Confidence sets with expected sizes for Multiclass Classification

Christophe Denis, Mohamed Hebiri

Confidence sets with expected sizes for Multiclass Classification

Christophe Denis* and Mohamed Hebiri†
LAMA, UMR-CNRS 8050,
Université Paris Est – Marne-la-Vallée

Abstract

Multiclass classification problems such as image annotation can involve a large number of classes. In this context, confusion between classes can occur, and single label classification may be misleading. We provide in the present paper a general device that, given an unlabeled dataset and a score function defined as the minimizer of some empirical and convex risk, outputs a set of class labels, instead of a single one. Interestingly, this procedure does not require that the unlabeled dataset explores the whole classes. Even more, the method is calibrated to control the expected size of the output set while minimizing the classification risk. We show the statistical optimality of the procedure and establish rates of convergence under the Tsybakov margin condition. It turns out that these rates are linear on the number of labels. We apply our methodology to convex aggregation of confidence sets based on the V-fold cross validation principle also known as the superlearning principle [vdLPH07]. We illustrate the numerical performance of the procedure on real data and demonstrate in particular that with moderate expected size, w.r.t. the number of labels, the procedure provides significant improvement of the classification risk.

Keywords : Multiclass classification, confidence sets, empirical risk minimization, cumulative distribution functions, convex loss, superlearning.

1 Introduction

The advent of high-throughput technology has generated tremendous amounts of large and high-dimensional classification data. This allows classification at unprecedented scales with hundreds or even more classes. The standard approach to classification in the multiclass setting is to use a classification rule for assigning a single label. More specifically, it consists in assigning a single label $Y \in \mathcal{Y}$, with $\mathcal{Y} = \{1, \dots, K\}$, to a given input example $X \in \mathcal{X}$ among a collection of labels. However, while a large number of classes can lead to precise characterizations of data points, similarities among classes also bear the risk of confusion and misclassification. Hence, assigning a single label can lead to wrong or ambiguous results.

In this paper, we address this problem by introducing an approach that yields sets of labels as outputs, namely, confidence sets. A confidence set Γ is a function that maps $\mathcal{X}$ onto $2^{\mathcal{Y}}$. A natural way to obtain a set of labels is to use ranked outputs of the classification rule. For example, one could take the classes that correspond to the top-level conditional probabilities $\mathbb{P}(Y = \cdot | X = x)$. Here, we provide a more general approach where we control the *expected* size of

*Christophe.Denis@u-pem.fr
†Mohamed.Hebiri@u-pem.fr

the confidence sets. For a confidence set Γ, the expected size of Γ is defined as $\mathbb{E}[|\Gamma(X)|]$, where $|\cdot|$ stands for the cardinality. For a sample X and given an expected set size β, we provide an algorithm that outputs a set $\hat{\Gamma}(X)$ such that $\mathbb{E}[|\hat{\Gamma}(X)|] \approx \beta$. Furthermore, the procedure aims at minimizing the classification error given by

$$\mathbb{P}\left(Y \notin \hat{\Gamma}(X)\right) \approx \min_{\Gamma\,:\,\mathbb{E}[|\Gamma(X)|]=\beta} \mathbb{P}\left(Y \notin \Gamma(X)\right) = \mathcal{R}^*_\beta.$$

We establish a close formula of the oracle confidence set $\Gamma^* = \operatorname{argmin}_{\Gamma\,:\,\mathbb{E}[|\Gamma(X)|]=\beta} \mathbb{P}\left(Y \notin \Gamma(X)\right)$ that involves the cumulative distribution functions of the conditional probabilities. Besides, we formulate a data-driven counterpart of Γ^* based on the minimization of the empirical risk. However, the natural risk function in the multiclass setting is non convex, and then minimizing it is proving to be a computational issue. As a remedy, convex surrogates are often used in machine learning, and more specifically in classification as reflected by the popularity of methods such as Boosting [FS97], logistic regression [FHT00] and support vector machine [Vap98]. In our problem this translates by considering some convex surrogate of the 0/1-loss in $\mathbb{P}\left(Y \notin \Gamma(X)\right)$; we introduce a new convex risk function which enables us to emphasize specific aspects of confidence sets. That convex risk function is partly inspired by the works in [Zha04, BJM06, YW10] that deal with binary classification.

Our approach displays two main features. First, our method can be implemented in a semi-supervised way [Vap98]. More precisely, the method is a two-steps procedure that requires the estimation of score functions (as minimizers of some convex risk function) in a first step and the estimation of the cumulative distribution of these scores in a second one. Hence, the first step requires labeled data whereas the second one involves only unlabeled samples. Notably, the unlabeled sample does not necessary consists of examples coming from the whole classes. This aspects is fondamental when we deal with a large number of classes, some of which been sparsely represented. Second, from the theoretical point of view, we provide an oracle inequality satisfied by $\hat{\Gamma}$, the empirical confidence set that results from the minimization of an empirical convex risk. The obtained oracle inequality shall enable us to derive rates of convergence on a particular class of confidence sets under the Tsybakov noise assumption on the data generating distribution. These rates are linear in the number of classes K and also depend on the regularity of the cumulative distribution functions of the conditional probabilities.

An obvious benefit of considering convex risk minimization is when we deal with aggregation. However, aggregating confidence sets is no simple matter. Another contribution of the present paper is about applying the above methodology to aggregation of confidence sets. More specifically, we provide a generalization of the superlearning algorithm [vdLPH07] initially introduced in the context of regression and binary classification. This algorithm relies on the V-fold cross-validation principle. We prove the consistency of this aggregation procedure and illustrate its relevance on real datasets. Let us point out that any arbitrary library of machine learning algorithms may be used in this aggregation procedure; we propose in the present paper to exploit support vector machines, random forest procedures or softmax regression since these methods are popular in machine learning and that each of them is associated to a different shape.

We end up this discussion by highlighting in two words what we perceive as being the main contributions of the present paper. We describe an optimal strategy for building confidence sets in multiclass classification setting, and we derive a new aggregation procedure for confidence sets that still allows controlling the expected size of the resulting confidence set.

Related works: The closest learning task to the present work is *classification with reject option* which is a particular setting in binary classification. Several papers fall within the scope of this area [Cho70, HW06, YW10, WY11, Lei14, DH15] and differ from each other by the goal they

consider. Among these references, our procedure is partially inspired by the paper [DH15] that also considers a semi-supervised approach to build confidence sets invoking some cumulative distribution functions (of the conditional probabilities themselves in their case). The similarity is however limited to the definition of oracle confidence sets, the oracle confidence in the present paper being an extension of the one defined in [DH15] to the multiclass setting. On the other hand, all the data-driven considerations are completely different (in particular, [DH15] focuses on plug-in rules) and importantly, we develop here new probabilistic results on sums of cumulative distribution functions of random variables, that are of own interest.

Assigning a set of labels instead of a single one for an input example is not new [VGS05, WLW04, dCDB09, LRW13, CCB16]. One of the most popular methods is based on *Conformal Prediction* approach [VGS99, Vov02, VGS05]. In the multiclass classification framework, the goal of this algorithm is to build the smallest set of labels such that its classification error is below a pre-specified level. Since our procedure aims at minimizing the classification error while keeping under control the size of the set, *Conformal Prediction* can be seen as a dual of our method. It is worth mentioning that conformal predictors approaches need two labeled datasets where we only need one labeled dataset, the second being unlabeled. We refer to the very interesting statistical study of *Conformal Predictors* in the binary case in the paper [Lei14].

Notation: First, we state general notation. Let $\mathcal{Y} = \{1, \dots, K\}$, with $K \geq 2$ being an integer. Let (X, Y) be the generic data-structure taking its values in $\mathcal{X} \times \mathcal{Y}$ with distribution $\mathbb{P}$. The goal in classification is to predict the label Y given an observation of X. This is performed based on a classifier (or classification rule) s which is a function mapping $\mathcal{X}$ onto $\mathcal{Y}$. 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

$$s^*(\cdot) = \operatorname*{argmax}_{k \in \mathcal{Y}} p_k(\cdot),$$

where $p_k(x) = \mathbb{P}(Y = k | X = x)$ for $x \in \mathcal{X}$ and $k \in \mathcal{Y}$.
Let us now consider more specific notation related to the multiclass confidence set setting. Let a confidence set be any measurable function that maps $\mathcal{X}$ onto $2^{\mathcal{Y}}$. Let Γ be a confidence set. This confidence set is characterized by two attributes. The first one is the risk associated to the confidence set

$$\mathcal{R}(\Gamma) = \mathbb{P}(Y \notin \Gamma(X)), \tag{1}$$

and is related to its accuracy. The second attribute is linked to the information given by the confidence set. It is defined as

$$\mathcal{I}(\Gamma) = \mathbb{E}(|\Gamma(X)|), \tag{2}$$

where $|\cdot|$ stands for the cardinality. Moreover, for some $\beta \in [1, K]$, we say that, for two confidence sets Γ and Γ' such that $\mathcal{I}(\Gamma) = \mathcal{I}(\Gamma') = \beta$, the confidence set Γ is "better" than Γ' if $\mathcal{R}(\Gamma) \leq \mathcal{R}(\Gamma')$.

Organization of the paper: The rest of the paper is organized as follows. Next section is devoted to the definition and the main properties of the oracle confidence set for multiclass classification. The empirical risk minimization procedure is provided in Section 3. Rates of convergence for the confidence set that results from this minimization can also be found in this section. We present an application of our procedure to aggregation of confidence sets 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.

2 Confidence set for multiclass classification

In the present section, we define a class of confidence sets that are suitable for multiclass classification and referred as *Oracle β-sets*. For some $\beta \in (0, K)$, these sets are shown to be optimal according to the risk (1) with an information (2) equal to β. Moreover, basic but fondamental properties of Oracle β-sets can be found in Proposition 1, while Proposition 3 provides another interpretation of these sets.

2.1 Notation and definition

First of all, we introduce in this section a class of confidence sets that specifies oracle confidence sets. Let $\beta \in (0, K)$ be a desired information level. The so-called *Oracle β-sets* are optimal according to the risk (1) among all the confidence sets Γ such that $\mathcal{I}(\Gamma) = \beta$. Throughout the paper we make the following assumption

(A1) *For all $k \in \{1, \dots, K\}$, the cumulative distribution function F_{p_k} of $p_k(X)$ is continuous.*

The definition of the *Oracle β-set* relies on the continuous and decreasing function G defined for any $t \in [0, 1]$ by

$$G(t) = \sum_{k=1}^{K} \bar{F}_{p_k}(t),$$

where for any $k \in \{1, \dots, K\}$, we denote by $\bar{F}_{p_k}$ the tail distribution function of $p_k(X)$, that is, $\bar{F}_{p_k} = 1 - F_{p_k}$ with F_{p_k} being the cumulative distribution function (c.d.f.) of $p_k(X)$. The generalized inverse G^{-1} of G is given by (see [vdV98]):

$$G^{-1}(\beta) = \inf\{t \in [0, 1] : \ G(t) \leq \beta\}, \ \ \forall \beta \in (0, K).$$

The functions G and G^{-1} are central in the construction of the Oracle β-sets. We then provide some of their useful properties in the following proposition.

Proposition 1. *The following properties on G hold*

i) *For every $t \in (0, 1)$ and $\beta \in (0, K)$, $G^{-1}(\beta) \leq t \Leftrightarrow \beta \geq G(t)$.*

ii) *For every $\beta \in (0, K)$, $G(G^{-1}(\beta)) = \beta$.*

iii) *Let ε be a random variable, independent of X, and distributed from a uniform distribution on $\{1, \dots, K\}$ and let U be uniformly distributed on $[0, K]$. Define*

$$Z = \sum_{k=1}^{K} p_k(X) \mathbf{1}_{\{\varepsilon = k\}}.$$

If the function G is continuous, then $G(Z) \overset{\mathcal{L}}{=} U$ and $G^{-1}(U) \overset{\mathcal{L}}{=} Z$.

The proof of Proposition 1 relies on Lemma 1 in the Appendix. Now, we are able to defined the Oracle β-set:

Definition 1. *Let $\beta \in (0, K)$, the Oracle β-set is given by*

$$\begin{aligned} \Gamma^*_\beta(X) &= \{k \in \{1, \dots, K\} : \ G(p_k(X)) \leq \beta\} \\ &= \left\{k \in \{1, \dots, K\} : \ p_k(X) \geq G^{-1}(\beta)\right\}. \end{aligned}$$

This definition of the Oracle β-set is intuitive and can be related to the binary classification with reject option setting [Cho70, HW06, DH15] in the following way: a label k is assigned to the Oracle β-set if the probability $p_k(X)$ is large enough. It is worth noting that the function G plays the same role as the c.d.f. of the score function in [DH15]. As emphasized by Proposition 1, their introduction allows to control exactly the information (2). Indeed, it follows from the definition of the Oracle β-set that for each $\beta \in (0, K)$

$$|\Gamma_\beta^*(X)| = \sum_{k=1}^{K} \mathbf{1}_{\{p_k(X) \geq G^{-1}(\beta)\}},$$

and then $\mathcal{I}(\Gamma_\beta^*) = \mathbb{E}\left[|\Gamma_\beta^*(X)|\right] = G(G^{-1}(\beta))$. Therefore, Proposition 1 ensures that

$$\mathcal{I}(\Gamma_\beta^*) = \beta.$$

This last display points out that the Oracle β-sets are indeed β-level (that is, its information equals β). In the next section, we focus on the study of the risk of these oracle confidence sets.

Remark 1. *Naturally, the definition of Oracle β-sets can be extended to any $\beta \in [0, K]$. However, the limit cases $\beta = 0$ and $\beta = K$ are of limited interest and are completely trivial. We then exclude these two limit cases from the present study.*

2.2 Properties of the oracle confidence sets

Let us first state the optimality of the Oracle β-set:

Proposition 2. *Let Assumption (A1) be satisfied. For any $\beta \in (0, K)$, we have both:*

1. *The Oracle β-set Γ_β^* satisfies the following property:*

$$\mathcal{R}\left(\Gamma_\beta^*\right) = \inf_{\Gamma} \mathcal{R}\left(\Gamma\right),$$

   where the infimum is taken over all β-level confidence sets.

2. *For any β-level confidence set Γ, the following holds*

$$0 \leq \mathcal{R}\left(\Gamma\right) - \mathcal{R}\left(\Gamma_\beta^*\right) = \mathbb{E}\left[\sum_{k \in (\Gamma_\beta^*(X)\, \Delta\, \Gamma(X))} \left|p_k(X) - G^{-1}(\beta)\right|\right], \tag{3}$$

   where the symbol Δ stands for the symmetric difference of two sets, that is, for two subsets A and B of $\{1, \ldots, K\}$, we write $A \,\Delta\, B = (A \setminus B) \cup (B \setminus A)$.

Several remarks can be made from Proposition 2. First, for $\beta \in (0, K)$, the Oracle β-set is optimal for the misclassification risk, over the class of β-level confidence sets. Moreover, the excess risk of any β-level confidence set relies on the behavior of the score functions p_k around the threshold $G^{-1}(\beta)$. Finally, we can note that if $K = 2$ and $\beta = 1$, which implies that $G^{-1}(\beta) = 1/2$, Equation (3) reduces to the misclassification excess risk in binary classification.

Remark 2. *One way to build a confidence set Γ with information β is to set Γ as the β top levels conditional probabilities. In the sequel this method is referred as the `max` procedure. This strategy is natural but actually suboptimal as shown by the first point of Proposition 2. As an illustration, we consider a simulation scheme with $K = 10$ classes. We generate (X, Y) according to a mixture model. More precisely,*

i) the label Y is distributed from a uniform distribution on $\{1, \dots, K\}$;

ii) conditional on $Y = k$, the feature X is generated according to a multivariate gaussian distribution with mean parameter $\mu_k \in \mathbb{R}^{10}$ and identity covariance matrix. For each $k = 1, \dots, K$, the vectors μ_k are i.i.d realizations of uniform distribution on $[0, 4]$.

For $\beta = 2$ we evaluate the risks of the Oracle β-set and the `max` *procedure and obtain respectively* 0.05 *and* 0.09 *(with very small variance).*

Remark 3. *An important motivation behind the introduction of confidence sets and in particular of Oracle β-sets is that they might outperform the Bayes rule which can be seen as the Oracle β-set associated to $\beta = 1$. This gap in performance is even larger when the number of classes K is large and there is a big confusion between classes. Such improvement will be illustrated in the numerical study (see Section 4.3).*

We end up this section by providing another characterization of the Oracle β-set.

Proposition 3. *For $t \in [0, 1]$, and Γ a confidence set, we define*

$$L_t(\Gamma) = \mathbb{P}\left(Y \notin \Gamma(X)\right) + t\mathcal{I}(\Gamma).$$

For $\beta \in [0, K]$, the following equality holds:

$$L_{G^{-1}(\beta)}(\Gamma^*_\beta) = \min_{\Gamma} L_{G^{-1}(\beta)}(\Gamma).$$

The proof of this proposition relies on the same arguments as those of Proposition 2. It is then omitted. Proposition 3 states that the Oracle β-set is defined as the minimizer, over all confidence sets Γ, of the risk function L_t when the tuning parameter t is set equal to $G^{-1}(\beta)$. Note moreover that the risk function L_t is a trade-off, controlled by the parameter t, between the risk of a confidence set on the one hand, and the information provided by this confidence set on the other hand. Hence, the risk function L_t can be viewed as a generalization to the multiclass case of the risk function provided in [Cho70, HW06] for binary classification with reject option setting.

3 Empirical risk minimization

In this section we introduce and study confidence sets which rely on the minimization of convex risks. Their definitions and main properties are given in Sections 3.1-3.2. As a consequence, we deduce a data-driven procedure described in Section 3.3 with several theoretical properties, such as rates of convergence, that we demonstrate in Section 3.4.

3.1 ϕ-risk

Let $f = (f_1, \dots, f_K) : \mathcal{X} \to \mathbb{R}^K$ be a score function and $G_f(.) = \sum_{k=1}^K \bar{F}_{f_k}(.)$. Assuming that the function G_f is continuous and given an information level $\beta \in (0, K)$, there exists $\delta \in \mathbb{R}$, such that $G_f(-\delta) = \beta$. Given this simple but important fact, we define the confidence set $\Gamma_{f,\delta}$ associated to f and δ as

$$\Gamma_{f,\delta}(X) = \{k \ : \ f_k(X) \geq -\delta\}. \tag{4}$$

In this way, the confidence set $\Gamma_{f,\delta}$ consists of top scores, and the threshold δ is fixed so that $\mathcal{I}\left(\Gamma_{f,\delta}\right) = \beta$. As a consequence, we naturally aim at solving the problem

$$\min_{f \in \mathcal{F}} \mathcal{R}(\Gamma_{f,\delta}),$$

where $\mathcal{F}$ is a class of functions. Due to computational issues, it is convenient to focus on a convex surrogate of the previous minimization problem. To this end, let $\phi : \mathbb{R} \to \mathbb{R}$ be a convex function. We define the ϕ-risk of f by

$$R_\phi(f) = \mathbb{E}\left[\sum_{k=1}^{K} \phi(Z_k f_k(X))\right], \tag{5}$$

where $Z_k = 2\,\mathbf{1}_{\{Y=k\}} - 1$ for all $k = 1, \ldots, K$. Therefore, our target score becomes

$$\bar{f} \in \operatorname*{argmin}_{f \in \mathcal{F}} R_\phi(f),$$

for the purpose of building the confidence $\Gamma_{\bar{f},\delta}$. In the sequel, we also introduce f^*, the minimizer over the class of all measurable functions, of the ϕ-risk. The notation suppresses the dependence on ϕ. It is worth mentioning at this point that the definition of the risk function R_ϕ is dictated by Equation (3) and suits for confidence sets. Moreover, this function differs from the classical risk function used in the multiclass setting (see [TB07]). The reason behind this is that building a confidence set is closer to K binary classification problems.

3.2 Classification calibration for confidence sets

Convexification of the risk in classification is a standard technique. In this section we adapt classical results and tools to confidence sets in the multiclass setting. We refer the reader to earlier papers as [Zha04, BJM06, YW10] for interesting developments.

One of the main important concept when we deal with convexification of the risk is the notion of calibration of the loss function ϕ. This property permits to connect confidence sets deduced from the convex risk and from the classification risk.

Definition 2. *We say that the function ϕ is confidence set calibrated if for all $\beta > 0$, there exists $\delta^* \in \mathbb{R}$ such that*

$$\Gamma_{f^*,\delta^*} = \Gamma^*_\beta,$$

with f^* being the minimizer of the ϕ-risk

$$f^* \in \operatorname*{argmin}_{f} R_\phi(f),$$

where the infimum is taken over the class of all measurable functions. Hence, the confidence set based on f^* coincides with the Bayes confidence set.

Given this, we can state the following proposition that gives a characterization of the confidence set calibration property in terms of the function G.

Proposition 4. *The function ϕ is confidence set calibrated if and only if for all $\beta \in (0, K)$, there exists $\delta^* \in \mathbb{R}$ such that $\phi'(\delta^*)$ and $\phi'(-\delta^*)$ both exists, $\phi'(\delta^*) < 0$, $\phi'(-\delta^*) < 0$ and*

$$G^{-1}(\beta) = \frac{\phi'(\delta^*)}{\phi'(\delta^*) + \phi'(-\delta^*)},$$

where ϕ' denotes the derivative of ϕ.

The proof of the proposition follows the lines of Theorem 1 in [YW10] with minor modifications. These characterizations of calibration for confidence sets generalize the notion of

calibration in the classification setting as well as the necessary and sufficient condition for the minimizer of the ϕ-risk to be calibrated. Indeed, if we pick $\delta = 0$ in Definition 2 and Proposition 4 we exactly come back to the calibration property in the classical classification setting (see [BJM06]). Note that commonly used loss functions as boosting ($x \mapsto \exp(-x)$), least squares ($x \mapsto (x-1)^2$) and logistic ($x \mapsto \log(1+\exp(-x))$) are examples of calibrated losses (see for instance [BJM06, WY11]). Now, for some score function f and some real number δ such that $G_f(-\delta) = \beta$, we define the excess risk

$$\Delta\mathcal{R}(\Gamma_{f,\delta}) = \mathcal{R}(\Gamma_{f,\delta}) - \mathcal{R}(\Gamma^*_\beta),$$

and the excess ϕ-risk

$$\Delta R_\phi(f) = R_\phi(f) - R_\phi(f^*).$$

We also introduce the marginal conditional excess ϕ-risk on $f = (f_1, \dots, f_K)$ as

$$\Delta R^k_\phi(f(X)) = p_k(X)(\phi(f_k(X)) - \phi(f^*_k(X))) + (1 - p_k(X))(\phi(-f_k(X)) - \phi(-f^*_k(X))),$$

for $k = 1, \dots, K$. The following theorem shows that the consistency in terms of ϕ-risk implies the consistency in terms of classification risk $\mathcal{R}$.

Theorem 1. *Assume that ϕ is confidence set calibrated and assume that there exists constants $C > 0$ and $s \geq 1$ such that*[1]

$$|p_k(X) - G^{-1}(\beta)|^s \leq C\Delta R^k_\phi(-\delta^*). \tag{6}$$

Let $\hat{f}_n$ be a sequence of scores. We assume that for each n, the function $G_{\hat{f}_n}$ is continuous. Let $\delta_n \in \mathbb{R}$ be such that $G_{\hat{f}_n}(-\delta_n) = \beta$, then

$$\Delta R_\phi(\hat{f}_n) \xrightarrow{P} 0 \qquad \Rightarrow \qquad \Delta\mathcal{R}(\Gamma_{\hat{f}_n,\delta_n}) \xrightarrow{P} 0.$$

The theorem ensures that the convergence in terms of ϕ risk implies the convergence in terms of risk $\mathcal{R}$. This convergence is made possible since we manage, in the proof, to bound the excess risk by (a power of) the excess ϕ-risk. The assumption needed in this theorem is also standard and is for instance satisfied for the boosting, least square and logistic losses with the parameter s being equal to 2 (see [BJM06]).

3.3 Data-driven procedure

In this section we provide the steps of the construction of our empirical confidence set that is deduced from the empirical risk minimization. Before going into further details, let us first mention that our procedure is semi-supervised in the sense that it requires two datasets, one of which being unlabeled. Hence we introduce a first data set $\mathcal{D}_n = \{(X_i, Y_i),\ i = 1, \dots, n\}$, which consists of independent copies of (X, Y). We define the empirical ϕ-risk associated to a score function f (which is the empirical counterpart of R_ϕ given in (5)):

$$\hat{R}_\phi(f) = \frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K}\phi(Z^i_k f_k(X_i)), \tag{7}$$

where $Z^i_k = 2\,\mathbf{1}_{\{Y_i=k\}} - 1$ for all $k = 1, \dots, K$. We also define the empirical risk minimizer over $\mathcal{F}$, a convex set of score functions, as

$$\hat{f} = \arg\min_{f\in\mathcal{F}} \hat{R}_\phi(f).$$

[1] With abuse of notation, we write $\Delta R^k_\phi(-\delta^*)$ instead of $\Delta R^k_\phi((-\delta^*, \dots, -\delta^*))$ since no confusion can occur.

At this point, we have in hands the optimal score function $\hat{f}$ and need to build the corresponding confidence set with the right expected size. However, let us before introduce an intermediate confidence set that would help comprehension since it mimics the oracle β-set Γ^* with regard to its construction using $\hat{f}$ instead of f^*. For this purpose, we define

$$F_{\hat{f}_k}(t) = \mathbb{P}_X\left(\hat{f}_k(X) \leq t | \mathcal{D}_n\right),$$

for $t \in \mathbb{R}$, where $\mathbb{P}_X$ is the marginal distribution of X. As for the c.d.f. of p_k and f_k, we make the following assumption:

(A2) *The cumulative distribution functions $F_{\hat{f}_k}$ with $k = 1, \ldots, K$ are continuous.*

At this point, we are able to define an empirical approximation of the Oracle β-set for $\beta \in (0, K)$:

$$\widetilde{\Gamma}_\beta(X) = \left\{k \in \{1, \ldots, K\} : \ \widetilde{G}(\hat{f}_k(X)) \leq \beta\right\}, \tag{8}$$

where for $t \in \mathbb{R}$

$$\widetilde{G}(t) = \sum_{k=1}^{K} \bar{F}_{\hat{f}_k}(t), \tag{9}$$

with $\bar{F}_{\hat{f}_k} = 1 - F_{\hat{f}_k}$. Since the function $\widetilde{G}$ depends on the unknown distribution of X, we consider a second but interestingly *unlabeled* dataset $\mathcal{D}_N = \{X_i, i = 1, \ldots, N\}$, independent of $\mathcal{D}_n$ in order to compute the empirical versions of the $\bar{F}_{\hat{f}_k}$'s. By now, we can define the empirical β-set based on $\hat{f}$:

Definition 3. *Let $\hat{f}$ be the minimizer of the empirical ϕ-risk given in* (7) *based on $\mathcal{D}_n$, and consider the unlabeled dataset $\mathcal{D}_N$. Let $\beta \in (0, K)$. The empirical β-set is given by*

$$\hat{\Gamma}_\beta(X) = \left\{k \in \{1, \ldots, K\} : \ \hat{G}(\hat{f}_k(X)) \leq \beta\right\}, \tag{10}$$

where

$$\hat{G}(.) = \frac{1}{N} \sum_{i=1}^{N} \sum_{k=1}^{K} \mathbf{1}_{\{\hat{f}_k(X_i) \geq .\}}.$$

The most important remark about the construction of this data-driven confidence set is that it is made in a semi-supervised way. Indeed, the estimation of the tail distribution functions of $\hat{f}_k$ requires only a set of *unlabeled* observations. This is particularly attractive in applications where the number of label observations is small (because labelling examples is time consuming or may be costly) and where one has in hand several unlabeled features that can be used to make prediction more accurate. As an important consequence is that the estimation of the tail distribution functions of $\hat{f}_k$ does not depend on the number of observations in each class label. That is to say, this unlabeled dataset can be unbalanced with respect to the classes, that can often occur in multiclass classification settings where the number of classes K is quite large. Next section deals with the theoretical performance of the empirical β-sets.

3.4 Rates of convergence

In this Section, we provide rates of convergence for the empirical confidence sets defined in Section 3.3. First, we state some additional notation. In the sequel, the symbols **P** and **E**

stand for generic probability and expectation, respectively. Moreover, given the empirical β-set $\hat{\Gamma}_\beta$ from Definition 3, we consider the risk $\mathbf{R}\left(\hat{\Gamma}_\beta\right) = \mathbf{P}\left(Y \notin \hat{\Gamma}_\beta(X)\right)$ and the information $\mathbf{I}\left(\hat{\Gamma}_\beta\right) = \mathbf{E}\left[|\hat{\Gamma}_\beta(X)|\right]$. Our first result ensures that $\hat{\Gamma}_\beta$ is of level β up to a term of order $K/\sqrt{N}$.

Proposition 5. *For each $\beta \in (0, K)$, the following equality holds*

$$\mathbf{I}\left(\hat{\Gamma}_\beta\right) = \beta + O\left(\frac{K}{\sqrt{N}}\right).$$

In order to precise rates of convergence for the risk, we need to formulate several assumptions. First, we consider loss functions ϕ which have in common that the modulus of convexity of their underlying risk function R_ϕ, defined by

$$\delta(\varepsilon) = \inf\left\{\frac{R_\phi(f) + R_\phi(g)}{2} - R_\phi\left(\frac{f+g}{2}\right),\ \sum_{k=1}^{K} \mathbb{E}_X\left[(f_k - g_k)^2(X)\right] \geq \varepsilon^2\right\}, \tag{11}$$

satisfies $\delta(\varepsilon) \geq c_1\epsilon^2$ for some $c_1 > 0$ (we refer to [BJM06, BM06] for more details on modulus of convexity for classification risk). Moreover, we assume that ϕ is classification calibrated and L-Lipschitz for $L > 0$. On the other hand, for $\alpha > 0$, we assume a margin condition $\mathrm{M}_\alpha^{\mathrm{k}}$ on each $p_k, k = 1, \ldots, K$.

$$\mathrm{M}_\alpha^{\mathrm{k}}: \quad \mathbb{P}_X\left(0 < |p_k(X) - G^{-1}(\beta)| \leq t\right) \leq c_2 t^\alpha, \quad \text{for some constant } c_2 > 0 \text{ and for all } t > 0.$$

It is important to note that since we assume that the distribution functions of $p_k(X)$ are continuous for each k, we have $\mathbb{P}_X\left(0 < |p_k(X) - G^{-1}(\beta)| \leq t\right) \to 0$ with $t \to 0$. Therefore, the margin condition only precise the rate of this decay to 0. Now, we provide the rate of convergence that we can expect for the empirical confidence sets defined by (10).

Theorem 2. *Assume that $\|f\|_\infty \leq B$ for all $f \in \mathcal{F}$. Let $M_n = \mathcal{N}(1/n, L_\infty, \mathcal{F})$ be the cardinality of the set of closed balls with radius $1/n$ in L_∞-norm needed to cover $\mathcal{F}$. Under the assumptions of Theorem 1, with the modulos of convexity $\delta(\varepsilon) \geq c_1\epsilon^2$ with $c_1 > 0$, if ϕ is L-Lipschitz and if the margin assumptions $\mathrm{M}_\alpha^{\mathrm{k}}$ are satisfied with $\alpha > 0$, then*

$$\mathbf{E}\left[|\Delta\mathcal{R}(\hat{\Gamma}_\beta)|\right] \leq C(B, L, s, \alpha) K^{1-\alpha/(\alpha+s)} \left\{\inf_{f\in\mathcal{F}} \Delta R_\phi(f) + \frac{K\log(M_n)}{n}\right\}^{\alpha/(\alpha+s)} + C'\frac{K}{\sqrt{N}},$$

where $C' > 0$ is an absolute constant and $C(B, L, s, \alpha) > 0$ is a constant that depends only on L, B, s and α.

From this theorem we obtain the following rate of convergence: $K\left(\frac{\log(M_n)}{n}\right)^{\alpha/(\alpha+s)} + \frac{K}{\sqrt{N}}$. This is the first bound for confidence sets in multiclass setting. The regularity parameter α plays a crucial role and governs the rate; larger values of α lead to faster rates. Moreover, this rate is linear in K, the number of classes, that seems to be the characteristic of multiclass classification as compared to binary classification. Note that the exponent $\alpha/(\alpha + s)$ is not classical and is due to the estimation of quantiles. The second part of the rates which is of order $K/\sqrt{N}$ relies on the estimation of the function $\tilde{G}$ given in (9). This part of the estimation is established under the mild Assumptions (A1) and (A2). For this term the proof of the linearity on K is tricky and is obtained thanks to a new technical probabilistic results on sums of cumulative distribution functions (see Lemma 1). Let us conclude this paragraph by mentioning that Theorem 2 applies for instance for $\mathcal{F}$ being the convex hull of a finite family of a score functions which is the scope of the next section.

4 Application to confidence sets aggregation

This Section is devoted to an aggregation procedure which relies on the superlearning principle. The specifics of our superlearning procedure is given in Section 4.1. In Section 4.2, we show the consistency of our procedure and finally we provide a numerical illustration in Section 4.3 of our algorithm.

4.1 Description of the procedure: superlearning

In this section, we describe the superlearning procedure for confidence sets. The superlearning procedure is based on the V-fold cross-validation procedure (see [vdLPH07]). Initially, this aggregation procedure has been introduced in the context of regression and binary classification. Let $V \geq 2$ be an integer and let $(B_v)_{1\leq v\leq V}$ be a regular partition of $\{1,\dots,n\}$, *i.e.*, a partition such that, for each $v = 1,\dots,V$, $\mathrm{card}(B_v) \in \{\lfloor n/V \rfloor, \lfloor n/V \rfloor + 1\}$, where we write $\lfloor x \rfloor$ for the floor of any real number x (that is, $x-1 \leq \lfloor x \rfloor \leq x$). For each $v \in \{1,\dots,V\}$, we denote by $D_n^{(v)}$ (respectively $D_n^{(-v)}$) the dataset $\{(X_i,Y_i), i \in B_v\}$ (respectively $\{(X_i,Y_i), i \notin B_v\}$), and define the corresponding empirical measures

$$\begin{aligned} P_n^{(v)} &= \frac{1}{\mathrm{card}(B_v)} \sum_{i\in B_v} \mathrm{Dirac}(X_i,Y_i), \quad \text{and} \\ P_n^{(-v)} &= \frac{1}{n-\mathrm{card}(B_v)} \sum_{i\notin B_v} \mathrm{Dirac}(X_i,Y_i). \end{aligned}$$

For a score algorithm f, that is to say a function which maps the empirical distribution to a score function. We define the cross-validated risk of f as

$$\widehat{R}_\phi^n(f) = \frac{1}{V}\sum_{v=1}^{V} R_\phi^{(P_n^{(v)})}\left(f(P_n^{(-v)})\right), \tag{12}$$

where for each $v \in \{1,\dots,V\}$, $R_\phi^{(P_n^{(v)})}(f(P_n^{(-v)}))$ is the empirical estimator of $R_\phi(f(P_n^{(-v)}))$, based on $D_n^{(v)}$ and conditionally on $D_n^{(-v)}$:

$$R_\phi^{(P_n^{(v)})}\left(f(P_n^{(-v)})\right) = \frac{1}{\mathrm{card}(B_v)} \sum_{i\in I_v}\sum_{k=1}^{K} \phi(Z_k^i f_k(P_n^{(-v)})(X_i)).$$

Next, we consider $\mathcal{F} = \left(f^1,\dots,f^M\right)$ a family of M score algorithms. We define the cross-validated score by

$$\widehat{f} \in \underset{f\in \mathrm{conv}(\mathcal{F})}{\mathrm{argmin}}\ \widehat{R}_\phi^n(f). \tag{13}$$

Finally, we consider the resulting cross-validated confidence set defined by

$$\hat{\Gamma}_\beta^{\mathrm{CV}}(X) = \left\{k \in \{1,\dots,K\} :\ \hat{G}(\hat{f}_k(X)) \leq \beta\right\},$$

that we analyse in the next section.

4.2 Consistency of the algorithm

In this Section, we show some results that illustrate the consistency of the superlearning procedure described above. To this end, we introduce the oracle counterpart of the cross-validated risk defined in (12). For a score f, we define

$$\tilde{R}_\phi^n(f) = \frac{1}{V}\sum_{v=1}^{V} R_\phi(f(P_n^{(-v)}),$$

that yields to the oracle counterpart of the cross-validated score defined in (13)

$$\tilde{f} \in \underset{f\in \mathrm{conv}(\mathcal{F})}{\mathrm{argmin}}\ \tilde{R}_\phi^n(f).$$

Here again, we assume that the loss function ϕ satisfies the properties described in Section 3.4 so that we can state the following result:

Proposition 6. *We assume that for each $f \in \mathrm{conv}(\mathcal{F})$ and $v \in \{1,\dots,V\}$, $\|f(P_n^{(-v)})\|_\infty \leq B$. Then the following holds*

$$\mathbf{E}\left[\tilde{R}_\phi^n(\hat{f}) - \tilde{R}_\phi^n(\tilde{f})\right] \leq C(B,L)\frac{KM\log(n)}{\lfloor n/V\rfloor},$$

where $C(B,L) > 0$ is a constant that depends only on B and L.

As usual when one deals with cross-validated estimators, the theorem compares $\hat{f}$ to the oracle counterpart $\tilde{f}$ in terms of the oracle cross-validated ϕ-risk. The theorem teaches us that, for sufficiently large n, $\hat{f}$ perform as well as $\tilde{f}$. However our main goal remains confidence sets. Therefore the next step consists in showing that the confidence set associated to the cross-validated score $\hat{f}$ has good properties in terms of the classification risk. Let $\Gamma_{f,\delta}$ be the confidence set that results from a choice of $\beta \in (0,K)$ and a score function $f \in \mathrm{conv}(\mathcal{F})$. We introduce the following excess risks

$$\begin{aligned}
\Delta\tilde{\mathcal{R}}^n(\Gamma_{f,\delta}) &= \frac{1}{V}\sum_{v=1}^{V}\mathcal{R}\left(\Gamma_{f\left(P_n^{(-v)}\right),\delta}\right) - \mathcal{R}^*,\\
\Delta\tilde{R}_\phi^n(f) &= \frac{1}{V}\sum_{v=1}^{V} R_\phi(f(P_n^{(-v)})) - R_\phi(f^*).
\end{aligned}$$

These quantities can be view as cross-validated counterparts of the excess risks introduced in Section 3.2. Hereafter, we provide a result which can be view as the cross-validation counterpart of Theorem 2. It is interpreted in a same way.

Proposition 7. *We assume that for each $v \in \{1,\dots,V\}$, $t \mapsto \mathbb{P}_X\left(\hat{f}\left(P_n^{(-v)}\right) \leq t|\mathcal{D}_n\right)$ is continuous. Under the assumptions of Proposition 6 and if the margin assumptions $\mathrm{M}_\alpha^{\mathrm{k}}$ are satisfied with $\alpha > 0$,*

$$\mathbf{E}\left[|\Delta\tilde{\mathcal{R}}^n(\hat{\Gamma}_\beta^{\mathrm{CV}})|\right] \leq C(B,L,\alpha,s)K^{1-\alpha/(\alpha+s)}\left\{\mathbf{E}\left[\Delta\tilde{R}_\phi^n(\tilde{f})\right] + \frac{KM\log(n)}{\lfloor n/V\rfloor}\right\}^{\alpha/(\alpha+s)} + C'\frac{K}{\sqrt{N}}.$$

The proof of Proposition 7 relies on Proposition 6 and similar arguments as in Theorem 2.

`Forest` ($K = 4$)

		β-set				
β		`rforest`	`softmax reg`	`svm`	`kknn`	`CV`
2	**R**	0.02 (0.02)	0.06 (0.02)	0.02 (0.01)	0.05 (0.03)	0.02 (0.01)
	I	2.00 (0.09)	2.00 (0.08)	2.00 (0.09)	2.00 (0.08)	2.00 (0.08)

`Plant` ($K = 100$)

		β-set				
β		`rforest`	`softmax reg`	`svm`	`kknn`	`CV`
2	**R**	0.18 (0.03)	0.77 (0.02)	0.32 (0.04)	0.20 (0.03)	0.17 (0.03)
	I	2.00 (0.09)	2.02 (0.18)	1.99 (0.10)	2.00 (0.08)	2.00 (0.08)
10	**R**	0.02 (0.01)	0.42 (0.04)	0.03 (0.02)	0.08 (0.03)	0.02 (0.01)
	I	9.95 (0.38)	10.06 (0.58)	9.98 (0.22)	9.98 (0.23)	9.96 (0.37)

Table 1: For each of the $B = 100$ repetitions and for each dataset, we derive the estimated risks **R** and information **I** of the different β-sets w.r.t. β. We compute the means and standard deviations (between parentheses) over the $B = 100$ repetitions. For each β, the β-sets are based on, from left to right, `rforest`, `softmax reg` and `svm`, `kknn` and `CV` which are respectively the random forest, the softmax regression, support vector machines, k nearest neighbors and the superlearning procedure. Top: the dataset is the `Forest` – the dataset is the `Plant`.

4.3 Application to real datasets

In this section, we provide an application of our aggregation procedure described in Section 4.1. For the numerical experiment we focus on the boosting loss and consider the library of algorithms constituted by the random forest, the softmax regression, the support vector machines and the k nearest neighbors (with $k = 11$) procedures. To be more specifics, we respectively exploit the `R` packages `randomForest`, `polspline`, `e1071` and `kknn`. All the `R` functions are used with standard tuning parameters. Finally the parameter V of the aggregation procedure is fixed to 5.

We evaluate the performance of the procedure on two real datasets: the *Forest type mapping* dataset and the *one-hundred plant species leaves* dataset coming from the UCI database. In the sequel we refer to these two datasets as `Forest` and `Plant` respectively. The `Forest` dataset consists of $K = 4$ classes and 523 labeled observations (we gather the train and test sets) with 27 features. Here the classes are unbalanced. In the `Plant` dataset, there are $K = 100$ classes and 1600 labeled observations. This dataset is balanced so that each class consists of 16 observations. The original dataset contains 3 covariates (each covariate consists of 64 features). In order to make the problem more challenging, we drop 2 covariates.

To get an indication of the statistical significance, it makes sense to compare our aggregated confidence set (referred as `CV`) to the confidence sets that result from each component of the library. Roughly speaking, we evaluate risks (and informations) of these confidence sets on each dataset. To do so, we use the *cross validation* principle. In particular, we run $B = 100$ times the procedure where we split the data each time in three: a sample of size n to build the scores $\hat{f}$; a sample of size N to estimate the function G and to get the confidence sets; and a sample of size M to evaluate the risk and the information. For both datasets, we make sure that in the sample of size n, there is the same number of observations in each class. As a benchmark, we notify that the misclassification risks of the best classifier from the library in the `Forest` dataset is evaluated at 0.15 , whereas in the `Plant` dataset, it is evaluated at 0.40. As planned, the performance of the classical methods are quite bad in this last dataset.

We set the sizes of the samples as $n = 200$, $N = 100$ and $M = 223$ for the `Forest` dataset,

and $n = 1000$, $N = 200$ and $M = 400$ for the `Plant` one. The results are reported in Table 1, and confirm our expectations. In particular, our main observation is that the aggregated confidence set (`CV`) outperforms all components of the library in the sense that it is at least as good as the best component in all of the experiments. Second, let us state some remarks that hold for all of the confidence sets and in particular our aggregated confidence set. First, we note that the information $\mathbf{I}(\Gamma)$ has the good level β which is supported by our theory. Moreover, we see that the risk gets drastically better with moderate β as compared to the *best* misclassification risk. For instance, for the `Plant`, the error rate of the confidence set with $\beta = 2$ based on random forests is 0.18 whereas the misclassification error rate of the best component is 0.40.

5 Conclusion

In multiclass classification setting, the present paper propose a new procedure that assigns a set of labels instead of a single label to each instance. This set has a controlled expected size (or information) and its construction relies on cumulative distribution functions and on an empirical risk minimization procedure. Theoretical guarantees, especially rates of convergence are also provided and rely on the regularity of these cumulative distribution functions. The obtained rates of convergence highlight a linear dependence w.r.t the number of classes K. The procedure described in Section 3 is defined as a two steps algorithm whose second step consists in the estimation of the function G (which is a sum of tail distribution functions). Interestingly, this step does not require a set of labeled data and neither to explore the whole classes, that is suitable for semi-supervised learning. Moreover, we apply our methodology to derive an aggregation algorithm which is based on the V-fold cross-validation principle. Future works will focus on the optimality in the minimax sense with respect to the classification error. In particular, we will investigate whether the rates of convergence are optimal in terms of their dependence on K. We believe that this dependence (linearity on K) is the correct one. However we will instigate whether this dependence can be reduced under some sparsity assumption.

6 Appendix

This section gathers the proofs of our results. Let us first add a notation that will be used throughout the Appendix: for any random variable (or vector) Z, we denote by $\mathbb{P}_Z$ the probability w.r.t. Z and by $\mathbb{E}_Z$, the corresponding expectation.

6.1 Technical Lemmas

We first lay out key lemmata, which are crucial to establish the main theory. We consider $K \geq 2$ be an integer, and $Z_1, \ldots, Z_K$, K random variables. Moreover we define function H by:

$$H(t) = \frac{1}{K} \sum_{k=1}^{K} F_k(t), \quad \forall t \in [0, 1],$$

where for all $k = 1, \ldots, K$, F_k is the cumulative distribution function of Z_k. Finally, let us define the generalized inverse H^{-1} of H:

$$H^{-1}(p) = \inf\{t : H(t) \geq p\}, \quad \forall p \in (0, 1).$$

Lemma 1. *Let ε distributed from a uniform distribution on $\{1,\dots,K\}$ and independent of Z_k, $k=1,\dots,K$. Let U distributed from a uniform distribution on $[0,1]$. We consider*

$$Z=\sum_{k=1}^{K} Z_k \mathbf{1}_{\{\varepsilon=k\}}.$$

If H is continuous then

$$H(Z)\stackrel{\mathcal{L}}{=}U \text{ and } H^{-1}(U)\stackrel{\mathcal{L}}{=}Z$$

Proof. First we note that for every $t\in[0,1]$, $\mathbb{P}\left(H(Z)\leq t\right)=\mathbb{P}\left(Z\leq H^{-1}(t)\right)$. Moreover, we have

$$\begin{aligned}\mathbb{P}\left(H(Z)\leq t\right) &= \sum_{k=1}^{K}\mathbb{P}(Z\leq H^{-1}(t),\varepsilon=k)\\ &= \frac{1}{K}\sum_{k=1}^{K}\mathbb{P}(Z_k\leq H^{-1}(t)) \text{ with } \varepsilon \text{ independent of } X\\ &= H(H^{-1}(t))\\ &= t \text{ with } H \text{ continuous.}\end{aligned}$$

To conclude the proof, we observe that

$$\begin{aligned}\mathbb{P}\left(H^{-1}(U)\leq t\right) &= \mathbb{P}\left(U\leq H(t)\right)\\ &= \frac{1}{K}\sum_{k=1}^{K}F_k(t)\\ &= \sum_{k=1}^{K}\mathbb{P}\left(Z_k\leq t,\varepsilon=k\right)\\ &= \mathbb{P}(Z\leq t).\end{aligned}$$

□

Lemma 2. *There exists an absolute constant $C'>0$ such that*

$$\sum_{k=1}^{K}\mathbf{P}\left(|\hat{G}(\hat{f}_k(X)-\widetilde{G}(\hat{f}_k(X))|\geq|\widetilde{G}(\hat{f}_k(X))-\beta|\right)\leq\frac{C'K}{\sqrt{N}}.$$

Proof. We define, for $\gamma>0$ and $k\in\{1,\dots K\}$

$$\begin{aligned}A_0^k &= \left\{\left|\widetilde{G}(\hat{f}_k(X))-\beta\right|\leq\gamma\right\}\\ A_j^k &= \left\{2^{j-1}\gamma<|\widetilde{G}(\hat{f}_k(X))-\beta|\leq 2^j\gamma\right\},\ j\geq 1.\end{aligned}$$

Since, for every k, the events $(A_j^k)_{j\geq 0}$ are mutually exclusive, we deduce

$$\sum_{k=1}^{K}\mathbf{P}\left(|\hat{G}(\hat{f}_k(X)-\widetilde{G}(\hat{f}_k(X))|\geq|\widetilde{G}(\hat{f}_k(X))-\beta|\right)=$$

$$\sum_{k=1}^{K}\sum_{j\geq 0}\mathbf{P}\left(|\hat{G}(\hat{f}_k(X)-\widetilde{G}(\hat{f}_k(X))|\geq|\widetilde{G}(\hat{f}_k(X))-\beta|,A_j^k\right). \quad (14)$$

Now, we consider a random variable ε uniformly distributed on $\{1,\dots,K\}$ and independent of $\mathcal{D}_n$ and X. Conditional on $\mathcal{D}_n$ and under Assumption $(A2)$, we apply Lemma 1 with $Z_k = \hat{f}_k(X)$, $Z = \sum_{k=1}^K Z_k \mathbf{1}_{\{\varepsilon=k\}}$ and then obtain that $\widetilde{G}(Z)$ is uniformly distributed on $[0,K]$. Therefore, for all $j \geq 0$ and $\gamma > 0$, we deduce

$$\begin{aligned}\frac{1}{K}\sum_{k=1}^K \mathbb{P}_X\left(|\widetilde{G}(\hat{f}_k(X)) - \beta| \leq 2^j\gamma|\mathcal{D}_n\right) &= \mathbb{P}_X\left(|\widetilde{G}(Z) - \beta| \leq 2^j\gamma|\mathcal{D}_n\right)\\ &\leq \frac{2^{j+1}\gamma}{K}.\end{aligned}$$

Hence, for all $j \geq 0$, we obtain

$$\sum_{k=1}^K \mathbf{P}(A_j^k) \leq 2^{j+1}\gamma. \tag{15}$$

Next, we observe that for all $j \geq 1$

$$\begin{aligned}&\sum_{k=1}^K \mathbf{P}\left(|\hat{G}(\hat{f}_k(X) - \widetilde{G}(\hat{f}_k(X))| \geq |\widetilde{G}(\hat{f}_k(X)) - \beta|, A_j^k\right) \leq\\ &\qquad\sum_{k=1}^K \mathbb{E}_{(\mathcal{D}_n,X)}\left[\mathbb{P}_{\mathcal{D}_N}\left(|\hat{G}(\hat{f}_k(X)) - \widetilde{G}(\hat{f}_k(X))| \geq 2^{j-1}\gamma|\mathcal{D}_n, X\right)\mathbf{1}_{A_j^k}\right].\end{aligned} \tag{16}$$

Now, since conditional on $(\mathcal{D}_n, X)$, $\hat{G}(\hat{f}_k(X))$ is an empirical mean of i.i.d random variables of common mean $\widetilde{G}(\hat{f}_k(X)) \in [0,K]$, we deduce from Hoeffding's inequality that

$$\mathbb{P}_{\mathcal{D}_N}\left(|\hat{G}(\hat{f}_k(X)) - \widetilde{G}(\hat{f}_k(X))| \geq 2^{j-1}\gamma|\mathcal{D}_n, X\right) \leq 2\exp\left(-\frac{N\gamma^2 2^{2j-1}}{K^2}\right).$$

Therefore, from Inequalities (14), (15) and (16), we get

$$\begin{aligned}&\sum_{k=1}^K \mathbf{P}\left(|\hat{G}(\hat{f}_k(X) - \widetilde{G}(\hat{f}_k(X))| \geq |\widetilde{G}(\hat{f}_k(X)) - \beta|\right)\\ &\qquad\leq \sum_{k=1}^K \mathbf{P}\left(A_0^k\right) + \sum_{j\geq 1} 2\exp\left(-\frac{N\gamma^2 2^{2j-1}}{K^2}\right)\left(\sum_{k=1}^K \mathbf{P}\left(A_j^k\right)\right)\\ &\qquad\leq 2\gamma + \gamma\sum_{j\geq 1} 2^{j+2}\exp\left(-\frac{N\gamma^2 2^{2j-1}}{K^2}\right).\end{aligned}$$

Finally, choosing $\gamma = \dfrac{K}{\sqrt{N}}$ in the above inequality, we finish the proof of the lemma. □

6.2 Proof of Proposition 2

Let $\beta > 0$ and Γ be a confidence set such that $\mathcal{I}(\Gamma) = \beta$. First, we note that the following decomposition holds

$$\begin{aligned}
\mathcal{R}(\Gamma) - \mathcal{R}(\Gamma^*_\beta) &= \sum_{j=1}^{K}\sum_{k=1}^{K} \mathbb{E}\left[\sum_{l=1}^{K} (\mathbf{1}_{\{Y \notin \Gamma(X)\}} - \mathbf{1}_{\{Y \notin \Gamma^*_\beta(X)\}}) \mathbf{1}_{\{Y=l\}} \mathbf{1}_{\{|\Gamma(X)|=k\}} \mathbf{1}_{\{|\Gamma^*_\beta(X)|=j\}}\right] \\
&= \sum_{j=1}^{K}\sum_{k=1}^{K} \mathbb{E}\left[\sum_{l=1}^{K} (\mathbf{1}_{\{l \notin \Gamma(X)\}} - \mathbf{1}_{\{l \notin \Gamma^*_\beta(X)\}}) p_l(X) \mathbf{1}_{\{|\Gamma(X)|=k\}} \mathbf{1}_{\{|\Gamma^*_\beta(X)|=j\}}\right] \\
&= \sum_{j=1}^{K}\sum_{k=1}^{K} \mathbb{E}\left[\sum_{l=1}^{K} (\mathbf{1}_{\{l \in \Gamma^*_\beta(X) \setminus \Gamma(X)\}} - \mathbf{1}_{\{l \in \Gamma(X) \setminus \Gamma^*_\beta(X)\}}) p_l(X) \mathbf{1}_{\{|\Gamma(X)|=k, |\Gamma^*_\beta(X)|=j\}}\right],
\end{aligned}$$

where we conditioned by X to get the second equality. From the last decomposition and with

$$\begin{aligned}
\mathbb{E}\left[|\Gamma(X)|\right] &= \mathbb{E}\left[|\Gamma^*_\beta(X)|\right] \\
&= \mathbb{E}\left[\sum_{j=1}^{K}\sum_{k=1}^{K} k \mathbf{1}_{\{|\Gamma(X)|=k\}} \mathbf{1}_{\{|\Gamma^*_\beta(X)|=j\}}\right] \\
&= \mathbb{E}\left[\sum_{j=1}^{K}\sum_{k=1}^{K} j \mathbf{1}_{\{|\Gamma(X)|=k\}} \mathbf{1}_{\{|\Gamma^*_\beta(X)|=j\}}\right],
\end{aligned}$$

we can express the excess risk as the sum of two terms:

$$\begin{aligned}
\mathcal{R}(\Gamma) - \mathcal{R}(\Gamma^*_\beta) &= \sum_{j=1}^{K}\sum_{k=1}^{K} \mathbb{E}\left[\left(\sum_{l=1}^{K} \mathbf{1}_{\{l \in \Gamma^*_\beta(X) \setminus \Gamma(X)\}} p_l(X) - jG^{-1}(\beta)\right) \mathbf{1}_{\{|\Gamma(X)|=k\}} \mathbf{1}_{\{|\Gamma^*_\beta(X)|=j\}}\right] \\
&\quad + \sum_{j=1}^{K}\sum_{k=1}^{K} \mathbb{E}\left[\left(kG^{-1}(\beta) - \sum_{l=1}^{K} \mathbf{1}_{\{l \in \Gamma(X) \setminus \Gamma^*_\beta(X)\}} p_l(X)\right) \mathbf{1}_{\{|\Gamma(X)|=k\}} \mathbf{1}_{\{|\Gamma^*_\beta(X)|=j\}}\right]. \quad (17)
\end{aligned}$$

Now, for $j, k \in \{1, \ldots, K\}$ on the event $\{|\Gamma(X)| = k, |\Gamma^*_\beta(X)| = j\}$, we have

$$k = \sum_{l=1}^{K} \mathbf{1}_{\{l \in \Gamma(X) \setminus \Gamma^*_\beta(X)\}} + \sum_{l=1}^{K} \mathbf{1}_{\{l \in \Gamma(X) \cap \Gamma^*_\beta(X)\}},$$

and

$$j = \sum_{l=1}^{K} \mathbf{1}_{\{l \in \Gamma^*_\beta(X) \setminus \Gamma(X)\}} + \sum_{l=1}^{K} \mathbf{1}_{\{l \in \Gamma(X) \cap \Gamma^*_\beta(X)\}}.$$

Therefore, since

$$l \in \Gamma^*_\beta \Leftrightarrow p_l(X) \geq G^{-1}(\beta),$$

Equality (17) yields the result.

6.3 Proof of Theorem 1

First we recall that $\hat{f}_n = (\hat{f}_{n,1}, \dots, \hat{f}_{n,K})$ is a sequence of score functions and $\delta_n \in \mathbb{R}$ is such that $G_{\hat{f}_n}(-\delta_n) = \beta$. We suppress the dependence on n to simplify notation and write $\hat{f} = (\hat{f}_1, \dots, \hat{f}_K)$ and δ for $\hat{f}$ and δ_n respectively. Moreover, since there is no doubt, we also suppress everywhere the dependence on X. We also define the events

$$B_k = \{\hat{f}_k \in (-\delta, -\delta^*) \text{ or } \hat{f}_k \in (-\delta^*, -\delta)\}, \tag{18}$$

for $k = 1, \dots, K$. We aim at controlling the excess risk $\Delta\mathcal{R}(\Gamma_{\hat{f},\delta})$. Since the risk $\mathcal{R}$ is decomposable, it is convenient to introduce "marginal excess risks":

$$\Delta\mathcal{R}^k(\Gamma_{f,\delta}) = \mathbf{1}_{\{k \in (\Gamma_{\hat{f},\delta} \,\Delta\, \Gamma^*_\beta)\}} |p_k - G^{-1}(\beta)|.$$

Recall also that by convexity of the loss function ϕ, we have that for all $x, y \in \mathbb{R}$,

$$\phi(y) - \phi(x) \geq \phi'(x)(y - x). \tag{19}$$

Assume that $\hat{f}_k \leq -\delta$ and $\hat{f}_k \leq -\delta^* \leq f^*_k$, which translates as $p_k - G^{-1}(\beta) \geq 0$, we get thanks to (19)

$$p_k(\phi(\hat{f}_k) - \phi(-\delta)) - (1 - p_k)(\phi(-\hat{f}_k) - \phi(\delta)) \geq (\phi'(\delta^*) - \phi'(-\delta^*))(p_k - G^{-1}(\beta))(\hat{f}_k + \delta^*) \geq 0.$$

Similarly, if $\hat{f}_k \geq -\delta$ and $\hat{f}_k \geq -\delta^* \geq f^*_k$, that is, $p_k - G^{-1}(\beta) \leq 0$ we have

$$p_k(\phi(\hat{f}_k) - \phi(-\delta)) - (1 - p_k)(\phi(-\hat{f}_k) - \phi(\delta)) \geq 0.$$

Note that in the following two cases

- $\hat{f}_k \leq -\delta$ and $f^*_k \leq -\delta^*$;
- $\hat{f}_k \geq -\delta$ and $f^*_k \geq -\delta^*$,

we have $\Delta\mathcal{R}^k(\Gamma_{\hat{f},\delta}) = 0$. Therefore, from the above inequalities, on B^c_k and by assumption (6) we get

$$\mathbf{1}_{B^c_k} \mathbf{1}_{\{k \in (\Gamma_{\hat{f},\delta} \,\Delta\, \Gamma^*_\beta)\}} |p_k - G^{-1}(\beta)|^s \leq C(\Delta R^k_\phi(\hat{f})). \tag{20}$$

Therefore, since $s \geq 1$, we have

$$\begin{aligned}
\left(\mathbb{E}\left[\sum_{k=1}^K \mathbf{1}_{B^c_k \cap \{k \in (\Gamma_{\hat{f},\delta} \,\Delta\, \Gamma^*_\beta)\}} |p_k - G^{-1}(\beta)|\right]\right)^s &\leq \frac{1}{K}\sum_{k=1}^K \mathbb{E}\left[\mathbf{1}_{B^c_k \cap \{k \in \Gamma_{\hat{f},\delta} \Delta \Gamma^*_\beta\}} K^s |p_k - G^{-1}(\beta)|^s\right] \\
&\leq CK^{s-1}\Delta R_\phi(\hat{f}).
\end{aligned}$$

Moreover

$$\begin{aligned}
\mathbb{E}\left[\sum_{k=1}^K \mathbf{1}_{B_k} \mathbf{1}_{\{k \in (\Gamma_{\hat{f},\delta} \,\Delta\, \Gamma^*_\beta)\}} |p_k - G^{-1}(\beta)|\right] &\leq \sum_{k=1}^K \mathbb{P}(B_k) \\
&\leq \mathbb{E}\left[|G_{\hat{f}}(-\delta) - G_{\hat{f}}(-\delta^*)|\right] \\
&\leq \mathbb{E}\left[|G_{f^*}(-\delta^*) - G_{\hat{f}}(-\delta^*)|\right].
\end{aligned}$$

Finally, we get the following bound

$$\Delta\mathcal{R}(\Gamma_{\hat{f},\delta}) \leq K^{\frac{s-1}{s}}\Delta R_\phi(\hat{f})^{1/s} + \mathbb{E}\left[|G_{f^*}(-\delta^*) - G_{\hat{f}}(-\delta^*)|\right].$$

Now we observe that

$$|\mathbf{1}_{\{\hat{f}_k \geq -\delta^*\}} - \mathbf{1}_{\{f_k^* \geq -\delta^*\}}| \leq \mathbf{1}_{\{|p_k - G^{-1}(\beta)|^s \leq C\Delta R_\phi^k(\hat{f})\}}. \tag{21}$$

Therefore, for each $\gamma > 0$, we have

$$\begin{aligned}\mathbb{E}_X\left[|G_{f^*}(-\delta^*) - G_{\hat{f}}(-\delta^*)|\right] &\leq \sum_{k=1}^K \mathbb{P}_X\left(|p_k(X) - G^{-1}(\beta)|^s \leq C\Delta R_\phi^k(\hat{f})\right) \\ &\leq \sum_{k=1}^K \mathbb{P}_X\left(|p_k(X) - G^{-1}(\beta)| \leq \gamma^{1/s}\right) + \mathbb{P}_X\left(\gamma \leq C\Delta R_\phi^k(\hat{f})\right).\end{aligned}$$

Now using Markov Inequality, we have that

$$\begin{aligned}\sum_{k=1}^K \mathbb{P}_X\left(\gamma \leq C\Delta R_\phi^k(\hat{f})\right) &\leq \frac{C}{\gamma}\sum_{k=1}^K \mathbb{E}_X\left[\Delta R_\phi^k(\hat{f})\right] \\ &\leq \frac{C\Delta R_\phi(\hat{f})}{\gamma}.\end{aligned}$$

The above inequality yields

$$\mathbb{E}_X\left[|G_{f^*}(-\delta^*) - G_{\hat{f}}(-\delta^*)|\right] \leq \frac{C\Delta R_\phi(\hat{f})}{\gamma} + \sum_{k=1}^K \mathbb{P}_X\left(|p_k(X) - G^{-1}(\beta)| \leq \gamma^{1/s}\right). \tag{22}$$

Hence, with Equation (21), we get

$$\Delta\mathcal{R}(\Gamma_{\hat{f},\delta}) \leq K^{\frac{s-1}{s}}\Delta R_\phi(\hat{f})^{1/s} + \frac{C\Delta R_\phi(\hat{f})}{\gamma} + \sum_{k=1}^K \mathbb{P}_X\left(|p_k(X) - G^{-1}(\beta)| \leq \gamma^{1/s}\right).$$

The term $\sum_{k=1}^K \mathbb{P}_X\left(|p_k(X) - G^{-1}(\beta)| \leq \gamma^{1/s}\right) \to 0$ when $\gamma \to 0$, given that the distribution function of the $p_k's$ are continuous. Then using the convergence in distribution of $\Delta R_\phi(\hat{f})$ to zero, the last inequality ensures the desired result.

6.4 Proof of Proposition 5

For any $\beta \in (0, K)$, and conditional on $\mathcal{D}_n$ we define

$$\widetilde{G}^{-1}(\beta) = \inf\{t \in \mathbb{R} :\ \widetilde{G}(t) \leq \beta\}. \tag{23}$$

We note that Assumption $(A2)$ ensures that $t \mapsto \widetilde{G}(t)$ is continuous and then

$$\widetilde{G}(\widetilde{G}^{-1}(\beta)) = \beta.$$

Now, we have

$$\begin{aligned} |\widetilde{\Gamma}_\beta(X)| &= \sum_{k=1}^{K} \mathbf{1}_{\{\widetilde{G}(\hat{f}_k(X))\leq\beta\}} \\ &= \sum_{k=1}^{K} \mathbf{1}_{\{\hat{f}_k(X)\geq\widetilde{G}^{-1}(\beta)\}}. \end{aligned}$$

Hence, the last equation implies that

$$\mathbb{E}_X\left[|\widetilde{\Gamma}_\beta(X)||\mathcal{D}_n\right] = \sum_{k=1}^{K}\mathbb{P}_X\left(\hat{f}_k(X)\geq\widetilde{G}^{-1}(\beta)|\mathcal{D}_n\right) = \widetilde{G}(\widetilde{G}^{-1}(\beta)) = \beta. \tag{24}$$

Therefore, we obtain $\mathbf{E}\left[|\widetilde{\Gamma}_\beta(X)|\right] = \beta$. Also, we can write

$$\begin{aligned} \left|\mathbf{E}\left[\left|\hat{\Gamma}_\beta(X)\right|\right] - \beta\right| &\leq \left|\mathbf{E}\left[|\hat{\Gamma}_\beta(X)| - |\widetilde{\Gamma}_\beta(X)|\right]\right| \\ &\leq \left|\mathbf{E}\left[\sum_{k=1}^{K}\left(\mathbf{1}_{\{\hat{G}(\hat{f}_k(X))\leq\beta\}} - \mathbf{1}_{\{\widetilde{G}(\hat{f}_k(X))\leq\beta\}}\right)\right]\right| \\ &\leq \mathbf{E}\left[|\hat{\Gamma}_\beta(X)\,\Delta\,\widetilde{\Gamma}_\beta(X)|\right] \\ &\leq \sum_{k=1}^{K}\mathbf{E}\left[|\mathbf{1}_{\{\hat{G}(\hat{f}_k(X))\leq\beta\}} - \mathbf{1}_{\{\widetilde{G}(\hat{f}_k(X))\leq\beta\}}|\right] \\ &\leq \sum_{k=1}^{K}\mathbf{P}\left(|\hat{G}(\hat{f}_k(X) - \widetilde{G}(\hat{f}_k(X))| \geq |\widetilde{G}(\hat{f}_k(X)) - \beta|\right). \end{aligned}$$

Hence, applying Lemma 2 in the above inequality, we obtain the desired result.

6.5 Proof of Theorem 2

When there is no doubt, we suppress the dependence on X. First, let us state a intermediate result that is also needed to prove the theorem.

Lemma 3. *Consider $\Gamma_{\hat{f},\delta}$ the confidence set based on the score function $\hat{f}$ with information β (that is, $G_{\hat{f}}(-\delta) = \beta$). Under assumptions M_α^k, the following holds*

$$\Delta\mathcal{R}(\Gamma_{\hat{f},\delta}) \leq C(\alpha,s)\left\{K^{1-1/(s+\lambda-\lambda s)}\Delta R_\phi(\hat{f})^{1/(s+\lambda-\lambda s)} + K^{1-\lambda/(s+\lambda-\lambda s)}\Delta R_\phi(\hat{f})^{\lambda/(s+\lambda-\lambda s)}\right\},$$

where $\lambda = \frac{\alpha}{\alpha+1}$ and $C(\alpha,s)$ is non negative constant which depends only on α and s.

Proof. For each $k = 1,\ldots,K$, we define the following events $S_k = B_k^c \cap \{k \in (\Gamma_{\hat{f},\delta}\,\Delta\,\Gamma_\beta^*)\}$ and $T_k = B_k \cap \{k \in (\Gamma_{\hat{f},\delta}\,\Delta\,\Gamma_\beta^*)\}$, where the B_k's are the events given in Eq. (18). Now we observe that

$$\Delta\mathcal{R}(\Gamma_{\hat{f},\delta}) = \mathbb{E}\left[\sum_{k=1}^{K}\mathbf{1}_{S_k}|p_k - G^{-1}(\beta)|\right] + \mathbb{E}\left[\sum_{k=1}^{K}\mathbf{1}_{T_k}|p_k - G^{-1}(\beta)|\right] = \Delta\mathcal{R}_1(\Gamma_{\hat{f},\delta}) + \Delta\mathcal{R}_2(\Gamma_{\hat{f},\delta}),$$

where

$$\begin{aligned}\Delta\mathcal{R}_1(\Gamma_{\hat{f},\delta}) &= \mathbb{E}\left[\sum_{k=1}^{K}\mathbf{1}_{S_k}|p_k - G^{-1}(\beta)|\right],\\ \Delta\mathcal{R}_2(\Gamma_{\hat{f},\delta}) &= \mathbb{E}\left[\sum_{k=1}^{K}\mathbf{1}_{T_k}|p_k - G^{-1}(\beta)|\right].\end{aligned}$$

The end of the proof consists in controlling each of these two terms. Let us first consider $\Delta\mathcal{R}_1(\Gamma_{\hat{f},\delta})$. For $\varepsilon > 0$, we have

$$\begin{aligned}\Delta\mathcal{R}_1(\Gamma_{\hat{f},\delta}) &= \mathbb{E}\left[\sum_{k=1}^{K}\mathbf{1}_{S_k}|p_k - G^{-1}(\beta)|\left(\mathbf{1}_{\{|p_k-G^{-1}(\beta)|\geq\varepsilon\}} + \mathbf{1}_{\{|p_k-G^{-1}(\beta)|\leq\varepsilon\}}\right)\right]\\ &\leq \mathbb{E}\left[\sum_{k=1}^{K}\mathbf{1}_{S_k}\varepsilon^{1-s}|p_k - G^{-1}(\beta)|^s\right] + \varepsilon\sum_{k=1}^{K}\mathbb{P}(S_k)\\ &\leq C\varepsilon^{1-s}\Delta R_\phi(\hat{f}) + \varepsilon\sum_{k=1}^{K}\mathbb{P}(S_k), \qquad (25)\end{aligned}$$

where we used the assumption (6) and more precisely (20) to deduce the last inequality. To control $\sum_{k=1}^{K}\mathbb{P}(S_k)$, we require the following result that is a direct application of Lemma 5 in [BJM06].

Lemma 4. *Under the assumptions M_α^k we have*

$$\sum_{k=1}^{K}\mathbb{P}\left(S_k\right) \leq C(\alpha)\left(K^{1/\alpha}\Delta\mathcal{R}_1(\Gamma_{\hat{f},\delta})\right)^{\frac{\alpha}{\alpha+1}},$$

where $C(\alpha) > 0$ is a constant that depends only on α.

Proof. The proof of this result relies on the following simple fact: for all $\varepsilon > 0$

$$\begin{aligned}\mathbb{E}\left[\mathbf{1}_{S_k}|p_k - G^{-1}(\beta)|\right] &\geq \varepsilon\mathbb{E}\left[\mathbf{1}_{S_k}\mathbf{1}_{\{|p_k-G^{-1}(\beta)|\geq\varepsilon\}}\right]\\ &\geq \varepsilon\left[\mathbb{P}\left(S_k\right) - c_2\varepsilon^\alpha\right].\end{aligned}$$

Choosing $\varepsilon = \left(\frac{1}{c_2K(\alpha+1)}\sum_{k=1}^{K}\mathbb{P}\left(S_k\right)\right)^{1/\alpha}$ we get the lemma. □

We go back to the proof of Lemma 3. Applying Lemma 4 to (25), we get

$$\Delta\mathcal{R}_1(\Gamma_{\hat{f},\delta}) \leq C\varepsilon^{1-s}\Delta R_\phi(\hat{f}) + \varepsilon C(\alpha)\left(K^{1/\alpha}\Delta\mathcal{R}_1(\Gamma_{\hat{f},\delta})\right)^{\frac{\alpha}{\alpha+1}}.$$

Choosing $\varepsilon = \frac{s-1}{sC(\alpha)}K^{(\lambda-1)}\Delta\mathcal{R}_1(\Gamma_{\hat{f},\delta})^{(1-\lambda)}$, we obtain

$$\Delta\mathcal{R}_1(\Gamma_{\hat{f},\delta}) \leq C_1(\alpha,s)K^{1-1/(s-\lambda s+\lambda)}\Delta R_\phi(\hat{f})^{1/(s+\lambda-\lambda s)}, \qquad (26)$$

for a non negative constant $C_1(\alpha,s)$ that depends only on α and s.

Let us now focus on the second term, $\Delta\mathcal{R}_2(\Gamma_{\hat{f},\delta})$. Since the assumptions of Theorem 1 are satisfied, we can use Eq. (22) for any $\gamma > 0$. Combined with the Margin assumptions M_α^k, we obtain

$$\begin{aligned}\Delta\mathcal{R}_2(\Gamma_{\hat{f},\delta}) &\leq \mathbb{E}_X\left[|G_{f^*}(-\delta^*) - G_{\hat{f}}(-\delta^*)|\right] \\ &\leq \frac{C\Delta R_\phi(\hat{f})}{\gamma} + c_2 K\gamma^{\alpha/s}.\end{aligned}$$

Therefore, optimizing in γ, we have

$$\Delta\mathcal{R}_2(\Gamma_{\hat{f},\delta}) \leq C_2(\alpha,s)K^{1-\lambda/(s+\lambda-\lambda s)}\Delta R_\phi(\hat{f})^{\lambda/(s+\lambda-\lambda s)},$$

for a non negative constant $C_2(\alpha,s)$ that depends only on α and s. The result stated in the lemma is deduced by combining the last equation with Eq. (26) and by setting $C(\alpha,s) = \max\{C_1(\alpha,s); C_2(\alpha,s)\}$. □

We now state another important lemma that describes the behavior of the empirical minimizer of the ϕ-risk on the class $\mathcal{F}$:

Lemma 5. *Let $\bar{f} \in \mathcal{F}$ be the minimizer of $R_\phi(f)$ over $\mathcal{F}$. Under the assumptions of Theorem 2, we have that*

$$\mathbf{E}\left[R_\phi(\hat{f}_n) - R_\phi(\bar{f})\right] \leq \frac{3KL}{n} + \frac{KC(B,L)\log(N_n)}{n},$$

where $C(B,L) > 0$ is a constant that only depends on the constant B given in the Proposition 6 and on the Lipschitz constant L.

Proof. First, according to Eq. (11), for each $f \in \mathcal{F}$, we can write

$$\frac{R_\phi(f) + R_\phi(\bar{f})}{2} - R_\phi\left(\frac{f+\bar{f}}{2}\right) \geq \delta\left(\sqrt{\sum_{k=1}^K \mathbb{E}_X\left[(f_k - \bar{f}_k)^2(X)\right]}\right),$$

Hence, by assumption on the modulus of convexity, we deduce

$$\frac{R_\phi(f) + R_\phi(\bar{f})}{2} - R_\phi\left(\frac{f+\bar{f}}{2}\right) \geq c_1\sum_{k=1}^K \mathbb{E}_X\left[(f_k - \bar{f}_k)^2(X)\right].$$

Since $R_\phi\left(\frac{f+\bar{f}}{2}\right) \geq R_\phi(\bar{f})$, we obtain

$$\sum_{k=1}^K \mathbb{E}_X\left[(f_k - \bar{f}_k)^2(X)\right] \leq \frac{1}{2c_1}R_\phi(f) - R_\phi(\bar{f}).$$

Now, denoting by $h(z, f(x)) = \sum_{k=1}^K \phi(z_k f_k(x)) - \phi(z_k\bar{f}_k(x))$, we get the following bound

$$\begin{aligned}\mathbb{E}_X\left[h^2(Zf(X))\right] &\leq KL^2\sum_{k=1}^K \mathbb{E}_X\left[(f_k - \bar{f}_k)^2(X)\right] \\ &\leq \frac{KL^2}{2c_1}\mathbb{E}_X\left[h(Zf(X))\right], \qquad (27)\end{aligned}$$

where L is the Lipschitz constant L for ϕ. On the other hand, we have the following decomposition

$$R_\phi(\hat{f}) - R_\phi(\bar{f}) = R_\phi(\hat{f}) + 2(\hat{R}_\phi(\hat{f}) - \hat{R}_\phi(\bar{f})) - 2(\hat{R}_\phi(\hat{f}) - \hat{R}_\phi(\bar{f})) - R_\phi(\bar{f}).$$

Also, since $\hat{R}_\phi(\hat{f}) - \hat{R}_\phi(\bar{f}) \leq 0$, we get

$$\begin{aligned} R_\phi(\hat{f}) - R_\phi(\bar{f}) &\leq (R_\phi(\hat{f}) - R_\phi(\bar{f})) - 2(\hat{R}_\phi(\hat{f}) - \hat{R}_\phi(\bar{f})) \\ &\leq \frac{3KL}{n} + \sup_{f \in \mathcal{F}_n} (R_\phi(f) - R_\phi(\bar{f})) - 2(\hat{R}_\phi(f) - \hat{R}_\phi(\bar{f})), \end{aligned}$$

where $\mathcal{F}_n$ is the ϵ-net of $\mathcal{F}$ w.r.t the L_∞-norm and with $\epsilon = 1/n$. Now, using Bernstein's Inequality, we have that for all $f \in \mathcal{F}_n$ and $t > 0$

$$\begin{aligned} &\mathbf{P}\left((R_\phi(f) - R_\phi(\bar{f})) - 2(\hat{R}_\phi(f) - \hat{R}_\phi(\bar{f})) \geq t\right) \leq \\ &\qquad \mathbf{P}\left(2((R_\phi(f) - R_\phi(\bar{f})) - (\hat{R}_\phi(f) - \hat{R}_\phi(\bar{f})) \geq t + R_\phi(f) - R_\phi(\bar{f})\right) \\ &\qquad\qquad \leq \exp\left(-\frac{n(t + \mathbb{E}\left[h(Z, f(X))\right])^2/8}{\mathbb{E}\left[h^2(Z, f(X))\right]) + (2KLB/3)(t + \mathbb{E}\left[h(Z, f(X))\right]))}\right). \end{aligned}$$

Using Eq. (27), we get for all $f \in \mathcal{F}_n$

$$\mathbf{P}\left((R_\phi(f) - R_\phi(\bar{f})) - 2(\hat{R}_\phi(f) - \hat{R}_\phi(\bar{f})) \geq t\right) \leq \exp\left(-\frac{nt}{8(KL^2/(2c_1) + KLB/3)}\right),$$

Therefore, using a union bound argument, and then integrating we deduce that

$$\begin{aligned} \mathbf{E}\left[R_\phi(\hat{f}) - R_\phi(\bar{f})\right] &\leq \frac{3KL}{n} + \mathbf{E}\left[\sup_{f \in \mathcal{F}_n} (R_\phi(f) - R_\phi(\bar{f})) - 2(\hat{R}_\phi(f) - \hat{R}_\phi(\bar{f}))\right] \\ &\leq \frac{3KL}{n} + \frac{KC(B, L)\log(M_n)}{n}. \end{aligned}$$

□

We are now ready to conclude the proof of the theorem. We have the following inequality

$$|\mathcal{R}(\hat{\Gamma}_\beta) - \mathcal{R}^*_\beta| \leq \Delta\mathcal{R}(\widetilde{\Gamma}_\beta) + |\mathcal{R}(\hat{\Gamma}_\beta) - \mathcal{R}(\widetilde{\Gamma}_\beta)|. \tag{28}$$

We deal which each terms in the r.h.s separately. First, we have from Jensen's Inequality that

$$\left(\mathbf{E}\left[\Delta\mathcal{R}(\widetilde{\Gamma}_\beta)\right]\right)^{\frac{s+\lambda-\lambda s}{\lambda}} \leq \mathbf{E}\left[\Delta\mathcal{R}(\widetilde{\Gamma}_\beta)^{\frac{s+\lambda-\lambda s}{\lambda}}\right].$$

Hence, from Lemma 3, we deduce

$$\left(\mathbf{E}\left[\Delta\mathcal{R}(\widetilde{\Gamma}_\beta)\right]\right)^{\frac{s+\lambda-\lambda s}{\lambda}} \leq C(\alpha, s)^{\frac{s+\lambda-\lambda s}{\lambda}} K^{\frac{s+\lambda-\lambda s}{\lambda}-1} \mathbf{E}\left[\Delta R_\phi(\hat{f})\right].$$

Moreover, from Lemma 5, we have that

$$\mathbf{E}\left[\Delta R_\phi(\hat{f})\right] \leq \inf_{f \in \mathcal{F}} \Delta R_\phi(f) + \frac{3KL}{n} + \frac{KC(B, L)\log(M_n)}{n}.$$

Therefore, we can write

$$\mathbf{E}\left[\Delta\mathcal{R}(\widetilde{\Gamma}_\beta)\right] \leq C(\alpha,s)K^{1-\lambda/(s+\lambda-\lambda s)}\left\{\inf_{f\in\mathcal{F}}\Delta R_\phi(f) + \frac{3KL}{n} + \frac{KC(B,L)\log(N_n)}{n}\right\}^{\lambda/(s+\lambda-\lambda s)}. \tag{29}$$

For the second term $|\mathcal{R}(\hat{\Gamma}_\beta) - \mathcal{R}(\widetilde{\Gamma}_\beta)|$ in (28), we observe that

$$\mathbf{1}_{\{Y\notin\hat{\Gamma}_\beta(X)\}} - \mathbf{1}_{\{Y\notin\widetilde{\Gamma}_\beta(X)\}} = \sum_{k=1}^{K}\mathbf{1}_{\{Y=k\}}\mathbf{1}_{\left\{k\notin\hat{\Gamma}_\beta(X)\right\}} - \sum_{k=1}^{K}\mathbf{1}_{\{Y=k\}}\mathbf{1}_{\left\{k\notin\widetilde{\Gamma}_\beta(X)\right\}}.$$

Therefore, we can write

$$\begin{aligned}\mathbf{E}\left[\mathbf{1}_{Y\notin\hat{\Gamma}_\beta(X)\}} - \mathbf{1}_{\{Y\notin\widetilde{\Gamma}_\beta(X)\}}\right] &= \sum_{k=1}^{K}\mathbf{E}\left[p_k(X)\left(\mathbf{1}_{\left\{k\notin\hat{\Gamma}_\beta(X)\right\}} - \mathbf{1}_{\{k\notin\widetilde{\Gamma}_\beta(X)\}}\right)\right]\\ &= \sum_{k=1}^{K}\mathbf{E}\left[p_k(X)\left(\mathbf{1}_{\left\{\hat{G}(\hat{f}_k(X))>\beta\right\}} - \mathbf{1}_{\{\tilde{G}(\hat{f}_k(X))>\beta\}}\right)\right].\end{aligned}$$

Since $0 \leq p_k(X) \leq 1$ for all $k \in \{1,\dots,K\}$, the last equality implies

$$\begin{aligned}\left|\mathbf{R}\left(\hat{\Gamma}_\beta\right) - \mathbf{R}\left(\tilde{\Gamma}_\beta\right)\right| &= \left|\mathbf{E}\left[\mathbf{1}_{Y\notin\hat{\Gamma}_\beta(X)\}} - \mathbf{1}_{\{Y\notin\widetilde{\Gamma}_\beta(X)\}}\right]\right|\\ &\leq \sum_{k=1}^{K}\mathbf{E}\left[\left|\mathbf{1}_{\left\{\hat{G}(\hat{f}_k(X))>\beta\right\}} - \mathbf{1}_{\{\tilde{G}(\hat{f}_k(X))>\beta\}}\right|\right]\\ &\leq \sum_{k=1}^{K}\mathbf{P}\left(|\hat{G}(\hat{f}_k(X) - \widetilde{G}(\hat{f}_k(X))| \geq |\widetilde{G}(\hat{f}_k(X)) - \beta|\right).\end{aligned}$$

Therefore, Lemma 2 implies

$$\left|\mathbf{R}\left(\hat{\Gamma}_\beta\right) - \mathbf{R}\left(\tilde{\Gamma}_\beta\right)\right| \leq \frac{C'K}{\sqrt{N}}. \tag{30}$$

Injecting Eqs. (29) and (30) to Eq. (28) we conclude the proof of the theorem.

6.6 Proof of Proposition 6

We begin with the following decomposition

$$\tilde{R}^n_\phi(\hat{f}) - \tilde{R}^n_\phi(\tilde{f}) = \tilde{R}^n_\phi(\hat{f}) + 2(\hat{R}^n_\phi(\hat{f}) - \hat{R}^n_\phi(\tilde{f})) - 2(\hat{R}^n_\phi(\hat{f}) - \hat{R}^n_\phi(\tilde{f})) - \tilde{R}^n_\phi(\tilde{f}),$$

since $\hat{R}^n_\phi(\hat{f}) - \hat{R}^n_\phi(\tilde{f}) \leq 0$, we get

$$\tilde{R}^n_\phi(\hat{f}) - \tilde{R}^n_\phi(\tilde{f}) \leq (\tilde{R}^n_\phi(\hat{f}) - \tilde{R}^n_\phi(\tilde{f})) - 2(\hat{R}^n_\phi(\hat{f}) - \hat{R}^n_\phi(\tilde{f})). \tag{31}$$

Now, we denote by $\mathcal{C}_n = \{(i_1/n,\dots,i_M/n),(i_1,\dots,i_M)\in\{0,\dots,n\}\}\cap\mathcal{C}_M$,
and $\mathcal{F}_n = \{f = \sum_{i=1}^{M} m_i f_i,\ \ m_1,\dots m_M) \in \mathcal{C}_n\}$. For each $f \in \mathrm{conv}(\mathcal{F})$, there exists $f_n \in \mathcal{F}_n$ such that

$$\begin{aligned}|\tilde{R}^n_\phi(f) - \tilde{R}^n_\phi(f_n)| &\leq \frac{2KLBM}{n}\\ |\hat{R}^n_\phi(f) - \hat{R}^n_\phi(f_n)| &\leq \frac{2KLBM}{n}.\end{aligned}$$

Therefore, with Equation (31), we obtain

$$\tilde{R}^n_\phi(\hat{f}) - \tilde{R}^n_\phi(\tilde{f}) \leq \frac{6LKB}{n} + \sup_{f \in \mathcal{F}_n} (\tilde{R}^n_\phi(f) - \tilde{R}^n_\phi(\bar{f})) - 2(\hat{R}^n_\phi(f) - \hat{R}^n_\phi(\bar{f})).$$

Now, Similar arguments as in [DvdL05] and Lemma 5 yield the proposition.

References

[BJM06] P. Bartlett, M. Jordan, and J. McAuliffe. Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101(473):138–156, 2006.

[BM06] P. Bartlett and S. Mendelson. Empirical minimization. *Probability Theory and Related Fields*, 135(3):311–334, 2006.

[CCB16] A. Choromanska, K. Choromanski, and M. Bojarski. On the boosting ability of top-down decision tree learning algorithm for multiclass classification. preprint, 2016.

[Cho70] C.K. Chow. On optimum error and reject trade-off. *IEEE Transactions on Information Theory*, 16:41–46, 1970.

[dCDB09] J. del Coz, J. Díez, and A. Bahamonde. Learning nondeterministic classifiers. *Journal of Machine Learning Research*, 10:2273–2293, 2009.

[DH15] C. Denis and M. Hebiri. Consistency of plug-in confidence sets for classification in semi-supervised learning. preprint, 2015.

[DvdL05] S. Dudoit and M. van der Laan. Asymptotics of cross-validated risk estimation in estimator selection and performance assessment. *Statistical Methodology*, 2(2):131–154, 2005.

[FHT00] J. Friedman, T. Hastie, and R. Tibshirani. Additive logistic regression: a statistical view of boosting. *The Annals of Statistics*, 28(2):337–407, 2000.

[FS97] Y. Freund and R. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences*, 55(1):119–139, 1997.

[HW06] R. Herbei and M. Wegkamp. Classification with reject option. *The Canadian Journal of Statistics*, 34(4):709–721, 2006.

[Lei14] J. Lei. Classification with confidence. *Biometrika*, 101(4):755–769, 2014.

[LRW13] J. Lei, J. Robins, and L. Wasserman. Distribution-free prediction sets. *Journal of the American Statistical Association*, 108(501):278–287, 2013.

[TB07] A. Tewari and P. Bartlett. On the consistency of multiclass classification methods. *Journal of Machine Learning Research*, 8:1007–1025, 2007.

[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.

[vdLPH07] M. van der Laan, E. Polley, and A. Hubbard. Super learner. *Statistical Applications in Genetics and Molecular Biology*, 6, 2007.

[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 *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.

[Vov02] V. Vovk. Asymptotic optimality of transductive confidence machine. In *Algorithmic learning theory*, volume 2533 of *Lecture Notes in Computer Science*, pages 336–350. Springer, Berlin, 2002.

[WLW04] T.-F. Wu, C.-J. Lin, and R. Weng. Probability estimates for multi-class classification by pairwise coupling. *Journal of Machine Learning Research*, 5:975–1005, 2004.

[WY11] M. Wegkamp and M. Yuan. Support vector machines with a reject option. *Bernoulli*, 17(4):1368–1385, 2011.

[YW10] M. Yuan and M. Wegkamp. Classification methods with reject option based on convex risk minimization. *Journal of Machine Learning Research*, 11:111–130, 2010.

[Zha04] T. Zhang. Statistical behavior and consistency of classification methods based on convex risk minimization. *The Annals of Statistics*, 32(1):56–85, 02 2004.