A nearest-neighbor-based ensemble classifier and its large-sample optimality

Mojirsheibani, Majid; Pouliot, William

DOI:
10.1080/00949655.2021.1882458

A nearest-neighbor-based ensemble classifier and its large-sample optimality

Majid Mojirsheibani[1] and William Pouliot[2]

Department of Mathematics, California State University Northridge, CA, 91330, USA[1]

Department of Economics, University of Birmingham, B15 2TT, UK[2]

Abstract

A nonparametric approach is proposed to combine several individual classifiers in order to construct an asymptotically more accurate classification rule in the sense that its misclassification error rate is, asymptotically, at least as low as that of the best individual classifier. The proposed method uses a nearest neighbor type approach to estimate the conditional expectation of the class associated with a new observation (conditional on the vector of individual predictions). Both mechanics and the theoretical validity of the proposed approach are discussed. As an interesting byproduct of our results, it is shown that the proposed method can also be applied to any single classifier in which case the resulting new classifier will be at least as good as the original one. Several numerical examples, involving both real and simulated data, are also given. These numerical studies further confirm the superiority of the proposed classifier.

Keywords: Nonparametric, asymptotics, classification.

1 Introduction

Consider the following standard two-group classification problem. Let $(\boldsymbol{\chi}, Y)$ be a random pair, where $\boldsymbol{\chi}$ takes values in a metric space $(\mathcal{F}, \rho)$ and $Y \in \{0, 1\}$, called the class label, must be predicted based on $\boldsymbol{\chi}$. Here, $\mathcal{F}$ is not necessarily confined to $\mathbb{R}^d$. In classification one wants to find a classifier, $g : \mathcal{F} \to \{0, 1\}$, whose misclassification error, $\mathbb{P}\{g(\boldsymbol{\chi}) \neq Y\}$, is as small as possible. The classifier with the lowest misclassification error, called the Bayes classifier, is given by $g_{\text{B}}(\chi) = 1$ if $\mathbb{P}\{Y = 1|\boldsymbol{\chi} = \chi\} > 1/2$, and

[1]Corresponding author. Email: majid.mojirsheibani@csun.edu

This work was supported by the NSF under Grant DMS-1916161 of Majid Mojirsheibani.

[2]Email: w.pouliot@bham.ac.uk

$g_{\text{B}}(\chi) = 0$ otherwise; see, for example, Cérou and Guyader [9], Abraham et al. [1], and Devroye, et al. [12]. Although our setup is expressed for the popular two-class problem, all the results in this paper can be extended in a straightforward manner to the multi-class classification problem; see Remark 2.

Of course, in practice the distribution of $(\boldsymbol{\chi}, Y)$ is virtually always unknown and, typically, one only has access to a training sample of n independent and identically distributed (iid) observations $\mathbb{T}_n = \{(\boldsymbol{\chi}_1, Y_1), \ldots, (\boldsymbol{\chi}_n, Y_n)\}$ from the distribution of $(\boldsymbol{\chi}, Y)$. Much of the theory of classification deals with the construction of sample-based classification rules $\widehat{g}_n$ based on $\mathbb{T}_n$ whose error rates are somehow as small as possible. Of course, the choice of $\widehat{g}_n$ is at the practitioner's discretion and there may be several different options; therefore, let $\widehat{g}_{n,1}, \ldots, \widehat{g}_{n,J}$ be $J \geq 1$ different classification rules for predicting Y. Here, $\widehat{g}_{n,1}$ may be a linear classifier, $\widehat{g}_{n,2}$ a kernel classifier, while $\widehat{g}_{n,3}$ may be Breiman's [8] random forest classifier, etc. The aim is then to combine these individual classifiers in such a way that the resulting classifier is in some sense at least as good as the best individual classifier.

There is a vast literature on combined or ensemble methods and there are many different approaches available; this is particularly true for the important problems of classification and regression function estimation. One may divide the existing methods into roughly two types: (a) those approaches that involve a large number of similar or homogeneous base models. Relevant examples here include Breiman's [7, 8] random forest, Lin and Jeon [17], Biau et al. [3], and Rahman et al. [19]. (b) Those approaches that combine a number of models or estimators that are constructed based on different theories or estimation methods. Results under (b) include Fischer and Mougeot [13], Biau et al. [4], Cholaquidis et al. [10], Balakrishnan and Mojirsheibani [2], Mojirsheibani [18], and LeBlanc and Tibshirani [16]. The methods employed in the cited papers under (b) are mainly nonlinear in nature, which is also the framework of the current paper. There is also a large body of literature on linear and convex aggregation methods; in fact, Chapter

3 of the monograph by Giraud [14] presents a detailed account of such methods along with many relevant references.

In passing, we also note that there are other taxonomies for characterizing combined classifiers. In fact, as explained in Rokach [20, 21], combined classification methods can be put into two main categories: *weighting* methods and *non-weighing* or *meta-learning*. Popular weighting methods include the majority voting employed by Breiman [7, 8] in the context of tree classification, and by Xu et al. [24] for the problem of handwriting recognition. Weighted-averaging of estimated class conditional probabilities that are produced by each classifier, has also been studied by several researchers; results along these lines include the work of Xu et al. [24], Breiman [6], and LeBlanc and Tibshirani [16]. There are also other weighting methods that can be found in Rokach [20, 21]. Meta-learning methods typically work by using the predicted values of the individual classifiers on the data. Relevant results along these lines include the stacked generaliztion of Wolpert [23], Breiman's [6] stacked method, and the nonlinear methods of Mojirsheibani [18], Balakrishnan and Mojirsheibani [2], Biau et al. [4], and Cholaquidis et al. [10]. For more on meta-learning methods, one may refer to Rokach [20].

In the next section, we consider the problem of combining several individual classifiers in such a way that the resulting ensemble is, asymptotically, at least as good as the best individual one. The paper is organized as follows. Section 2 presents the main results, where both the mechanics and the theoretical validity of the proposed approach are discussed. Numerical studies involving both simulated as well as real data are carried out in section 3; these studies further confirm the good finite-sample performance of the proposed approach. All proofs are deferred to section 4.

2 Main results

In order to motivate our proposed method, consider the following hypothetical oversimplified setup. Let $g_1, \cdots, g_J$ be J classifiers for predicting Y based on $\boldsymbol{\chi}$ (no data yet). Here, each g_j is a map of the form $g_j : \mathcal{F} \to \{0,1\}$. Define the combined classifier $G^* : \{0,1\}^J \to \{0,1\}$, for predicting the same Y, by

$$G^*\big(g_1(\boldsymbol{\chi}), \cdots, g_J(\boldsymbol{\chi})\big) \;=\; \begin{cases} 1 & \text{if } \mathbb{E}\left[(2Y-1)\big|g_1(\boldsymbol{\chi}), \cdots, g_J(\boldsymbol{\chi})\right] > 0 \\ 0 & \text{otherwise.} \end{cases} \tag{1}$$

Then we have the following elementary result

Proposition 1 *The combined classifier G^* in (1) is optimal, i.e.,*

$$\mathbb{P}\Big\{G^*\big(g_1(\boldsymbol{\chi}), \cdots, g_J(\boldsymbol{\chi})\big) \neq Y\Big\} \;=\; \inf_{G:\ \{0,1\}^J \to \{0,1\}} \mathbb{P}\Big\{G\big(g_1(\boldsymbol{\chi}), \cdots, g_J(\boldsymbol{\chi})\big) \neq Y\Big\}.$$

We also observe that in view of Proposition 1, and without further ado, one has

$$\mathbb{P}\left\{G^*\big(g_1(\boldsymbol{\chi}), \cdots, g_J(\mathbf{X})\big) \neq Y\right\} \;\leq\; \min_{1 \leq j \leq J} \mathbb{P}\big\{g_j(\boldsymbol{\chi}) \neq Y\big\}; \tag{2}$$

in other words, G^* is at least as good as the best classifier among $g_1, \ldots, g_J$. When $J = 1$, we may simply write g instead of g_1 in which case (2) reduces to

$$\mathbb{P}\{G^*(g(\boldsymbol{\chi})) \neq Y\} \;\leq\; \mathbb{P}\{g(\boldsymbol{\chi}) \neq Y\}\,,$$

where the equality holds when g is the Bayes classifier, i.e., $g(\boldsymbol{\chi}) = 1$ if $\mathbb{E}[(2Y-1)|\boldsymbol{\chi}] > 0$, otherwise $g(\boldsymbol{\chi}) = 0$.

Next, suppose that we have J individual classifiers, $\widehat{g}_{n,1}, \ldots, \widehat{g}_{n,J}$, constructed based on the data $\mathbb{T}_n$ for predicting Y. As explained in the introduction, these could be J very different classifiers; for example, $\widehat{g}_{n,1}$ may be a linear classifier, $\widehat{g}_{n,2}$ a kernel classifier, $\widehat{g}_{n,3}$ may be a random forest classifier or the support vector machine, etc. Then (1) in conjunction with Proposition 1 suggests considering a combined classifier, G_n^*, of the following form

$$G_n^*\big(\widehat{g}_{n,1}(\boldsymbol{\chi}), \cdots, \widehat{g}_{n,J}(\boldsymbol{\chi})\big) = \begin{cases} 1 & \text{if } \mathbb{E}\left[(2Y-1)\big|\widehat{g}_{n,1}(\boldsymbol{\chi}), \cdots, \widehat{g}_{n,J}(\boldsymbol{\chi})\right] > 0 \\ 0 & \text{otherwise.} \end{cases} \tag{3}$$

Although (3) is not available in practice, the following counterpart of Proposition 1 shows that (3) is in fact theoretically optimal in the important sense that its overall error rate is the smallest:

Proposition 2 *Let G_n^* be the combined classifier in (3). Then*

$$\mathbb{P}\Big\{G_n^*\big(\widehat{g}_{n,1}(\boldsymbol{\chi}),\cdots,\widehat{g}_{n,J}(\boldsymbol{\chi})\big)\neq Y\Big\} \;=\; \inf_{G:\ \{0,1\}^J\to\{0,1\}} \mathbb{P}\Big\{G\big(\widehat{g}_{n,1}(\boldsymbol{\chi}),\cdots,\widehat{g}_{n,J}(\boldsymbol{\chi})\big)\neq Y\Big\},$$

and in particular $\mathbb{P}\big\{G_n^*\big(\widehat{g}_{n,1}(\boldsymbol{\chi}),\cdots,\widehat{g}_{n,J}(\boldsymbol{\chi})\big)\neq Y\big\} \;\le\; \min_{1\le j\le J}\mathbb{P}\big\{\widehat{g}_{n,j}(\boldsymbol{\chi})\neq Y\big\}$.

Obviously, the theoretically optimal classifier G_n^* is not useful in practice because the conditional expectation on the right hand side of (3) is virtually always unknown. Therefore, in what follows, the aim is to construct estimates of (3) whose error rates can be arbitrarily close to that of G_n^*, as the sample size n grows larger and larger. In what follows, we propose a rather simple-to-implement nearest neighbor (NN) type method that works as follows. Randomly split the data $\mathbb{T}_n$ into a training sample $\mathbb{T}_m$ of size m and a testing sequence $\mathbb{T}_\ell$ of size $\ell = n-m$, where $\mathbb{T}_m \cup \mathbb{T}_\ell = \mathbb{T}_n$ and $\mathbb{T}_m \cap \mathbb{T}_\ell = \emptyset$. Let $\widehat{g}_{m,1},\cdots,\widehat{g}_{m,J}$ be the J individual classifiers constructed based on $\mathbb{T}_m$ only, and consider the k-NN type combined classifier $G_{n,k}$, $1\le k\le \ell$, given by

$$G_{n,k}\big(\widehat{g}_{m,1}(\boldsymbol{\chi}),\cdots,\widehat{g}_{m,J}(\boldsymbol{\chi})\big) \;=\; \begin{cases} 1 & \text{if } \ \sum_{i:\,(\boldsymbol{\chi}_i,Y_i)\in\mathbb{T}_\ell}(2Y_i-1)\cdot\mathcal{I}_m(k,\boldsymbol{\chi},\boldsymbol{\chi}_i)>0 \\ 0 & \text{otherwise,} \end{cases} \tag{4}$$

where $\mathcal{I}_m(k,\boldsymbol{\chi},\boldsymbol{\chi}_i) = \mathbb{I}_{\{(\widehat{g}_{m,1}(\boldsymbol{\chi}_i),\cdots,\widehat{g}_{m,J}(\boldsymbol{\chi}_i)) \text{ is among the } k \text{ nearest neighbors of } (\widehat{g}_{m,1}(\boldsymbol{\chi}),\cdots,\widehat{g}_{m,J}(\boldsymbol{\chi}))\}}$; (5)

here, the distance between two J-dimensional vectors in $\{0,1\}^J$ is measured with respect to the Hamming distance, i.e., the number of discrepancies between the corresponding components of the two vectors. In the case of ties, the nearest neighbor to be selected is determined by random chance; thus, for example, if $(\widehat{g}_{m,1}(\boldsymbol{\chi}_i),\cdots,\widehat{g}_{m,J}(\boldsymbol{\chi}_i))$ is the third nearest neighbor of $(\widehat{g}_{m,1}(\boldsymbol{\chi}),\cdots,\widehat{g}_{m,J}(\boldsymbol{\chi}))$ for $i=2,7$, and 10, then we randomly choose one of these three candidates (and its corresponding Y_i) to be used as the third nearest

neighbor of $(\widehat{g}_{m,1}(\boldsymbol{\chi}), \cdots, \widehat{g}_{m,J}(\boldsymbol{\chi}))$. Then, the proposed combined classifier, denoted by $G_{n,\widehat{k}}$, is given by (4) with k replaced by $\widehat{k}$ that minimizes the re-substitution error, i.e.,

$$\widehat{k} = \underset{1\leq k\leq \ell}{\operatorname{argmin}} \frac{1}{\ell} \sum_{i:\,(\boldsymbol{\chi}_i, Y_i)\,\in\,\mathbb{D}_\ell} \mathbb{I}_{\left\{G_{n,k}\left(\widehat{g}_{m,1}(\boldsymbol{\chi}_i), \cdots, \widehat{g}_{m,J}(\boldsymbol{\chi}_i)\right) \neq Y_i\right\}} \tag{6}$$

To study the asymptotic optimality of $G_{n,\widehat{k}}$ we first state two assumptions. Define the quantity

$$\mathcal{S}_{m,\ell}(\boldsymbol{\chi}) \;=\; \sum_{i:\;\boldsymbol{\chi}_i\in\mathbb{T}_\ell} \mathbb{I}_{\left\{\left(\widehat{g}_{m,1}(\boldsymbol{\chi}_i), \cdots, \widehat{g}_{m,J}(\boldsymbol{\chi}_i)\right) = \left(\widehat{g}_{m,1}(\boldsymbol{\chi}), \cdots, \widehat{g}_{m,J}(\boldsymbol{\chi})\right)\right\}}, \tag{7}$$

and consider the following assumptions:

Assumption A.

The quantity $\mathcal{S}_{m,\ell}(\boldsymbol{\chi})$ diverges with n: $\mathcal{S}_{m,\ell}(\boldsymbol{\chi}) \to \infty$, in probability, as n (and thus ℓ) $\to \infty$.

Assumption B.

For the classifier G_n^* in (3), one has $\mathbb{P}\big\{G_n^*\big(\widehat{g}_{n,1}(\boldsymbol{\chi}), \cdots, \widehat{g}_{n,J}(\boldsymbol{\chi})\big) \neq Y\big\} \to c$, for some constant c in $[0, 1]$, as $n \to \infty$.

Assumption A above is not unrealistic at all; to appreciate this, observe that conditional on $\mathbb{T}_m$ and $\boldsymbol{\chi}$, the quantity $\mathcal{S}_{m,\ell}(\boldsymbol{\chi})$ merely represents the total number of "successes" in ℓ independent Bernoulli trials, which is, intuitively, expected to diverge as $\ell \to \infty$. In fact, alternative versions of this assumption have already been used in the literature (e.g., Devroye et al. [12]; p. 94)). The following result summarizes the asymptotic optimality of the proposed combined classifier $G_{n,\widehat{k}}$.

Theorem 1 *Let $G_{n,\widehat{k}}$ be the nearest neighbor combined classifier defined in (4), where $\widehat{k}$ is the minimizer of the empirical error in (6). If Assumptions A and B hold then*

$$\mathbb{P}\Big\{G_{n,\widehat{k}}\big(\widehat{g}_{m,1}(\boldsymbol{\chi}), \cdots, \widehat{g}_{m,J}(\boldsymbol{\chi})\big) \neq Y\Big\} - \inf_{G:\{0,1\}^J\to\{0,1\}} \mathbb{P}\Big\{G\big(\widehat{g}_{n,1}(\boldsymbol{\chi}), \cdots, \widehat{g}_{n,J}(\boldsymbol{\chi})\big) \neq Y\Big\} \longrightarrow 0,$$

as $n \to \infty$. In particular, $G_{n,\widehat{k}}$ is asymptotically at least as good as the best individual classifier, i.e.,

$$\limsup_{n\to\infty} \max_{1\leq j\leq J} \Big[\mathbb{P}\Big\{G_{n,\widehat{k}}\big(\widehat{g}_{m,1}(\boldsymbol{\chi}), \cdots, \widehat{g}_{m,J}(\boldsymbol{\chi})\big) \neq Y\Big\} - \mathbb{P}\Big\{\widehat{g}_{n,j}(\boldsymbol{\chi})\big) \neq Y\Big\}\Big] \;\leq\; 0\,,$$

where $\widehat{g}_{n,j}$ is the j-th individual classifier constructed based on the full data $\mathbb{T}_n$.

Remark 1 Let $G_{n,k}$ and $\mathcal{S}_{m,\ell}(\boldsymbol{\chi})$ be as in (4) and (7), respectively. Now, if we choose k to be equal to $\mathcal{S}_{m,\ell}(\boldsymbol{\chi})$ in $G_{n,k}$, provided that $\mathcal{S}_{m,\ell}(\boldsymbol{\chi}) > 0$, then our proposed combined classifier reduces to that of Balakrishnan and Mojirsheibani [2]. To appreciate this, observe that in this case (4) becomes

$$
\begin{aligned}
& G_{n,\,\mathcal{S}_{m,\ell}(\boldsymbol{\chi})}\big(\widehat{g}_{m,1}(\boldsymbol{\chi}),\cdots,\widehat{g}_{m,J}(\boldsymbol{\chi})\big) \\
= & \begin{cases} 1 & \text{if } \mathbb{I}_{\{\mathcal{S}_{m,\ell}(\boldsymbol{\chi})>0\}} \cdot \sum_{i:\,(\boldsymbol{\chi}_i,Y_i)\in\mathbb{T}_\ell} (2Y_i-1)\cdot \mathcal{I}_m\big(\mathcal{S}_{m,\ell}(\boldsymbol{\chi}),\boldsymbol{\chi},\boldsymbol{\chi}_i\big) \;>0 \\ 0 & \text{otherwise,} \end{cases} \\
& \qquad \big(\text{ where } \mathcal{I}_m\big(\mathcal{S}_{m,\ell}(\boldsymbol{\chi}),\boldsymbol{\chi},\boldsymbol{\chi}_i\big) \text{ is as in (5) with } k \text{ replaced by } \mathcal{S}_{m,\ell}(\boldsymbol{\chi})\big) \\
= & \begin{cases} 1 & \text{if } \mathbb{I}_{\{\mathcal{S}_{m,\ell}(\boldsymbol{\chi})>0\}} \sum_{i:\,(\boldsymbol{\chi}_i,Y_i)\in\mathbb{T}_\ell} (2Y_i-1)\cdot \mathbb{I}_{\left\{(\widehat{g}_{m,1}(\boldsymbol{\chi}_i),\cdots,\widehat{g}_{m,J}(\boldsymbol{\chi}_i))=(\widehat{g}_{m,1}(\boldsymbol{\chi}),\cdots,\widehat{g}_{m,J}(\boldsymbol{\chi}))\right\}} >0 \\ 0 & \text{otherwise,} \end{cases}
\end{aligned} \tag{8}
$$

where we have used the fact that, in view of the definition of $\mathcal{S}_{m,\ell}(\boldsymbol{\chi})$,

$$
\mathbb{I}_{\{\mathcal{S}_{m,\ell}(\boldsymbol{\chi})>0\}} \times \mathbb{I}_{\left\{(\widehat{g}_{m,1}(\boldsymbol{\chi}_i),\cdots,\widehat{g}_{m,J}(\boldsymbol{\chi}_i)) \text{ is among the } \mathcal{S}_{m,\ell}(\boldsymbol{\chi}) \text{ nearest neighbors of } (\widehat{g}_{m,1}(\boldsymbol{\chi}),\cdots,\widehat{g}_{m,J}(\boldsymbol{\chi}))\right\}}
$$

is equal to 1 if and only if $\mathbb{I}_{\left\{(\widehat{g}_{m,1}(\boldsymbol{\chi}_i),\cdots,\widehat{g}_{m,J}(\boldsymbol{\chi}_i)) = (\widehat{g}_{m,1}(\boldsymbol{\chi}),\cdots,\widehat{g}_{m,J}(\boldsymbol{\chi}))\right\}} = 1$. It is straightforward to see that (8) is equivalent to

$$
\begin{cases} 1 & \text{if } \frac{\mathbb{I}_{\{\mathcal{S}_{m,\ell}(\boldsymbol{\chi})>0\}}}{\mathcal{S}_{m,\ell}(\boldsymbol{\chi})} \left[\sum_{i:\,(\boldsymbol{\chi}_i,Y_i)\in\mathbb{T}_\ell} Y_i \cdot \mathbb{I}_{\left\{(\widehat{g}_{m,1}(\boldsymbol{\chi}_i),\cdots,\widehat{g}_{m,J}(\boldsymbol{\chi}_i)) = (\widehat{g}_{m,1}(\boldsymbol{\chi}),\cdots,\widehat{g}_{m,J}(\boldsymbol{\chi}))\right\}}\right] > \frac{1}{2} \\ 0 & \text{otherwise,} \end{cases} \tag{9}
$$

which is the combined classifier of Balakrishnan and Mojirsheibani [2], where, by convention, $0/0 := 0$. The classifier in (9) is essentially a weighted average of all $Y_i \in \mathbb{T}_\ell$, where the weights are indicator functions $\mathbb{I}_{\left\{(\widehat{g}_{m,1}(\boldsymbol{\chi}_i),\cdots,\widehat{g}_{m,J}(\boldsymbol{\chi}_i))=(\widehat{g}_{m,1}(\boldsymbol{\chi}),\cdots,\widehat{g}_{m,J}(\boldsymbol{\chi}))\right\}}$, each of which will be 1 if and only if $\widehat{g}_{m,j}(\boldsymbol{\chi}_i) = \widehat{g}_{m,j}(\boldsymbol{\chi})$ for all $j \in \{1,2,\ldots,J\}$. Unfortunately, from a practical point of view, if there are a few weak/poor classifiers among $\widehat{g}_{m,1},\ldots,\widehat{g}_{m,J}$, then one could end up with $\widehat{g}_{m,j}(\boldsymbol{\chi}_i) \neq \widehat{g}_{m,j}(\boldsymbol{\chi})$ for a large number of $\boldsymbol{\chi}_i$'s in $\mathbb{T}_\ell$ and this can hold true even when $\boldsymbol{\chi}_i$ and $\boldsymbol{\chi}$ belong to the same class. This means that many of the weights (i.e., the indicator functions) in (9) will be zero, which in turn reduces the predictive performance of (9). Our proposed method in this paper circumvents this vulnerability of the combined classifier of Balakrishnan and Mojirsheibani [2] in (9) by allowing a few weak classifiers to "misbehave" or be incorrect in their predictions. Thus, unlike (9), the new classifier $G_{n,\widehat{k}}$ is not seriously affected by the impact of a few poor classifiers.

Remark 2 The results of this section can be extended to the M-group ($M \geq 2$) classification problem in a straightforward manner. More specifically, let $(\boldsymbol{\chi}, Y)$ be a random pair where $\boldsymbol{\chi}$ is as before, but $Y \in \{1,\ldots,M\}$. A classifier of the form $g^*(\chi) := \text{argmax}_{1\leq k\leq M}\, \mathbb{P}\{Y = k|\boldsymbol{\chi} = \chi\}$ is optimal in the sense that $\mathbb{P}\{g^*(\boldsymbol{\chi}) \neq$

$Y\} = \min_{g:\mathcal{F}\to\{1,\ldots,M\}} \mathbb{P}\{g(\boldsymbol{\chi}) \neq Y\}$; see, for example, Devroye and Györfi [11], ch. 10. In this case, with $\mathcal{I}_m(k, \boldsymbol{\chi}, \boldsymbol{\chi}_i)$ as in (5), we have the following counterpart of (4):

$$G_{n,k}\big(\widehat{g}_{m,1}(\boldsymbol{\chi}), \cdots, \widehat{g}_{m,J}(\boldsymbol{\chi})\big) \;=\; \operatorname*{argmax}_{1\le j\le M} \sum_{i:\,(\boldsymbol{\chi}_i, Y_i)\in\mathbb{T}_\ell} \mathbb{I}_{\{Y_i=j\}} \cdot \mathcal{I}_m(k, \boldsymbol{\chi}, \boldsymbol{\chi}_i)$$

and the proposed combined classifier is given by $G_{n,\widehat{k}}$, where $\widehat{k}$ is as in (6). It can be shown that, under assumption A and the version of B corresponding to the M-group problem, the conclusion of Theorem 1 continues to hold in the general M-group problem in the sense that

$$\mathbb{P}\Big\{G_{n,\widehat{k}}\big(\widehat{g}_{m,1}(\boldsymbol{\chi}), \cdots, \widehat{g}_{m,J}(\boldsymbol{\chi})\big) \neq Y\Big\} - \inf_{G:\,\{1,\ldots,M\}^J\to\{1,\ldots,M\}} \mathbb{P}\Big\{G\big(\widehat{g}_{n,1}(\boldsymbol{\chi}), \cdots, \widehat{g}_{n,J}(\boldsymbol{\chi})\big) \neq Y\Big\} \longrightarrow 0, \quad \text{as } n\to\infty.$$

Remark 3 In the case where $J = 1$, Theorem 1 essentially implies that, given any initial classifier $\widehat{g}_n$, the error of the new classifier $G_{n,\widehat{k}}$ defined via (4) and (6) can always be asymptotically less than or equal to that of $\widehat{g}_n$. To the best of our knowledge, this is a new result in the literature.

Remark 4 Since, for each $k = 1, \ldots, \ell$, the nearest neighbor type combined classifier $G_{n,k}$ in (4) is constructed based on one sample split, its performance can be affected by the particular split used. Thus, unless n is very large, a "bad" split can in practice result in a poor choice of $\widehat{k}$ in (6) which will lead to a poor corresponding classifier $G_{n,\widehat{k}}$. This practical issue suggests using several random splits and taking their average. More precisely, with N sample splits, each split will produce an estimate $\widehat{k}_b$ of k, where $b = 1, \ldots, N$, and the corresponding predicted class Y (corresponding to $\boldsymbol{\chi}$) is given by $\widehat{Y}_b = G_{n,\widehat{k}_b}\big(\widehat{g}_{m,1}(\boldsymbol{\chi}), \cdots, \widehat{g}_{m,J}(\boldsymbol{\chi})\big)$. Since each $\widehat{Y}_b$ is either 0 or 1, the overall predicted value of Y is taken to be 1 if $N^{-1}\sum_{b=1}^{N} \widehat{Y}_b > 1/2$, and 0 otherwise. As for the choice of N, our experience shows that one can expect good results with N as small as 15 or 20.

3 Numerical examples

In what follows, we study the prediction of the class variable, Y ($= 0$ or 1, corresponding to the random covariate $\boldsymbol{\chi}$, using the ensemble methods proposed in this paper.

Example A *(Simulated data)*.

In this example we consider the prediction of $Y = 0$ or 1, based on $\boldsymbol{\chi} \in \mathbb{R}^{10}$, where $\mathbb{P}\{Y = 1\} = \mathbb{P}\{Y = 0\} = 0.5$. Here, we have taken $\boldsymbol{\chi} \sim N_{10}\big(\mathbf{0}, 8\boldsymbol{\Sigma}\big)$ whenever

$Y = 1$ (i.e., class 1), where $\mathbf{\Sigma} = (\sigma_{ij})_{i,j=1,\cdots,10}$, with $\sigma_{ij} = 2^{|i-j|}$. On the other hand, if $Y = 0$ then $\boldsymbol{\chi}$ has a 10-dim standard Cauchy distribution with independent components, i.e., the components of $\boldsymbol{\chi}$ are iid random variables with the pdf $f(x) = (\pi(1 + x^2))^{-1}$, $-\infty < x < \infty$. As for the choice of the individual classifiers, we have considered the following six classifiers: a 1-Nearest Neighbor (1-NN) classifier, (ii) a 7-NN classifier, (iii) the Support Vector Machine (SVM) of Boser et al. [5], (iv) Breiman's [8] Random Forest, (v) a Gaussian kernel classifier with a bandwidth of $n^{-0.2}$, and (vi) the Linear Discriminant Analysis (LDA). Next we considered three different combined classifiers: The proposed combined classifier $G_{n,\widehat{k}}$ defined in (4), where $\widehat{k}$ is the minimizer of the empirical error in (6), the combined classifier of Balakrishnan and Mojirsheibani [2] given in (9), and the majority-vote classifier. To trace the performance of various methods, several sample sizes were used: $n = 50, 100, 200, \ldots, 800, 900$, with $\frac{n}{2}$ observations from each population. As for the sample splits, we took $m = 0.65n$ and $\ell = n - m$, but any other fraction in the 0.55 to 0.85 range seems to work just as well. Next, for each sample size, we constructed the six individual classifiers based on $\mathbb{T}_m$, which were then used to construct the above combined classifiers based on $\mathbb{T}_\ell$. For each value of n, a total of 25 sample splits were used; this is in view of Remark 4. To assess the performance of various classifiers, we also generated 1000 additional observations, with 500 from each of the two populations; these were used as test samples for each classifier. Finally, the whole process above was repeated a total of 300 times, yielding 300 estimates of the misclassification errors of each classifiers discussed above. The average errors (over 300 Monte Carlo runs) are summarized in Table 1 along with their standard errors in parentheses.

As Table 1 shows, the SVM and random forest classifiers perform very well for smaller sample sizes, but as n reaches 100, the proposed combined classifier $G_{n,\widehat{k}}$ (as well as B-M) start performing better than all other classifiers. The boldface values represent the smallest error rates for each n. We also note that as n gets larger and larger, most classifiers start performing better and better (except for the LDA which is a wrong classifier in the presence of Cauchy populations); however, the proposed combined classifier is consistently superior

Table 1: Misclassification errors of the three combined classifiers and the six individual classifiers for the simulated data of Example A. Here $\widehat{G}_{n,\widehat{k}}$ is the proposed classifier, B-M is the combined classifier of Balakrishnan and Mojirsheibani [2], and Vote is the combined classifier based on majority voting. The boldfaced values represent the lowest errors for each n.

n	$\widehat{G}_{n,\widehat{k}}$	B-M	Vote	SVM	Forest	1-NN	7-NN	Kernel	LDA
50	.3342	.3404	.3732	**.3020**	.3180	.3616	.4698	.3339	.4864
	(.0029)	(.0023)	(.0015)	(.0016)	(.0018)	(.0015)	(.0013)	(.0014)	(.0013)
100	**.2802**	.2828	.3366	.2847	.2875	.3329	.4374	.3126	.4906
	(.0017)	(.0015)	(.0016)	(.0010)	(.0017)	(.0016)	(.0017)	(.0013)	(.0012)
200	**.1709**	.1821	.2646	.2506	.2181	.2696	.3341	.2635	.4849
	(.0014)	(.0013)	(.0010)	(.0009)	(.0011)	(.0011)	(.0011)	(.0010)	(.0013)
300	**.1491**	.1623	.2463	.2428	.2078	.2534	.2897	.2510	.4873
	(.0013)	(.0011)	(.0010)	(.0009)	(.0011)	(.0009)	(.0012)	(.0009)	(.0013)
400	**.1452**	.1514	.2248	.2385	.1909	.2399	.2570	.2387	.4843
	(.0010)	(.0009)	(.0013)	(.0008)	(.0009)	(.0009)	(.0011)	(.0009)	(.0015)
500	**.1394**	.1405	.2105	.2262	.1739	.2275	.2354	.2265	.4945
	(.0009)	(.0008)	(.0009)	(.0012)	(.0011)	(.0010)	(.0010)	(.0010)	(.0013)
600	**.1360**	**.1360**	.2001	.2212	.1676	.2221	.2195	.2213	.4845
	(.0008)	(.0007)	(.0009)	(.0007)	(.0008)	(.0007)	(.0010)	(.0007)	(.0014)
700	**.1310**	.1343	.1961	.2199	.1656	.2230	.2097	.2221	.4891
	(.0010)	(.0008)	(.0007)	(.0010)	(.0009)	(.0009)	(.0009)	(.0009)	(.0012)
800	**.1249**	.1278	.1861	.2119	.1583	.2137	.2066	.2133	.4851
	(.0008)	(.0007)	(.0009)	(.0008)	(.0008)	(.0009)	(.0009)	(.0009)	(.0013)
900	**.1190**	.1243	.1856	.2076	.1505	.2122	.1969	.2118	.4920
	(.0009)	(.0008)	(.0006)	(.0010)	(.0010)	(.0008)	(.0008)	(.0008)	(.0013)

to all the other ones. Furthermore, in some cases this superiority is quite notable; see, for example, the row corresponding to n=300, where the error of $\widehat{G}_{n,\widehat{k}}$ is only 0.1491 as compared to the best individual classifier, the random forest, with an error of 0.2078. Such rather large discrepancies can also be noticed for many other values of n in Table 1.

Another feature of the proposed classifier $\widehat{G}_{n,\widehat{k}}$ is that, unlike linear combined classifiers, it can even be used to improve the predictive performance of any single classifier (although the improvement may be quite incremental in some cases). To appreciate this, observe that according to Theorem 1, the proposed combined classifier $\widehat{G}_{n,\widehat{k}}$ can asymptotically outperform each constituent classifier $\widehat{g}_{n,1}, \ldots, \widehat{g}_{n,J}$, where $J \geq 1$. Now, taking J=1, this theorem states that, given a single classifier $\widehat{g}_n$, the proposed approach can produce an improved version of $\widehat{g}_n$. To put this to the test, we applied $\widehat{G}_{n,\widehat{k}}$ to some of the six

classifiers in Table 1; in fact, we applied it to each of the two classifiers that have already performed quite well and are difficult to outperform, i.e., the SVM and random forest classifiers. The results appear in Table 2. This table shows that as n increases, the

Table 2: Effects of applying the combined classifier $\widehat{G}_{n,\widehat{k}}$ to J=1 classifier only. Here $\widehat{\text{SVM}}$ and $\widehat{\text{Forest}}$ represent the classifiers obtained by applying $\widehat{G}_{n,\widehat{k}}$ to SVM and random forest, respectively. For each n, the boldfaced values represent the smaller of the two error rates when comparing SVM and $\widehat{\text{SVM}}$ in columns 3 and 4, and when comparing random forest and $\widehat{\text{Forest}}$ in columns 4 and 5.

n	SVM	$\widehat{\text{SVM}}$	Forest	$\widehat{\text{Forest}}$
50	**0.3020**	0.3076	**0.3180**	0.3193
	(.0016)	(.0017)	(.0018)	(.0024)
100	**0.2847**	0.2871	0.2875	**0.2844**
	(.0010)	(.0011)	(.0017)	(.0016)
200	0.2506	**0.2411**	0.2181	**0.2032**
	(.0009)	(.0008)	(.0011)	(.0013)
300	0.2428	**0.2380**	0.2078	**0.1993**
	(.0009)	(.0009)	(.0011)	(.0008)
400	0.2385	**0.2302**	0.1909	**0.1815**
	(.0009)	(.0007)	(.0009)	(.0009)
500	**0.2262**	0.2265	0.1739	**0.1732**
	(.0012)	(.0010)	(.0011)	(.0010)
600	0.2212	**0.2207**	0.1676	**0.1768**
	(.0007)	(.0008)	(.0008)	(.0008)
700	0.2199	**0.2188**	0.1656	**0.1643**
	(.0010)	(.0008)	(.0009)	(.0009)
800	0.2119	**0.2031**	0.1583	**0.1492**
	(.0008)	(.0007)	(.0008)	(.0008)
900	0.2076	**0.2070**	0.1505	**0.1480**
	(.0010)	(.0007)	(.0010)	(.0009)

proposed classifier can still improve upon the performance of each of these two classifiers, individually, despite the fact that both SVM and random forest are well known to be superb classifiers. It is true that the improvement is rather minimal, but the main message here is that $\widehat{G}_{n,\widehat{k}}$ is more than just a combined classifier, it can also improve the predictive power of a single classifier.

Example B *(Wisconsin Breast Cancer Data)*.

This real data set has 683 fully observed instances, 444 of which have been labeled *benign*,

which is class 1, and the rest are *malignant*, i.e., class 0. There are also 9 numerical covariates associated with each instance. A full description of this data set is available from the UCI Machine Learning Repository of data sets: `https://archive.ics.uci.edu/ml/datasets.php`. Also, see Wolberg and Mangasarian (1990).

To carry out the analysis, 500 of the 683 instances were randomly selected to be used as the training data, whereas the remaining 183 were set aside as the test sequence to be used to estimate the error rates of different classifiers. To study the performance of various classifiers as a function of the sample size n, we considered 5 different sample sizes $n = 100, \ldots, 500$ (since n can only go up to 500 here) and, for each value of n, six individual classifiers were constructed which were then used to construct the proposed combined classifier. Here, as in Example A, the sample splits were taken to be $m = 0.65n$ and $\ell = n - m$. Finally, the error rates of all classifiers were estimated using the test sequence of 183 instances that were set aside. This whole process was repeated 100 times. Table 3 reports the average error rates of various classifiers over 100 runs; the standard errors appear in parentheses. The boldfaced values represent the smallest error rates for each n. The combined classifier B-M of Table 1 is not included here for the simple reason that it is always inferior to $\widehat{G}_{n,\widehat{k}}$. As Table 3 shows, the proposed combined classifier can outperform the individual classifiers. This can be noticed by comparing the error of the best individual classifier (the random forest in this case) with that of $\widehat{G}_{n,\widehat{k}}$ that appears in the first column.

Example C *(German Credit Data)*.

Here we consider a real data set consisting of 1000 individuals, 700 of whom have been labeled as having "good credit", i.e., class 1, whereas the remaining 300 have "bad credit", which is class 0. There are 24 numerical covariates associated with each person. A full description of this data set is available from the UCI repository of machine learning data sets at `https://archive.ics.uci.edu/ml/index.php`.

Table 3: Misclassification errors of various classifiers for the *Wisconsin Breast Cancer* data of Example B. Here $\widehat{G}_{n,\widehat{k}}$ is our proposed classifier and Vote is the combined classifier based on majority voting. The boldfaced values represent the lowest errors for each n.

n	$\widehat{G}_{n,\widehat{k}}$	Vote	SVM	Forest	1-NN	15-NN	Kernel	LDA
100	**.0362**	.0387	.0541	.0387	.0469	.0449	.0459	.0453
	(.0013)	(.0012)	(.0017)	(.0014)	(.0016)	(.0012)	(.0016)	(.0016)
200	**.0312**	.0347	.0511	.0337	.0427	.0384	.0422	.0392
	(.0012)	(.0013)	(.0018)	(.0014)	(.0015)	(.0014)	(.0016)	(.0013)
300	**.0279**	.0304	.0439	.0303	.0396	.0346	.0391	.0375
	(.0011)	(.0011)	(.0015)	(.0012)	(.0012)	(.00012)	(.0011)	(.0012)
400	**.0277**	.0316	.0445	.0298	.0408	.0338	.0409	.0385
	(.0010)	(.0012)	(.0015)	(.0010)	(.0014)	(.0012)	(.0013)	(.0012)
500	**.0272**	.0327	.0422	.0295	.0416	.0336	.0416	.0395
	(.0010)	(.0011)	(.0014)	(.0010)	(.0012)	(.0012)	(.0012)	(.0012)

To carry out the analysis, first we randomly selected, and set aside, 300 of the 1000 observations to be used as the test sequence to estimate the error rates of various classifiers. As for the training sample size, seven values were considered: $n = 100, 200, \ldots, 700$, (since n cannot go beyond 700=1000−300). This grid of values of n allows us to somewhat monitor the performance of different classifiers as n increases. Then, given a sample of size n, each of the six individual classifiers of Example B were constructed and used to construct the proposed combined classifiers. Here, once again, the sample splits were taken to be $m = 0.65n$ and $\ell = n - m$. Finally, the error rates of various classifiers were estimated using the test sample of 300 observations. The entire process above was repeated 100 times and the average misclassification error rates were calculated. The results are summarized in Table 4. As this table shows, the combined classifier $\widehat{G}_{n,\widehat{k}}$ has the ability to perform well and, in fact, slightly outperform the best individual classifier (which is random forest) as the sample size increases to about 600.

4 Proofs

PROOF OF THEOREM 1

Let G_n^* and $G_{n,k}$ be as in (3) and (4), respectively. Similarly, let G_m^* be as in (3), but

Table 4: Misclassification errors of different classifiers for the *German Credit* data of Example C. Here $\widehat{G}_{n,\widehat{k}}$ is our proposed classifier and Vote is the combined classifier based on majority voting. The boldfaced values represent the lowest errors for each n.

n	$\widehat{G}_{n,\widehat{k}}$	Vote	SVM	Forest	1-NN	15-NN	Kernel	LDA
100	.2914	.3012	.3053	**.2755**	.3746	.3061	.3718	.3150
	(.0028)	(.0024)	(.0022)	(.0024)	(.0032)	(.0023)	(.0032)	(.0033)
200	.2691	.2922	.2975	**.2613**	.3590	.3030	.3572	.2974
	(.0026)	(.0026)	(.0022)	(.0023)	(.0025)	(.0023)	(.0025)	(.0030)
300	.2608	.2838	.2954	**.2523**	.3534	.2984	.3521	.2893
	(.0024)	(.0022)	(.0024)	(.0020)	(.0026)	(.00023)	(.0027)	(.0022)
400	.2472	.2794	.2909	**.2458**	.3494	.2973	.3467	.2860
	(.0024)	(.0029)	(.0023)	(.0027)	(.0027)	(.0024)	(.0026)	(.0023)
500	.2438	.2738	.2886	**.2414**	.3417	.2955	.3399	.2775
	(.0021)	(.0024)	(.0024)	(.0021)	(.0023)	(.0023)	(.0024)	(.0022)
600	**.2398**	.2749	.2914	.2411	.3424	.2994	.3395	.2774
	(.0023)	(.0022)	(.0023)	(.0022)	(.0023)	(.0021)	(.0024)	(.0024)
700	**.2372**	.2756	.2869	.2376	.3407	.2990	.3385	.2800
	(.0021)	(.0022)	(.0026)	(.0021)	(.0023)	(.0022)	(.0022)	(.0024)

with n replaced with m everywhere in (3). Also, let $G_{n,\widehat{k}}$ be the combined classifier given by (4) and (6), and define the quantities

$$L(G_n^*) = \mathbb{P}\Big\{G_n^*\big(\widehat{g}_{n,1}(\boldsymbol{\chi}),\cdots,\widehat{g}_{n,J}(\boldsymbol{\chi})\big)\neq Y\Big\} \tag{10}$$

$$L(G_m^*) = \mathbb{P}\Big\{G_m^*\big(\widehat{g}_{m,1}(\boldsymbol{\chi}),\cdots,\widehat{g}_{m,J}(\boldsymbol{\chi})\big)\neq Y\Big\} \tag{11}$$

$$\widehat{L}_\ell(G_{n,k}) = \frac{1}{\ell}\sum_{i:\,(\boldsymbol{\chi}_i,Y_i)\,\in\,\mathbb{T}_\ell}\mathbb{I}_{\left\{G_{n,k}\left(\widehat{g}_{m,1}(\boldsymbol{\chi}_i),\cdots,\widehat{g}_{m,J}(\boldsymbol{\chi}_i)\right)\neq Y_i\right\}} \tag{12}$$

$$L(G_{n,k}|\mathbb{T}_n) = \mathbb{P}\Big\{G_{n,k}\big(\widehat{g}_{m,1}(\boldsymbol{\chi}),\cdots,\widehat{g}_{m,J}(\boldsymbol{\chi})\big)\neq Y\,\Big|\,\mathbb{T}_n\Big\} \tag{13}$$

$$L(G_{n,\widehat{k}}) = \mathbb{P}\Big\{G_{n,\widehat{k}}\big(\widehat{g}_{m,1}(\boldsymbol{\chi}),\cdots,\widehat{g}_{m,J}(\boldsymbol{\chi})\big)\neq Y\Big\} \tag{14}$$

$$L(G_{n,\widehat{k}}|\mathbb{T}_n) = \mathbb{P}\Big\{G_{n,\widehat{k}}\big(\widehat{g}_{m,1}(\boldsymbol{\chi}),\cdots,\widehat{g}_{m,J}(\boldsymbol{\chi})\big)\neq Y\,\Big|\,\mathbb{T}_n\Big\}, \tag{15}$$

where $\widehat{k}$ is as in (6) and $1\leq k\leq \ell$. Therefore, in view of Proposition 2, and the notation in (10) and (14), one must show $L(G_{n,\widehat{k}})-L(G_n^*)\to 0$, as $n\to\infty$. Now, start by writing

$$\begin{aligned} L(G_{n,\widehat{k}})-L(G_n^*) &= \mathbb{E}\Big[L(G_{n,\widehat{k}}|\mathbb{T}_n)-\min_{1\leq k\leq \ell}L(G_{n,k}|\mathbb{T}_n)\Big] \\ &\quad+\Big[\mathbb{E}\min_{1\leq k\leq \ell}L(G_{n,k}|\mathbb{T}_n)-L(G_m^*)\Big]+\big[L(G_n^*)-L(G_m^*)\big] \end{aligned}$$

$$:= R_{n,1} + R_{n,2} + R_{n,3}. \tag{16}$$

Then, fix $\mathbb{T}_n$ and define the classes of sets $\mathscr{A}_{n,1} = \{A_{n,1,1}, \ldots, A_{n,1,\ell}\}$ and $\mathscr{A}_{n,0} = \{A_{n,0,1}, \ldots, A_{n,0,\ell}\}$ where, for $1 \leq k \leq \ell$,

$$A_{n,1,k} = \Big\{ \big(\widehat{g}_{m,1}(\chi), \cdots, \widehat{g}_{m,J}(\chi)\big), \text{as } \chi \text{ varies over } \mathcal{F} \,\Big|\, G_{n,k}\big(\widehat{g}_{m,1}(\chi), \cdots, \widehat{g}_{m,J}(\chi)\big) = 1 \Big\} \times \{0\}$$
$$A_{n,0,k} = \Big\{ \big(\widehat{g}_{m,1}(\chi), \cdots, \widehat{g}_{m,J}(\chi)\big), \text{as } \chi \text{ varies over } \mathcal{F} \,\Big|\, G_{n,k}\big(\widehat{g}_{m,1}(\chi), \cdots, \widehat{g}_{m,J}(\chi)\big) = 0 \Big\} \times \{1\}$$

Furthermore, for $1 \leq k \leq \ell$, let

$$\begin{aligned}
\lambda(A_{n,b,k}|\mathbb{T}_n) &= \mathbb{P}\Big\{ \Big(\widehat{g}_{m,1}(\boldsymbol{\chi}), \cdots, \widehat{g}_{m,J}(\boldsymbol{\chi}), Y\Big) \in A_{n,b,k} \,\Big|\, \mathbb{T}_n \Big\}, \quad b = 0, 1 \\
\widehat{\lambda}_\ell(A_{n,b,k}) &= \ell^{-1} \sum_{i:\,(\boldsymbol{\chi}_i, Y_i) \in \mathbb{T}_\ell} \mathbb{I}_{\left\{ \left(\widehat{g}_{m,1}(\boldsymbol{\chi}_i), \cdots, \widehat{g}_{m,J}(\boldsymbol{\chi}_i), Y_i\right) \in A_{n,b,k} \right\}}, \quad b = 0, 1,
\end{aligned}$$

and observe that

$$\begin{aligned}
& L(G_{n,\widehat{k}}|\mathbb{T}_n) - \min_{1\leq k\leq \ell} L(G_{n,k}|\mathbb{T}_n) \\
&= L(G_{n,\widehat{k}}|\mathbb{T}_n) - \widehat{L}_\ell(G_{n,\widehat{k}}) + \widehat{L}_\ell(G_{n,\widehat{k}}) - \min_{1\leq k\leq \ell} L(G_{n,k}|\mathbb{T}_n) \\
&\leq 2 \max_{1\leq k\leq \ell} \Big| \widehat{L}_\ell(G_{n,k}) - L(G_{n,k}|\mathbb{T}_n) \Big| \\
&\leq 2 \sum_{b=0,1} \max_{1\leq k\leq \ell} \Big| \widehat{\lambda}_\ell(A_{n,b,k}) - \lambda(A_{n,b,k}|\mathbb{T}_n) \Big| \\
&\leq 4 \sup_{B\in\mathscr{B}} \Big| \widehat{\lambda}_\ell(B) - \lambda(B|\mathbb{T}_m) \Big| \\
&\quad \text{(where } \mathscr{B} \text{ is the collection of the Borel sets of } \mathbb{R}^{J+1}\text{)} \\
&\leq 4 \sum_{\mathbf{z}\in\{0,1\}^{J+1}} \Big| \widehat{\lambda}_\ell(\{\mathbf{z}\}) - \lambda(\{\mathbf{z}\}|\mathbb{T}_m) \Big|,
\end{aligned} \tag{17}$$

where, for $\mathbf{z} \in \mathbb{R}^{J+1}$, we have $\widehat{\lambda}_\ell(\{\mathbf{z}\}) = \ell^{-1} \sum_{i:\,(\boldsymbol{\chi}_i, Y_i) \in \mathbb{T}_\ell} \mathbb{I}_{\left\{ \left(\widehat{g}_{m,1}(\boldsymbol{\chi}_i), \cdots, \widehat{g}_{m,J}(\boldsymbol{\chi}_i), Y_i\right) = \mathbf{z} \right\}}$ and $\lambda(\{\mathbf{z}\}|\mathbb{T}_m) = \mathbb{P}\left\{ \left(\widehat{g}_{m,1}(\boldsymbol{\chi}), \cdots, \widehat{g}_{m,J}(\boldsymbol{\chi}), Y\right) = \mathbf{z} \,\middle|\, \mathbb{T}_m \right\}$. Therefore, in view of (17), for every $\epsilon > 0$, one has

$$\begin{aligned}
& \mathbb{P}\Big\{ L(G_{n,\widehat{k}}|\mathbb{T}_n) - \min_{1\leq k\leq \ell} L(G_{n,k}|\mathbb{T}_n) > \epsilon \Big\} \\
&\leq \mathbb{P}\left\{ \sum_{\mathbf{z}\in\{0,1\}^{J+1}} \Big| \widehat{\lambda}_\ell(\{\mathbf{z}\}) - \lambda(\{\mathbf{z}\}|\mathbb{T}_m) \Big| > \epsilon/4 \right\}
\end{aligned}$$

$$\leq \sum_{\mathbf{z}\in\{0,1\}^{J+1}} \mathbb{E}\left[\mathbb{P}\left\{\left|\widehat{\lambda}_\ell(\{\mathbf{z}\}) - \lambda(\{\mathbf{z}\}|\mathbb{T}_m)\right| > 2^{-(J+3)}\,\epsilon \,\middle|\, \mathbb{T}_m\right\}\right]$$

$$\leq 2^{J+1}\left[2\,e^{-C\,\ell\epsilon^2}\right], \quad \text{with } C = 2^{-(2J+5)}, \tag{18}$$

where (18) follows from an application of Hoeffding's (1963) inequality in view of the fact that, conditional on $\mathbb{T}_m$, the term $\widehat{\lambda}_\ell(\{\mathbf{z}\})$ is the average of ℓ independent indicator functions, $\mathbb{I}_{\left\{\left(\widehat{g}_{m,1}(\boldsymbol{\chi}_i),\cdots,\widehat{g}_{m,J}(\boldsymbol{\chi}_i),Y_i\right)=\mathbf{z}\right\}}$, corresponding to the ℓ pairs $(\boldsymbol{\chi}_i, Y_i) \in \mathbb{T}_\ell$. Now, (18) together with the Borel-Cantelli lemma yields $L(G_{n,\widehat{k}}\big|\mathbb{T}_n) - \min_{1\leq k\leq \ell} L(G_{n,k}\big|\mathbb{T}_n) \to_{a.s.} 0$, as $n \to \infty$. Thus, by Lebesgue's dominated convergence theorem,

$$R_{n,1} := \mathbb{E}\Big[L(G_{n,\widehat{k}}\big|\mathbb{T}_n) - \min_{1\leq k\leq \ell} L(G_{n,k}\big|\mathbb{T}_n)\Big] \to 0, \quad \text{as } n \to \infty.$$

To deal with the term $R_{n,2}$ in (16), start by defining the quantity

$$\nu_n(\boldsymbol{\chi}) \;=\; \mathcal{S}_{m,\ell}(\boldsymbol{\chi}) + \ell \cdot \mathbb{I}_{\left\{\mathcal{S}_{m,\ell}(\boldsymbol{\chi})=0\right\}}, \tag{19}$$

where $\mathcal{S}_{m,\ell}(\boldsymbol{\chi})$ is as in (7). Here, we note that $\nu_n(\boldsymbol{\chi})$ takes values in $\{1,\ldots,\ell\}$. Also put

$$\widehat{\mathcal{P}}_{m,\ell}(\boldsymbol{\chi}) = \frac{\sum_{i:\,\boldsymbol{\chi}_i\in\mathbb{T}_\ell} Y_i\,\mathbb{I}_{\left\{\left(\widehat{g}_{m,1}(\boldsymbol{\chi}_i),\cdots,\widehat{g}_{m,J}(\boldsymbol{\chi}_i)\right)=\left(\widehat{g}_{m,1}(\boldsymbol{\chi}),\cdots,\widehat{g}_{m,J}(\boldsymbol{\chi})\right)\right\}}}{\mathcal{S}_{m,\ell}(\boldsymbol{\chi})} \tag{20}$$

and observe that upon replacing k by $\nu_n(\boldsymbol{\chi})$ in (4), we can write

$$G_{n,\,\nu_n(\boldsymbol{\chi})}\big(\widehat{g}_{m,1}(\boldsymbol{\chi}),\cdots,\widehat{g}_{m,J}(\boldsymbol{\chi})\big)$$

$$= \begin{cases} 1 & \text{if } \frac{1}{\nu_n(\boldsymbol{\chi})}\sum_{i:\,(\boldsymbol{\chi}_i,Y_i)\in\mathbb{T}_\ell}(2Y_i-1)\cdot\mathcal{I}_m\big(\nu_n(\boldsymbol{\chi}),\boldsymbol{\chi},\boldsymbol{\chi}_i\big) > 0 \\ 0 & \text{otherwise,} \end{cases}$$

(where the term $\mathcal{I}_m\big(\nu_n(\boldsymbol{\chi}),\boldsymbol{\chi},\boldsymbol{\chi}_i\big)$ is defined via (5))

$$= \begin{cases} 1 & \text{if } \frac{\mathbb{I}_{\{\mathcal{S}_{m,\ell}(\boldsymbol{\chi})>0\}}}{\mathcal{S}_{m,\ell}(\boldsymbol{\chi})}\sum_{i:\,(\boldsymbol{\chi}_i,Y_i)\in\mathbb{T}_\ell}(2Y_i-1)\cdot\mathcal{I}_m\big(\mathcal{S}_{m,\ell}(\boldsymbol{\chi}),\boldsymbol{\chi},\boldsymbol{\chi}_i\big) \\ & \qquad + \frac{\mathbb{I}_{\{\mathcal{S}_{m,\ell}(\boldsymbol{\chi})=0\}}}{\ell}\sum_{i:\,(\boldsymbol{\chi}_i,Y_i)\in\mathbb{T}_\ell}(2Y_i-1)\cdot\mathcal{I}_m(\ell,\boldsymbol{\chi},\boldsymbol{\chi}_i) > 0 \\ 0 & \text{otherwise,} \end{cases} \tag{21}$$

(where (21) follows from the definition of $\nu_n(\boldsymbol{\chi})$ in (19))

$$= \begin{cases} 1 & \text{if } \frac{\mathbb{I}_{\{\mathcal{S}_{m,\ell}(\boldsymbol{\chi})>0\}}}{\mathcal{S}_{m,\ell}(\boldsymbol{\chi})}\sum_{i:\,(\boldsymbol{\chi}_i,Y_i)\in\mathbb{T}_\ell}(2Y_i-1)\,\mathbb{I}_{\left\{\left(\widehat{g}_{m,1}(\boldsymbol{\chi}_i),\cdots,\widehat{g}_{m,J}(\boldsymbol{\chi}_i)\right)=\left(\widehat{g}_{m,1}(\boldsymbol{\chi}),\cdots,\widehat{g}_{m,J}(\boldsymbol{\chi})\right)\right\}} \\ & \qquad + (2\overline{Y}_\ell-1)\,\mathbb{I}_{\{\mathcal{S}_{m,\ell}(\boldsymbol{\chi})=0\}} > 0 \\ 0 & \text{otherwise,} \end{cases} \tag{22}$$

$$= \begin{cases} 1 & \text{if } \widehat{\mathcal{P}}_{m,\ell}(\boldsymbol{\chi}) \cdot \mathbb{I}_{\{\mathcal{S}_{m,\ell}(\boldsymbol{\chi})>0\}} + \overline{Y}_\ell \cdot \mathbb{I}_{\{\mathcal{S}_{m,\ell}(\boldsymbol{\chi})=0\}} > 1/2 \\ 0 & \text{otherwise,} \end{cases} \tag{23}$$

where $\widehat{\mathcal{P}}_{m,\ell}(\boldsymbol{\chi})$ is as in (20) and $\overline{Y}_\ell = \ell^{-1} \sum_{i:\, Y_i \in \mathbb{T}_\ell} Y_i$. Here, (22) follows from (21) because of the following simple facts:

(i) For each $\boldsymbol{\chi}_i \in \mathbb{T}_\ell$, the product of the two indicator functions, $\mathbb{I}_{\left\{\mathcal{S}_{m,\ell}(\boldsymbol{\chi})>0\right\}}$ and

$\mathbb{I}_{\left\{(\widehat{g}_{m,1}(\boldsymbol{\chi}_i), \cdots, \widehat{g}_{m,J}(\boldsymbol{\chi}_i)) \text{ is among the } \mathcal{S}_{m,\ell}(\boldsymbol{\chi}) \text{ neighbors of } (\widehat{g}_{m,1}(\boldsymbol{\chi}), \cdots, \widehat{g}_{m,J}(\boldsymbol{\chi}))\right\}}$,

(where the latter indicator function is just the term $\mathcal{I}_m\big(\mathcal{S}_{m,\ell}(\boldsymbol{\chi}), \boldsymbol{\chi}, \boldsymbol{\chi}_i\big)$ in (21)), will be equal to 1 if and only if $\mathbb{I}_{\left\{(\widehat{g}_{m,1}(\boldsymbol{\chi}_i), \cdots, \widehat{g}_{m,J}(\boldsymbol{\chi}_i)) = (\widehat{g}_{m,1}(\boldsymbol{\chi}), \cdots, \widehat{g}_{m,J}(\boldsymbol{\chi}))\right\}} = 1$.

(ii) For each $\boldsymbol{\chi}_i \in \mathbb{T}_\ell$, the term $\mathcal{I}_m(\ell, \boldsymbol{\chi}, \boldsymbol{\chi}_i)$ in (21), which is just the indicator function

$\mathbb{I}_{\left\{(\widehat{g}_{m,1}(\boldsymbol{\chi}_i), \cdots, \widehat{g}_{m,J}(\boldsymbol{\chi}_i)) \text{ is among the } \ell \text{ nearest neighbors of } (\widehat{g}_{m,1}(\boldsymbol{\chi}), \cdots, \widehat{g}_{m,J}(\boldsymbol{\chi}))\right\}}$,

is always equal to 1; this is because here the number of nearest neighbors, ℓ, is the same as the entire sample size, ℓ, (the size of $\mathbb{T}_\ell$).

To complete the proof, we also need the following lemma which puts bounds on the term $R_{n,2}$ in (16) based on the expected value of the expression that appears on the right side of (23). More specifically,

Lemma 1 *Let $R_{n,2}$ be as in (16) and put $\overline{Y}_\ell = \ell^{-1} \sum_{i:\, Y_i \in \mathbb{T}_\ell} Y_i$. Then*

$$0 \;\le\; R_{n,2} \;\;\le\;\; 2\,\mathbb{E}\left|\widehat{\mathcal{P}}_{m,\ell}(\boldsymbol{\chi}) \cdot \mathbb{I}_{\{\mathcal{S}_{m,\ell}(\boldsymbol{\chi})>0\}} + \overline{Y}_\ell \cdot \mathbb{I}_{\{\mathcal{S}_{m,\ell}(\boldsymbol{\chi})=0\}} \;-\; \mathcal{P}_m(\boldsymbol{\chi})\right|,$$

where $\mathcal{P}_m(\boldsymbol{\chi}) = \mathbb{P}\left\{Y = 1 \middle| \, \widehat{g}_{m,1}(\boldsymbol{\chi}), \cdots, \widehat{g}_{m,J}(\boldsymbol{\chi})\right\}$.

Now, to show $R_{n,2} \to 0$, let $\mathcal{P}_m(\boldsymbol{\chi})$ be as in Lemma 1 and observe that by Lemma 1 and the fact that $0 \le \overline{Y}_\ell \le 1$, one has

$$\begin{aligned} R_{n,2} \;\;&\le\;\; 2\,\mathbb{E}\left|\widehat{\mathcal{P}}_{m,\ell}(\boldsymbol{\chi}) \cdot \mathbb{I}_{\{\mathcal{S}_{m,\ell}(\boldsymbol{\chi})>0\}} \;-\; \mathcal{P}_m(\boldsymbol{\chi})\right| + 2\,\mathbb{E}\Big(\mathbb{I}_{\{\mathcal{S}_{m,\ell}(\boldsymbol{\chi})=0\}}\Big) \\ &\le\;\; 2\sqrt{\mathbb{E}\left[\mathbb{E}\left(\left|\widehat{\mathcal{P}}_{m,\ell}(\boldsymbol{\chi}) \cdot \mathbb{I}_{\{\mathcal{S}_{m,\ell}(\boldsymbol{\chi})>0\}} \;-\; \mathcal{P}_m(\boldsymbol{\chi})\right|^2 \middle| \, \mathbb{T}_m,\, \boldsymbol{\chi},\, \{\boldsymbol{\chi}_i\}_i \in \mathbb{T}_\ell\right)\right]} \end{aligned} \tag{24}$$

$$+\ 2\,\mathbb{P}\left\{\mathcal{S}_{m,\ell}(\boldsymbol{\chi})=0\right\}.$$

However, conditional on $\mathbb{T}_m$, $\boldsymbol{\chi}$, and $\{\boldsymbol{\chi}_i\}_i \in \mathbb{T}_\ell$, the random variable $\mathcal{S}_{m,\ell}(\boldsymbol{\chi})\cdot\widehat{\mathcal{P}}_{m,\ell}(\boldsymbol{\chi})$, which is equal to the numerator of the right side of (20), has the binomial distribution $\mathsf{Bin}\big(\mathcal{S}_{m,\ell}(\boldsymbol{\chi})\,,\,\mathcal{P}_m(\boldsymbol{\chi})\big)$ whenever $\mathcal{S}_{m,\ell}(\boldsymbol{\chi})>0$. Therefore, the expression under the square-root sign in (24) can be bounded as follows

$$\begin{aligned}
&\mathbb{E}\left[\mathbb{E}\left(\left|\widehat{\mathcal{P}}_{m,\ell}(\boldsymbol{\chi})\cdot\mathbb{I}_{\{\mathcal{S}_{m,\ell}(\boldsymbol{\chi})>0\}}\ -\ \mathcal{P}_m(\boldsymbol{\chi})\right|^2\,\middle|\,\mathbb{T}_m,\,\boldsymbol{\chi},\,\{\boldsymbol{\chi}_i\}_i\in\mathbb{T}_\ell\right)\right]\\
&\quad\le\ \mathbb{E}\left[\mathbb{E}\left(\left|\frac{\mathcal{S}_{m,\ell}(\boldsymbol{\chi})\cdot\widehat{\mathcal{P}}_{m,\ell}(\boldsymbol{\chi})\cdot\mathbb{I}_{\{\mathcal{S}_{m,\ell}(\boldsymbol{\chi})>0\}}}{\mathcal{S}_{m,\ell}(\boldsymbol{\chi})}\ -\ \mathcal{P}_m(\boldsymbol{\chi})\right|^2\right.\right.\\
&\qquad\qquad\left.\left.\times\,\mathbb{I}_{\{\mathcal{S}_{m,\ell}(\boldsymbol{\chi})>0\}}\,\middle|\,\mathbb{T}_m,\,\boldsymbol{\chi},\,\{\boldsymbol{\chi}_i\}_i\in\mathbb{T}_\ell\right)+\mathbb{I}_{\left\{\mathcal{S}_{m,\ell}(\boldsymbol{\chi})=0\right\}}\right]\\
&\quad=\ \mathbb{E}\left[\mathbb{E}\left(\left|\frac{\mathsf{Bin}\Big(\mathcal{S}_{m,\ell}(\boldsymbol{\chi}),\,\mathcal{P}_m(\boldsymbol{\chi})\Big)}{\mathcal{S}_{m,\ell}(\boldsymbol{\chi})}\ -\ \mathcal{P}_m(\boldsymbol{\chi})\right|^2\right.\right.\\
&\qquad\qquad\left.\left.\times\,\mathbb{I}_{\left\{\mathcal{S}_{m,\ell}(\boldsymbol{\chi})>0\right\}}\,\middle|\,\mathbb{T}_m,\,\boldsymbol{\chi},\,\{\boldsymbol{\chi}_i\}_i\in\mathbb{T}_\ell\right)+\mathbb{I}_{\left\{\mathcal{S}_{m,\ell}(\boldsymbol{\chi})=0\right\}}\right]\\
&\quad=\ \mathbb{E}\left[\frac{\mathcal{P}_m(\boldsymbol{\chi})\big(1-\mathcal{P}_m(\boldsymbol{\chi})\big)}{\mathcal{S}_{m,\ell}(\boldsymbol{\chi})}\times\mathbb{I}_{\left\{\mathcal{S}_{m,\ell}(\boldsymbol{\chi})>0\right\}}\right]+\mathbb{P}\left\{\mathcal{S}_{m,\ell}(\boldsymbol{\chi})=0\right\},
\end{aligned}$$

where the last line follows from the usual binomia variance formula. This last expression, in conjunction with (24), and the fact that $\mathcal{P}_m(\boldsymbol{\chi})\big(1-\mathcal{P}_m(\boldsymbol{\chi})\big)\le 1/4$ immediately yields

$$R_{n,2}\ \ \le\ \ 2\sqrt{\mathbb{E}\left[\big(4\,\mathcal{S}_{m,\ell}(\boldsymbol{\chi})\big)^{-1}\,\mathbb{I}_{\left\{\mathcal{S}_{m,\ell}(\boldsymbol{\chi})>0\right\}}\right]+\mathbb{P}\left\{\mathcal{S}_{m,\ell}(\boldsymbol{\chi})=0\right\}}\ +\ 2\,\mathbb{P}\left\{\mathcal{S}_{m,\ell}(\boldsymbol{\chi})=0\right\}. \tag{25}$$

But, upon replacing $\{\mathcal{S}_{m,\ell}(\boldsymbol{\chi})>0\}$ by $\{1\le\mathcal{S}_{m,\ell}(\boldsymbol{\chi})\le k\}\cup\{\mathcal{S}_{m,\ell}(\boldsymbol{\chi})>k\}$, for any arbitrary integer $k\ge 1$, one finds $\mathbb{E}\big[\frac{1}{\mathcal{S}_{m,\ell}(\boldsymbol{\chi})}\,\mathbb{I}_{\{\mathcal{S}_{m,\ell}(\boldsymbol{\chi})>0\}}\big]\le\mathbb{P}\{1\le\mathcal{S}_{m,\ell}(\boldsymbol{\chi})\le k\}+k^{-1}$ holds for all $k\ge 1$. Therefore, by first choosing k large enough and then applying Assumption A, the bound in (25) can be made as small as desired. The proof of Theorem 1 now follows since, in view of Assumption B, we have $R_{n,3}\to 0$, as n (and thus m) $\to\infty$, where $R_{n,3}$ is as in (16). This completes the proof of Theorem 1.

□

PROOF OF LEMMA 1

Let $L(G_m^*)$ be as defined in (11). Also, let $G_{n,\,\nu_n(\boldsymbol{\chi})}$ be as given by the right side of (23) and put $L(G_{n,\,\nu_n(\boldsymbol{\chi})}\big|\mathbb{T}_n) = \mathbb{P}\big\{G_{n,\,\nu_n(\boldsymbol{\chi})}\big(\widehat{g}_{m,1}(\boldsymbol{\chi}),\cdots,\widehat{g}_{m,J}(\boldsymbol{\chi})\big) \neq Y\,\big|\mathbb{T}_n\big\}$. Now, observe that $\mathbb{E}\left[\min_{1\leq k\leq \ell} L(G_{n,k}\big|\mathbb{T}_n)\right] \leq \mathbb{E}\left[L(G_{n,\,\nu_n(\boldsymbol{\chi})}\big|\mathbb{T}_n)\right] = L(G_{n,\,\nu_n(\boldsymbol{\chi})})$, from which one obtains

$$R_{n,2} \;\leq\; L(G_{n,\,\nu_n(\boldsymbol{\chi})}) - L(G_m^*). \tag{26}$$

Furthermore, $R_{n,2} \geq 0$ which follows from the fact that

$$\begin{aligned}
\mathbb{E}\left[\min_{1\leq k\leq \ell} L(G_{n,k}\big|\mathbb{T}_n)\right] &= \mathbb{E}\left[L\left(\operatorname*{argmin}_{G_{n,k}\in\{G_{n,1},\ldots,G_{n,\ell}\}} L\big(G_{n,k}|\mathbb{T}_n\big)\,\middle|\,\mathbb{T}_n\right)\right] \\
&= L\left(\operatorname*{argmin}_{G_{n,k}\in\{G_{n,1},\ldots,G_{n,\ell}\}} L\big(G_{n,k}|\mathbb{T}_n\big)\right) \\
&\geq L(G_m^*),
\end{aligned}$$

where the last line follows from Proposition 2 with n replaced by m. Next, observe that

$$\begin{aligned}
R_{n,2} &\leq L(G_{n,\,\nu_n(\boldsymbol{\chi})}) - L(G_m^*)\,, \quad \text{(by (26))} \\
&= \mathbb{P}\left\{G_m^*\big(\widehat{g}_{m,1}(\boldsymbol{\chi}),\cdots,\widehat{g}_{m,J}(\boldsymbol{\chi})\big) = Y\right\} - \mathbb{P}\big\{G_{n,\nu_n(\boldsymbol{\chi})}\big(\widehat{g}_{m,1}(\boldsymbol{\chi}),\cdots,\widehat{g}_{m,J}(\boldsymbol{\chi})\big) = Y\big\} \\
&= \sum_{k=0,1} \mathbb{E}\Big(\mathbb{I}_{\left\{\left[G_m^*\left(\widehat{g}_{m,1}(\boldsymbol{\chi}),\cdots,\widehat{g}_{m,J}(\boldsymbol{\chi})\right)=k\right]\cap[Y=k]\right\}} - \mathbb{I}_{\left\{\left[G_{n,\nu_n(\boldsymbol{\chi})}\left(\widehat{g}_{m,1}(\boldsymbol{\chi}),\cdots,\widehat{g}_{m,J}(\boldsymbol{\chi})\right)=k\right]\cap[Y=k]\right\}}\Big) \\
&= \sum_{k=0,1} \mathbb{E}\Bigg[\mathbb{E}\Bigg(\mathbb{I}_{\left\{G_m^*\left(\widehat{g}_{m,1}(\boldsymbol{\chi}),\cdots,\widehat{g}_{m,J}(\boldsymbol{\chi})\right)=k\right\}}\cdot\mathbb{I}_{\{Y=k\}} \\
&\qquad - \mathbb{I}_{\left\{G_{n,\nu_n(\boldsymbol{\chi})}\left(\widehat{g}_{m,1}(\boldsymbol{\chi}),\cdots,\widehat{g}_{m,J}(\boldsymbol{\chi})\right)=k\right\}}\cdot\mathbb{I}_{\{Y=k\}}\,\Big|\,\widehat{g}_{m,1}(\boldsymbol{\chi}),\cdots, \\
&\qquad\qquad \widehat{g}_{m,J}(\boldsymbol{\chi}),\left(\widehat{g}_{m,1}(\boldsymbol{\chi}_i),\cdots,\widehat{g}_{m,J}(\boldsymbol{\chi}_i),Y_i\right)_{i:\;\boldsymbol{\chi}_i\in\mathbb{T}_\ell}\Bigg)\Bigg] \\
&= \sum_{k=0,1} \mathbb{E}\Bigg[\left(\mathbb{I}_{\left\{G_m^*\left(\widehat{g}_{m,1}(\boldsymbol{\chi}),\cdots,\widehat{g}_{m,J}(\boldsymbol{\chi})\right)=k\right\}} - \mathbb{I}_{\left\{G_{n,\nu_n(\boldsymbol{\chi})}\left(\widehat{g}_{m,1}(\boldsymbol{\chi}),\cdots,\widehat{g}_{m,J}(\boldsymbol{\chi})\right)=k\right\}}\right) \\
&\qquad\qquad \times \mathbb{E}\Big(\mathbb{I}_{\{Y=k\}}\,\Big|\,\widehat{g}_{m,1}(\boldsymbol{\chi}),\cdots\widehat{g}_{m,J}(\boldsymbol{\chi})\Big)\Bigg] \\
&\qquad \text{(because } Y \text{ is independent of } \widehat{g}_{m,1}(\boldsymbol{\chi}_i),\cdots,\widehat{g}_{m,J}(\boldsymbol{\chi}_i),Y_i\,,\;\; i:\boldsymbol{\chi}_i\in\mathbb{T}_\ell) \\
&= \mathbb{E}\Bigg[\left(\mathbb{I}_{\left\{G_m^*\left(\widehat{g}_{m,1}(\boldsymbol{\chi}),\cdots,\widehat{g}_{m,J}(\boldsymbol{\chi})\right)=1\right\}} - \mathbb{I}_{\left\{G_{n,\nu_n(\boldsymbol{\chi})}\left(\widehat{g}_{m,1}(\boldsymbol{\chi}),\cdots,\widehat{g}_{m,J}(\boldsymbol{\chi})\right)=1\right\}}\right)\cdot\mathcal{P}_m(\boldsymbol{\chi})\Bigg]
\end{aligned}$$

$$
\begin{aligned}
&\quad - \mathbb{E}\Big[\Big(\mathbb{I}_{\left\{G_m^*\left(\widehat{g}_{m,1}(\boldsymbol{\chi}),\cdots,\widehat{g}_{m,J}(\boldsymbol{\chi})\right)=1\right\}} - \mathbb{I}_{\left\{G_{n,\nu_n(\boldsymbol{\chi})}\left(\widehat{g}_{m,1}(\boldsymbol{\chi}),\cdots,\widehat{g}_{m,J}(\boldsymbol{\chi})\right)=1\right\}}\Big)\cdot\Big(1-\mathcal{P}_m(\boldsymbol{\chi})\Big)\Big] \\
&\quad \text{(where, as in Lemma 1, } \mathcal{P}_m(\boldsymbol{\chi}) = \mathbb{P}\big\{Y=1\big|\,\widehat{g}_{m,1}(\boldsymbol{\chi}),\cdots\widehat{g}_{m,J}(\boldsymbol{\chi})\big\}.) \\
&= \mathbb{E}\Big[\Big(\mathbb{I}_{\left\{G_m^*\left(\widehat{g}_{m,1}(\boldsymbol{\chi}),\cdots,\widehat{g}_{m,J}(\boldsymbol{\chi})\right)=1\right\}} - \mathbb{I}_{\left\{G_{n,\nu_n(\boldsymbol{\chi})}\left(\widehat{g}_{m,1}(\boldsymbol{\chi}),\cdots,\widehat{g}_{m,J}(\boldsymbol{\chi})\right)=1\right\}}\Big)\cdot\Big(2\,\mathcal{P}_m(\boldsymbol{\chi})-1\Big)\Big] \\
&= 2\,\mathbb{E}\Big[\mathbb{I}_{\left\{G_m^*\left(\widehat{g}_{m,1}(\boldsymbol{\chi}),\cdots,\widehat{g}_{m,J}(\boldsymbol{\chi})\right)\neq G_{n,\nu_n(\boldsymbol{\chi})}\left(\widehat{g}_{m,1}(\boldsymbol{\chi}),\cdots,\widehat{g}_{m,J}(\boldsymbol{\chi})\right)\right\}}\cdot\Big|\mathcal{P}_m(\boldsymbol{\chi})-0.5\Big|\Big] \\
&\leq 2\,\mathbb{E}\Big|\widehat{\mathcal{P}}_{m,\ell}(\boldsymbol{\chi})\cdot\mathbb{I}_{\left\{\mathcal{S}_{m,\ell}(\boldsymbol{\chi})>0\right\}} + \overline{Y}_\ell\cdot\mathbb{I}_{\left\{\mathcal{S}_{m,\ell}(\boldsymbol{\chi})=0\right\}} - \mathcal{P}_m(\boldsymbol{\chi})\Big|,
\end{aligned}
$$

where the last line follows from the definition of $G_{n,\nu_n(\boldsymbol{\chi})}\big(\widehat{g}_{m,1}(\boldsymbol{\chi}),\cdots,\widehat{g}_{m,J}(\boldsymbol{\chi})\big)$ in the far right side of (23), the definition of $G_m^*\big(\widehat{g}_{m,1}(\boldsymbol{\chi}),\cdots,\widehat{g}_{m,J}(\boldsymbol{\chi})\big)$, and the fact that the inequality $\big|\mathcal{P}_m(\boldsymbol{\chi})-0.5\big| \leq \Big|\widehat{\mathcal{P}}_{m,\ell}(\boldsymbol{\chi})\cdot\mathbb{I}_{\left\{\mathcal{S}_{m,\ell}(\boldsymbol{\chi})>0\right\}} + \overline{Y}_\ell\,\mathbb{I}_{\left\{\mathcal{S}_{m,\ell}(\boldsymbol{\chi})=0\right\}} - \mathcal{P}_m(\boldsymbol{\chi})\Big|$ holds on the set $\Big\{G_m^*\big(\widehat{g}_{m,1}(\boldsymbol{\chi}),\cdots,\widehat{g}_{m,J}(\boldsymbol{\chi})\big) \neq G_{n,\nu_n(\boldsymbol{\chi})}\big(\widehat{g}_{m,1}(\boldsymbol{\chi}),\cdots,\widehat{g}_{m,J}(\boldsymbol{\chi})\big)\Big\}$.

□

PROOF OF PROPOSITION 1

The proof Proposition 1 is similar to (and, in fact, much simpler than) that of Proposition 2, and will not be given.

PROOF OF PROPOSITION 2

The proof of this Proposition is similar to (and easier than) the proof of Lemma 1 and goes as follows. Let $G:\{0,1\}^J \to \{0,1\}$ be any combined classifier with error $L(G) = \mathbb{P}\big\{G\big(\widehat{g}_{n,1}(\boldsymbol{\chi}),\cdots,\widehat{g}_{n,J}(\boldsymbol{\chi})\big)\neq Y\big\}$. Then, with G_n^* as in (3), one has

$$
\begin{aligned}
&L(G) - L(G_n^*) \\
&= \mathbb{P}\Big\{G_n^*\big(\widehat{g}_{n,1}(\boldsymbol{\chi}),\cdots,\widehat{g}_{n,J}(\boldsymbol{\chi})\big)=Y\Big\} - \mathbb{P}\Big\{G\big(\widehat{g}_{n,1}(\boldsymbol{\chi}),\cdots,\widehat{g}_{n,J}(\boldsymbol{\chi})\big)=Y\Big\} \\
&= \sum_{k=0,1}\mathbb{E}\Big[\mathbb{E}\Big(\mathbb{I}_{\left\{G_n^*\left(\widehat{g}_{n,1}(\boldsymbol{\chi}),\cdots,\widehat{g}_{n,J}(\boldsymbol{\chi})\right)=k\right\}}\cdot\mathbb{I}_{\left\{Y=k\right\}} \\
&\qquad\qquad - \mathbb{I}_{\left\{G\left(\widehat{g}_{n,1}(\boldsymbol{\chi}),\cdots,\widehat{g}_{n,J}(\boldsymbol{\chi})\right)=k\right\}}\cdot\mathbb{I}_{\left\{Y=k\right\}}\Big|\,\widehat{g}_{n,1}(\boldsymbol{\chi}),\cdots,\widehat{g}_{n,J}(\boldsymbol{\chi})\Big)\Big] \\
&= \mathbb{E}\Big[\Big(\mathbb{I}_{\left\{G_n^*\left(\widehat{g}_{n,1}(\boldsymbol{\chi}),\cdots,\widehat{g}_{n,J}(\boldsymbol{\chi})\right)=1\right\}} - \mathbb{I}_{\left\{G\left(\widehat{g}_{n,1}(\boldsymbol{\chi}),\cdots,\widehat{g}_{n,J}(\boldsymbol{\chi})\right)=1\right\}}\Big)\cdot\Big(2\,\mathcal{P}_n(\boldsymbol{\chi})-1\Big)\Big]
\end{aligned}
$$

$$= 2\,\mathbb{E}\left[\mathbb{I}_{\left\{G\left(\widehat{g}_{n,1}(\boldsymbol{\chi}),\cdots,\widehat{g}_{n,J}(\boldsymbol{\chi})\right)\neq G_n^*\left(\widehat{g}_{n,1}(\boldsymbol{\chi}),\cdots,\widehat{g}_{n,J}(\boldsymbol{\chi})\right)\right\}} \times \left|\mathcal{P}_n(\boldsymbol{\chi}) - 0.5\right|\right]$$

$$\geq 0\,, \qquad \text{where } \mathcal{P}_n(\boldsymbol{\chi}) = \mathbb{P}\big\{Y = 1\big|\,\widehat{g}_{n,1}(\boldsymbol{\chi}), \cdots \widehat{g}_{n,J}(\boldsymbol{\chi})\big\}.$$

□

Acknowledgements

This work was supported by the NSF under Grant DMS-1916161 of Majid Mojirsheibani.

Disclosure statement

The authors report that there are no potential conflicts of interest.

References

[1] Abraham, C., Biau, G., and Cadre, B. On the kernel rule for functional classification. *Annals of the Institute of Statistical Mathematics*, 2006;58:619-633.

[2] Balakrishnan, N. and Mojirsheibani, M. A simple method for combining estimates to improve the overall error rates in classification. *Computational Statistics*, 2015;30:1033-1049.

[3] Biau, G., Devroye, L. and Lugosi, G. Consistency of random forests and other averaging classifiers. *Journal of Machine Learning Research*, 2008;9:2015-2033.

[4] Biau, G., Fischer, A., Guedj, B. and Malley, J. D. COBRA: A combined regression strategy. *Journal of Multivariate Analysis*, 2016;146:18-28 (with supplementary material).

[5] Boser, B., Guyon, I., and Vapnik, V. (1992) A training algorithm for optimal margin classifiers. In *Proceedings of the Fifth Annual Workshop on Computational Learning Theory,* Vol. 5, pp. 144 - 152; 1992.

[6] Breiman, L. Stacked regression. *Machine Learning*, 1995;24:49-64.

[7] Breiman, L. Bagging predictors. *Machine Learning*, 1996;24:123-140.

[8] Breiman, L. Random Forests. *Machine Learning*, 2001;45:5-32.

[9] Cérou, F. and Guyader, A. Nearest neighbor classification in infinite dimensions. *ESAIM: Probability and Statistics*, 2006;10:340-355.

[10] Cholaquidis, A., Fraiman, R., Kalemkerian, J., and Llop, P. A nonlinear aggregation type classifier. *Journal of Multivariate Analysis*, 146, 269-281.

[11] Devroye, L. and Györfi, L. *Nonparametric density estimation: The L_1 view.* NY: John Wiley & Sons; 1985.

[12] Devroye, L., Györfi, L., and Lugosi, G. *A Probabilistic Theory of Pattern Recognition*, NY: Springer; 1996.

[13] Fischer A. and Mougeot, M. Aggregation using input-output trade-off. *Journal of Statistical Planning and Inference*, 2019;200:1-19.

[14] Giraud, C. *Introduction to High-Dimensional Statistics*, Chapman & Hall/CRC; 2014.

[15] Hoeffding, W. Probability Inequalities for Sums of Bounded Random Variables. *Journal of the American Statistical Association,* 1963;58:13-30.

[16] LeBlanc, M. and Tibshirani, R. Combining estimates in regression and classification. *Journal of the American Statistical Association,* 1996;91:1641-1650.

[17] Lin, Y. and Jeon, Y. Random forests and adaptive nearest neighbors. *Journal of the American Statistical Association*, 2006;101:578-590.

[18] Mojirsheibani, M. Combining classifiers based on discretization. *Journal of the American Statistical Association,* 1999:94:600-609.

[19] Rahman, R., Dhruba, S.R., Ghosh S., and Pal, R. Functional random forest with applications in dose-response predictions. *Scientific Reports*, 2019;9(Article number: 1628).

[20] Rokach, L. Taxonomy for characterizing ensemble methods in classification tasks: a review and annotated bibliography. *Computational Statistics & Data Analysis*,2009;53:4046-4072.

[21] Rokach, L. Ensemble-based classifiers. *Artificial Intelligence Review*, 2010;33:1-39.

[22] Wolberg, W.H. and Mangasarian, O.L. Multisurface method of pattern separation for medical diagnosis applied to breast cytology. In *Proceedings of the National Academy of Sciences*, 1990;87:9193-9196.

[23] Wolpert, D. Stacked Generalization. *Neural Networks*, 1992;5:241-259.

[24] Xu, L., Kryzak, A., and Suen, C.Y. Methods of combining multiple classifiers and their applications to handwriting recognition. *IEEE Transactions on Systems, Man and Cybernetics*, 1992;22:418-435.