F_{\hat{f}}(\hat{f}(X_{\bullet})) \geq 1 - \varepsilon \right) \right| \leq \frac{\tilde{C}}{\sqrt{N}}, \quad (32)$$

and

$$\left| \mathbf{R} \left( \hat{\Gamma}_{\varepsilon}^{\bullet} \right) - \mathbf{R} \left( \tilde{\Gamma}_{\varepsilon}^{\bullet} \right) \right| \leq \frac{C}{\sqrt{N}}, \quad (33)$$

where  $C, \tilde{C} > 0$  are constants which do not depend on  $n$ . For all  $x \in [1/2, 1]$ ,

$$F_{\hat{f}}(x) = \mathbb{E}_{\mathcal{D}_n} \left[ \mathbb{P} \left( \hat{f}(X) \leq x | \mathcal{D}_n \right) \right],$$

Hence, conditional on  $\mathcal{D}_n$ ,  $\hat{F}_{\hat{f}}(x)$  is the empirical cumulative distribution function of  $\hat{f}(X)$ , where  $\hat{f}$  is view as a deterministic function. Therefore, for all  $\gamma \geq \sqrt{\log(2)/2N}$ , Dvoretzky-Kiefer-Wolfowitz Inequality yields

$$\begin{aligned}\mathbb{P}_{\mathcal{D}_N} \left( |\hat{F}_{\hat{f}}(\hat{f}(X_{\bullet})) - F_{\hat{f}}(\hat{f}(X_{\bullet}))| \geq \gamma | \mathcal{D}_n, X_{\bullet} \right) &\leq \mathbb{P}_{\mathcal{D}_N} \left( \sup_{x \in [1/2, 1]} |\hat{F}_{\hat{f}}(x) - U_n(x)| \geq \gamma | \mathcal{D}_n \right) \\ &\leq 2 \exp(-2N\gamma^2),\end{aligned}$$

where  $U_n(x) = \mathbb{P}(\hat{f}(X_\bullet) \leq x | \mathcal{D}_n)$ .

Applying Lemma 1, we get

$$\left| \mathbb{E}_{(\mathcal{D}_N, X_\bullet)} \left[ \mathbf{1}_{\{\hat{F}_f(\hat{f}(X_\bullet)) \geq 1-\varepsilon\}} - \mathbf{1}_{\{F_f(\hat{f}(X_\bullet)) \geq 1-\varepsilon\}} | \mathcal{D}_n \right] \right| \leq \frac{\tilde{C}}{\sqrt{N}},$$

where  $\tilde{C}$  does not depend on  $n$ . Hence, we obtain Inequality (32). In the same way, we have

$$\left| \mathbf{P} \left( \hat{F}_f(\hat{f}(X_\bullet)) \geq 1 - \varepsilon, \hat{s}(X_\bullet) \neq Y_\bullet \right) - \mathbf{P} \left( F_f(\hat{f}(X_\bullet)) \geq 1 - \varepsilon, \hat{s}(X_\bullet) \neq Y_\bullet \right) \right| \leq \frac{\tilde{C}}{\sqrt{N}}.$$

Therefore, Inequality (33) holds for some constant  $C > 0$ .

Since, by Assumption (A2)  $\mathbf{P} \left( F_f(\hat{f}(X_\bullet)) \geq 1 - \varepsilon \right) = \varepsilon$ , Inequality (32) yields

$$\mathbf{P} \left( \hat{F}_f(\hat{f}(X_\bullet)) \geq 1 - \varepsilon \right) = \varepsilon + O(N^{-1/2}).$$

Now, we conclude the point 1) of the theorem. Since Inequality (33) ensures that

$$\left| \mathbf{R} \left( \hat{\Gamma}_\varepsilon^\bullet \right) - \mathbf{R} \left( \tilde{\Gamma}_\varepsilon^\bullet \right) \right| \rightarrow 0, \quad n, N \rightarrow +\infty,$$

it remains to prove that

$$\left| \mathbf{R} \left( \tilde{\Gamma}_\varepsilon^\bullet \right) - \mathcal{R} \left( \Gamma_\varepsilon^\bullet \right) \right| \rightarrow 0, \quad n \rightarrow +\infty.$$

Applying Proposition 3, we obtain for  $\delta_n > 0$ ,  $\delta_n \rightarrow 0$

$$\left| \mathbf{R} \left( \tilde{\Gamma}_\varepsilon^\bullet \right) - \mathcal{R} \left( \Gamma_\varepsilon^\bullet \right) \right| \leq 2\delta_n + 2\mathbf{P} \left( |\hat{\eta}(X_\bullet) - \eta^*(X_\bullet)| \geq \delta_n \right) + \left| F_f(\alpha_\varepsilon) - F_f^*(\alpha_\varepsilon) \right|.$$

Since,  $\hat{\eta}(X_\bullet) \rightarrow \eta^*(X_\bullet)$  in probability when  $n \rightarrow +\infty$ ,  $\hat{f}(X_\bullet) \rightarrow f^*(X_\bullet)$  in distribution and  $\left| F_f(\alpha_\varepsilon) - F_f^*(\alpha_\varepsilon) \right| \rightarrow 0$ . Moreover,  $\mathbf{P} \left( |\hat{\eta}(X_\bullet) - \eta^*(X_\bullet)| \geq \delta_n \right) \rightarrow 0$  which concludes the point 1) of the proof.

Finally, to prove 2), it remains to show that

$$\left| \mathbf{R} \left( \tilde{\Gamma}_\varepsilon^\bullet \right) - \mathcal{R} \left( \Gamma_\varepsilon^\bullet \right) \right| = O(a_n^{-\gamma_\varepsilon/2}),$$

We first note that,

$$\begin{aligned} \left| F_f(\alpha_\varepsilon) - F_f^*(\alpha_\varepsilon) \right| &\leq \mathbf{E} \left[ \left| \mathbf{1}_{\{\hat{f}(X_\bullet) \geq \alpha_\varepsilon\}} - \mathbf{1}_{\{f^*(X_\bullet) \geq \alpha_\varepsilon\}} \right| \right] \\ &\leq \mathbf{E} \left[ \left| \mathbf{1}_{\{\hat{f}(X_\bullet) \geq \alpha_\varepsilon\}} - \mathbf{1}_{\{f^*(X_\bullet) \geq \alpha_\varepsilon\}} \right| \mathbf{1}_{\{|\hat{\eta}(X_\bullet) - \eta^*(X_\bullet)| \geq |\eta^*(X_\bullet) - 1/2|\}} \right] \\ &\quad + \mathbf{E} \left[ \left| \mathbf{1}_{\{\hat{f}(X_\bullet) \geq \alpha_\varepsilon\}} - \mathbf{1}_{\{f^*(X_\bullet) \geq \alpha_\varepsilon\}} \right| \mathbf{1}_{\{|\hat{\eta}(X_\bullet) - \eta^*(X_\bullet)| < |\eta^*(X_\bullet) - 1/2|\}} \right] \\ &\leq \mathbf{P} \left( |\hat{\eta}(X_\bullet) - \eta^*(X_\bullet)| \geq |\eta^*(X_\bullet) - \alpha_\varepsilon| \right) \\ &\quad + \mathbf{P} \left( |\hat{\eta}(X_\bullet) - \eta^*(X_\bullet)| \geq |\eta^*(X_\bullet) - (1 - \alpha_\varepsilon)| \right) \\ &\quad + \mathbf{E} \left[ \left| \mathbf{1}_{\{\hat{\eta}(X_\bullet) \geq \alpha_\varepsilon\}} - \mathbf{1}_{\{\eta^*(X_\bullet) \geq \alpha_\varepsilon\}} \right| \right]. \end{aligned}$$

Therefore, applying both Proposition 3, Lemma 1 and the above inequality, we get the desired result.

## References

- [AT07] J.-Y. Audibert and A. Tsybakov. Fast learning rates for plug-in classifiers. *Ann. Statist.*, 35(2):608–633, 2007.
- [BW08] P. Bartlett and M. Wegkamp. Classification with a reject option using a hinge loss. *J. Mach. Learn. Res.*, 9:1823–1840, 2008.
- [Cho70] C. Chow. On optimum error and reject trade-off. *IEEE Transactions on Information Theory*, 16:41–46, 1970.
- [GKKW02] L. Györfi, M. Kohler, A. Krzyżak, and H. Walk. *A distribution-free theory of non-parametric regression*. Springer Series in Statistics. Springer-Verlag, New York, 2002.
- [GRKC09] Y. Grandvalet, A. Rakotomamonjy, J. Keshet, and S. Canu. Support vector machines with a reject option. In *Advances in Neural Information Processing Systems 21 (NIPS 2008)*, pages 537–544. MIT Press, 2009.
- [HW06] R. Herbei and M. Wegkamp. Classification with reject option. *Canad. J. Statist.*, 34(4):709–721, 2006.
- [Lei14] J. Lei. Classification with confidence. *Biometrika*, 101(4):755–769, 2014.
- [Mas90] P. Massart. The tight constant in the Dvoretzky-Kiefer-Wolfowitz inequality. *Ann. Probab.*, 18(3):1269–1283, 1990.
- [NZH10] M. Naadeem, J. Zucker, and B. Hanczar. Accuracy-rejection curves (ARCs) for comparing classification methods with a reject option. In *MLSB*, pages 65–81, 2010.
- [Tsy04] A. Tsybakov. Optimal aggregation of classifiers in statistical learning. *Ann. Statist.*, 32(1):135–166, 2004.
- [Vap98] V. Vapnik. *Statistical learning theory*. Adaptive and Learning Systems for Signal Processing, Communications, and Control. John Wiley & Sons Inc., New York, 1998. A Wiley-Interscience Publication.
- [vdV98] A. van der Vaart. *Asymptotic statistics*, volume 3 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge, 1998.
- [VGS99] V. Vovk, A. Gammerman, and C. Saunders. Machine-learning applications of algorithmic randomness. In *Proceedings of the 16<sup>th</sup> International Conference on Machine Learning*, pages 444–453. 1999.
- [VGS05] V. Vovk, A. Gammerman, and G. Shafer. *Algorithmic learning in a random world*. Springer, New York, 2005.
- [WY11] M. Wegkamp and M. Yuan. Support vector machines with a reject option. *Bernoulli*, 17(4):1368–1385, 2011.