A study on majority-voting classifiers with guarantees on the probability of error

A. Carè * M.C. Campi * F.A. Ramponi **
A.T.J.R. Cobbenhagen ***

* Dept. of Information Engineering, University of Brescia, Brescia, Italy (e-mail: {algo.care,marco.campi,federico.ramponi}@unibs.it).
** Dept. of Electronics, Information and Bioengineering, Politecnico di Milano, Milano, Italy (e-mail: simone.garatti@polimi.it).
*** Dept. of Mechanical Engineering, Eindhoven University of Technology, Eindhoven, The Netherlands (e-mail: a.t.j.r.cobbenhagen@tue.nl).

Abstract: The Guaranteed Error Machine (GEM) is a classification algorithm that allows the user to set a-priori (i.e., before data are observed) an upper bound on the probability of error. Due to its strong statistical guarantees, GEM is of particular interest for safety critical applications in control engineering. Empirical studies have suggested that a pool of GEM classifiers can be combined in a majority voting scheme to boost the individual performances. In this paper, we investigate the possibility of keeping the probability of error under control in the absence of extra validation or test sets. In particular, we consider situations where the classifiers in the pool may have different guarantees on the probability of error, for which we propose a data-dependent weighted majority voting scheme. The preliminary results presented in this paper are very general and apply in principle to any weighted majority voting scheme involving individual classifiers that come with statistical guarantees, in the spirit of Probably Approximately Correct (PAC) learning.

Copyright © 2020 The Authors. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0)

Keywords: Classification, Machine learning, Multi-agent systems, Randomized methods, Optimisation.

1. INTRODUCTION

A binary classifier is a function \( \hat{y} : \mathbb{R}^k \rightarrow \{0,1\} \); in supervised classification the classifier is trained (i.e., it is suitably chosen by an algorithm) from a set of previously recorded pairs \( (x_i,y_i) \), where \( x_i \in \mathbb{R}^k \) are vectors of features (e.g., health indicators extracted from a blood analysis; measures of a signal that enters a complex system; etc.) and \( y_i \in \{0,1\} \) are labels denoting the corresponding class (e.g., healthy/ill in the case of a blood analysis; bounded/unbounded in the case of a system response). When fed with a new vector of features \( x \) the classifier \( \hat{y}(\cdot) \) provides an automated prediction \( \hat{y}(x) \) for the corresponding \( y \). An important quantity to assess the quality of the classifier is the probability of error \( PE(\hat{y}) \), that is, the probability of the event \( \{ \hat{y}(x) \neq y \} \).

The Guaranteed Error Machine (GEM) is an algorithm to construct classifiers \( \hat{g}(\cdot) \) that was proposed in Campi [2010] and has a built-in mechanism to keep \( PE(\hat{g}) \) strictly under control, which makes it particularly attractive for safety critical applications (see e.g. Baronio et al. [2017]) and for control systems engineering when a quantitative assessment of sample-based control schemes is required (see e.g. Manganini et al. [2015]).

The bound on the error is achieved by adopting a ternary output \( \{0,1,\text{unknown}\} \), so that the classifier can refrain from classification: in the case of an abstention, the classifier makes no error irrespective of the value of \( y \). The key idea of GEM is that the complexity of a classifier can be tuned by the user by selecting the value of a parameter \( k \). A large value of \( k \) leads to classifiers that more often return a 0 or 1 value, but these classifiers misclassify more frequently, whereas smaller values of \( k \) correspond to more risk-averse classifiers that return \text{unknown} with higher probability. An alternative use of GEM, that has been exploited and studied in Carè et al. [2018], is that of constructing a set of GEM classifiers for increasing values of \( k \), and then selecting the one with the smallest value of \( k \), say \( \hat{k} \), for which no \text{unknowns} are returned. The probability of error of the resulting classifier cannot be bounded a-priori, but the connection between \( \hat{k} \) and \( PE(\hat{g}) \) is so strong that the observed value of \( \hat{k} \) comes with an informative confidence interval for \( PE(\hat{g}) \). Such confidence interval is easily computed without resorting to any extra validation or test set, so that GEM classifiers are said to be “self-testing classifiers”.

The construction of GEM comes with some degrees of freedom. For example, the construction starts from a data point \( (x_0,y_0) \) that is either known a-priori or randomly chosen from the available data set. A reasonable approach to reduce the arbitrariness of the choice is to consider all
the possible constructions simultaneously and then build a classifier based on majority voting among the available classifiers (which are called the base classifiers).

Although experimental studies in Manganini et al. [2015] suggest that majority voting classification schemes can actually boost the performance of GEM-like classifiers, it remains an open problem how to translate theoretical guarantees on the base classifiers to guarantees on the majority-voting classifier in such a way that they remain valid and practically meaningful: this is the problem that motivates our paper. Remarkably, GEM motivated our studies because of the tightness and practical usefulness of the resulting bounds, but, as we shall see, the results of this paper can be applied directly to any other type of base classifiers coming with confidence intervals on their probability of error in the spirit of Probably Approximately Correct (PAC) learning, see e.g. Graepel et al. [2005].

1.1 Contribution and structure of the paper

In Section 2 we introduce the mathematical framework of this paper and we observe that, at least in principle, majority voting schemes can worsen the performance with respect to the base classifiers. The conclusion is that one should aim at error estimation procedures that are able to detect whether or not the specific situation at hand is favourable to majority voting. We move some preliminary steps in this direction by connecting the error of the majority voting classifier to an agreement parameter. Section 3 is the main section of this paper, where we consider majority voting, and weighted majority voting, among base classifiers that are individually guaranteed to satisfy certain error thresholds at a given confidence level. We show the interesting fact that the proliferation of base classifiers does not necessarily lead to a loss of control on the probability of error. This is true even if the weights assigned to the base classifiers depend on the training set, under the condition that the voting scheme is adequately constrained. Conclusions are drawn in Section 4.

1.2 Existing literature and related research

The combination of classifiers into a voting-based classification scheme has a long tradition, and the literature on the topic is vast, in the wake of the celebrated gradient boosting machines and random forests, see Zhu [2015] for a review. On the theoretical side, we mention here the the notable studies of Schapire et al. [1998], Schapire and Freund [2012], where the generalization properties of majority voting schemes are related to the concept of margin in a statistical learning framework, and refer to Ruta and Gabrys [2002], Kuncheva et al. [2003], Kuncheva [2014] for studies on the best and worst cases. It is also worth observing that together with celebrated successes there came misconceptions: as it was noted in Varden and Morris [2013], the appeal to the Condorcet’s Jury Theorem is a common but weak and often misplaced argument to support voting decision schemes.

Our approach in this paper, similarly to Schapire et al. [1998] and differently from Ruta and Gabrys [2002], assumes that no validation set is available. On the other hand, differently from Schapire et al. [1998] and any other work that we are aware of, our focus is on transforming the existing guarantees for the base classifiers into guarantees for the voting classification scheme. At a conceptual level, the approach in our Section 3 has points in common with the so-called PAC-Bayes approaches, where a Kullback-Leibler divergence is used to penalise the departure from a reference distribution (which is called "prior" in the PAC-Bayes terminology). However, to apply the PAC-Bayes bounds of Lacasse et al. [2007] and Germain et al. [2015] the user needs either a validation set or the knowledge of the mechanism that generates the classifiers ("reconstruction function"). With our approach, instead, the user needs just to know the error thresholds for the base classifiers and their confidence level.

2. PRELIMINARY DEFINITIONS AND RESULTS

Let $ \Delta = X \times Y $ denote the set of all possible data points $(x, y)$, where $ x $ is the feature vector and $ y $ is a binary label, i.e. $ y \in \{0, 1\} $ (a data point $(x, y)$ will be denoted in compact form as $ \delta $). We assume that $ \Delta $ is equipped with a probability measure $ P_\Delta $ unknown to the user.

A training set $ T $ in $ \Delta $ is a random sample of data points $ \delta^{(1)}, \ldots, \delta^{(N)} $, that we assume to be drawn according to the product measure $ P_\Delta^N $ (i.i.d. assumption). A classifier $ \hat{y} $ is a map from $ X $ to $ \{0, 1, \text{unknown}\} $, and a classification algorithm $ A $ is a map from training sets $ T $ to classifiers. In the following we will consider an ensemble of classification algorithms $ \{A_c\} $, indexed by $ c \in C $. Most often, $ C $ will be a finite set, $ C = \{1, 2, \ldots, M\} $, but it can also be an infinite set. The classifier obtained by applying the classification scheme $ A_c $ to $ T $ will be denoted by $ \hat{y}_c $, and the dependence of $ \hat{y}_c $ from $ T $ will be left implicit.

We define the error function $ \mathcal{E}(\delta, c) $ as follows:

$$
\mathcal{E}(\delta, c) = \begin{cases}
0, & \text{if } \hat{y}_c(x) = y \\
1, & \text{if } \hat{y}_c(x) \in \{0, 1\} \text{ and } \hat{y}_c(x) \neq y.
\end{cases}
$$

The probability of error of a classifier $ \hat{y}_c $ is formally defined as

$$
\text{PE}(\hat{y}_c) = E_\Delta [\mathcal{E}(\delta, c)] = \int_\Delta \mathcal{E}(\delta, c) \, dP_\Delta(\delta).
$$

In practice, it is seldom the case that the classifiers $ \hat{y}_c $, $ c \in C $, agree unanimously on the label to assign to a feature vector $ x $. Hence, to let a qualified majority emerge from the labels $ \hat{y}_c(x) $, $ c \in C $, and produce an unambiguous answer in $ \{0, 1\} $, for any given $ x $ let us introduce the partition $ C = S_0(x) \cup S_1(x) \cup S_u(x) $, where $ S_0(x) = \{ c \in C : \hat{y}_c(x) = 0 \} $, $ S_1(x) = \{ c \in C : \hat{y}_c(x) = 1 \} $, and $ S_u(x) = \{ c \in C : \hat{y}_c(x) = \text{unknown} \} $. Furthermore, let us introduce a probability measure $ P_C $ on $ C $. $ P_C $ is always the same irrespective of $ x $ and it is used to establish the majority in a more general sense than the usual, “democratic” one (the latter is recovered when $ P_C $ is uniform). We define the voting majority to be

$$
S_+(x) = \begin{cases}
S_0(x), & \text{if } P_C(S_0(x)) \geq P_C(S_1(x)), \\
S_1(x), & \text{if } P_C(S_0(x)) < P_C(S_1(x)),
\end{cases}
$$

and the voting minority as $ S_-(x) = C - (S_+(x) \cup S_u(x)) $. The majority classifier is now defined as follows.

**Definition 1.** (Majority classifier). For all $ x $,

$$
\hat{y}_*(x) = \begin{cases}
0, & \text{if } P_C(S_0(x)) \geq P_C(S_1(x)), \\
1, & \text{if } P_C(S_0(x)) < P_C(S_1(x)).
\end{cases}
$$
The definition yields a well-posed binary classifier irrespective of $P_C(S_u(x))$, and takes into account, so to say, only the “actual voters”. The fact that, in the case of a tie, we have assigned $\tilde{y}_c(x) = 0$ to be the voting majority’s label (see equation (1)) is just for simplicity, and variations on the theme are clearly possible.

So far, the measure $P_C$ should have been thought of as a weighing tool to compute the majority. Now $P_C$ will be used to construct a peculiar, non-deterministic classifier that when asked to classify a point $x$ first samples an index $c$ from $C$ according to $P_C$ and then outputs the label $\tilde{y}_c(x)$. Such a classifier is known in the literature under the name of “Gibbs classifier”.

**Definition 2.** (Gibbs classifier) We denote by $\tilde{y}_C$ the random map $x \mapsto \tilde{y}_C(x)$, where $c$ is a random variable distributed according to $P_C$.

The role of the Gibbs classifier $\tilde{y}_C$ in this paper will be only instrumental to bounding $P_E(\tilde{y}_C)$. Observing that the Gibbs classifier misclassifies a given point $\hat{d}$ with probability $E_C[\mathcal{E}[\hat{d}, c]]$, and given that $\hat{d}$ and $c$ are independent since $c$ is always extracted from the same $P_C$ for every $x$, its probability of error is naturally defined as $P_E(\tilde{y}_C) = E_A[E_C[\mathcal{E}[\hat{d}, c]]]$. By an application of Fubini’s theorem, we obtain the following useful equivalence:

$$P_E(\tilde{y}_C) = E_C[E_A[\mathcal{E}[\hat{d}, c]]] = E_C[P_E(\tilde{y}_C)].$$

The following theorem relates $P_E(\tilde{y}_C)$ and $P_E(\tilde{y}_C)$.

**Theorem 1.** Define $A = \inf_{x \in X} P_C(S_u(x))$. It holds that $P_E(\tilde{y}_C) \leq \frac{1}{A} \cdot P_E(\tilde{y}_C)$.

**Proof.** Letting $1\{E\}$ be the indicator function of the event $E$, we have

$$P_E(\tilde{y}_C) \geq E_A[E_C[\mathcal{E}[\hat{d}, c]] \cdot 1\{c \in S_u(x)\}]$$

Note that, whenever $S_u(x)$ is non-empty, $\tilde{y}_C(x) = \tilde{y}_C(x)$ for all $c \in S_u(x)$; let $c^*_x$ be any element of $S_u(x)$. Then, (3) is equal to

$$E_A[E_C[\mathcal{E}[\hat{d}, c^*_x]] \cdot 1\{c \in S_u(x)\}]$$

and

$$= E_A[\mathcal{E}[\hat{d}, c^*_x]] \cdot \inf_{x \in X} E_C[1\{c \in S_u(x)\}]$$

$$\geq P_E(\tilde{y}_C) \cdot \inf_{x \in X} P_C(S_u(x)) = P_E(\tilde{y}_C) \cdot A.$$

The term $A$ is an “agreement index” that quantifies the size of the voting majority in the worst case. The crucial fact for the theory here developed is that, since the distribution $P_C$ is chosen by the user, such index is (at least in principle) an observable quantity. In the rest of the paper, the observable quantity $A$ will allow us to express our results on the error of the majority classifier in a concise way. However, its global nature ($A$ is computed as an infimum over the whole $X$) can be limiting in practice: the next two theorems, which can be proven similarly as Theorem 1, stand as alternative building blocks for a more practical approach where guarantees on the probability of error are based on the agreement of the base classifiers at a given point $x$.

Theorem 2. Define $A_x = P_C(S_u(x))$ and let $a \in (0, 1]$. It holds that

$$P_\Delta(\tilde{y}_C(x) \neq y \land A_x \geq a) \leq \frac{1}{A} \cdot P_E(\tilde{y}_C).$$

**Theorem 3.** Consider the case where the base classifiers deliver no unknowns. Assume that $P_\Delta(A_x \in [a, \bar{a}]) \geq \mu$, with $0.5 < \bar{a} \leq a \leq 1$. Then,

$$P_\Delta(\tilde{y}_C(x) \neq y \land A_x \in [a, \bar{a}]) \leq \frac{P_E(\tilde{y}_C) - (1 - \bar{a})\mu}{2\bar{a} - 1}.$$

Going back to Theorem 1, we remark that, without restrictions on the probability with which $\tilde{y}_C(x)$ = unknown, the misclassification probability can worsen to an arbitrary extent with respect to the performance of the individual classifiers. This is not surprising as we are forcing the majority classifier to be a binary classifier, so that for values of $x$ such that almost all the classifiers are uncertain, a decision is made based on the voice of a small subset of them that could always be wrong. We remark also that, if there are no unknowns in the set $\{\tilde{y}_C(x): c \in C, x \in X\}$, then we necessarily have $A \geq \frac{1}{2}$, from which the well-known bound $P_E(\tilde{y}_C) \leq 2P_E(\tilde{y}_C)$ is obtained. It can be shown that this bound is tight and there are situations where the performance of the majority worsens with respect to the performance of any individual classifier. Such possibility calls for mathematical results that enable the user to detect whether or not the situation at hand is favourable to majority voting.

### 3. A POOL OF GUARANTEED CLASSIFIERS

We start by considering the situation where the probability of error of each classifier $\tilde{y}_C$ is bounded by a given $\epsilon_\nu$. From equation (2) and Theorem 1 it follows immediately that

$$P_E(\tilde{y}_C) \leq \frac{1}{A} E_C[\epsilon_\nu].$$

This simple relation suggests that, when some classifiers come with small $\epsilon_\nu$ while others come with a large $\epsilon_\nu$, it is advisable to concentrate the distribution $P_C$, according to which the majority is computed, on the best guaranteed classifiers. For example, we might want to define $P_C$ so that it concentrates on the classifier with minimum $\epsilon_\nu$, say $\tilde{c}$; then, assuming there is a unique such classifier and there are no unknowns, the threshold given by the right-hand side of (4) becomes equal to $\tilde{c}$, being $A = 1$ by construction.

The possibility of choosing $P_C$ according to the thresholds $\epsilon_\nu$ is complicated by the fact that, in real-life machine learning, the values of $\epsilon_\nu$ are typically stochastic estimates that are not valid with absolute certainty but only with some confidence. In particular, we study here the situation where the $\epsilon_\nu$ are functions of the training set $T$, as it is expressed by the following condition.

**Condition 1.** For all $c \in C$, there exists a function $\epsilon_\nu: \Delta^N \rightarrow [0, 1]$ such that

$$\Delta^N(PE(\tilde{y}_C) > \epsilon_\nu(T)) \leq \beta.$$

---

1 Throughout the paper we assume that the error function is measurable over $\Delta \times P_C$ with respect to the product measure $P_\Delta \times P_C$.

2 This reminds us of a famous quote from Cicero’s: *E terto in dissensione civili, cum boni plus quam multi valent, expendendos cives, non numerandos patro.* (De re publica VI.1, ed. J.G.F. Powell.)
Remark 1. Clearly, Condition 1 can be trivially met by any classification scheme that divides the training set into two parts and uses only the first part to train the classifier, while the second part is used to estimate the probability of error at a sufficient confidence level. More interestingly, Condition 1 is met, with values of $\epsilon_c$ and $\beta$ that are often satisfactory, by a modified version of the GEM algorithm of Campi [2010], where the a-priori selected complexity parameter $k$ is increased by 1 until there are no more unknowns, see Carè et al. [2018] for an application of this idea. Moreover, it is satisfied by classification schemes to which the more general theory of the Wait-and-Judge scenario approach (Campi and Garatti [2018], Carè et al. [2019]) can be applied to relate probability of error and complexity of the classifier. Also compression schemes, see Graepel et al. [2005], Margellos et al. [2015], Campi et al. [2018], satisfy such a condition, although the resulting bounds are often more conservative and less applicable than with GEM. In all of these cases, the dependence of the error threshold $\epsilon_c$ on $T$ is typically reduced to the dependence on a complexity parameter $k$ that can be easily computed from $T$.

For short, we will denote by $\hat{c}_c$, the value $\epsilon_c(T)$. We now assume that $P_C$ is a T-dependent probability measure that the user can choose at will, for example, but not necessarily, by using the information carried by the values of $\hat{c}_c$.

When $C$ is finite, the following proposition can be often used to ensure that (4) holds true with high confidence

**Proposition 1.** Under Condition 1, if $C$ is finite, $C = \{1, \ldots, M\}$, then

$$P_N^T \left\{ P_E(\hat{y}_c) \leq \frac{1}{A} E_C(\hat{c}_c) + \varrho \right\} \geq 1 - M \beta. \quad (6)$$

**Proof.** Clearly, $P_N^T \left\{ P_E(\hat{y}_c) \leq E_C(\hat{c}_c) \right\} \geq P_N^T \left\{ P_E(\hat{y}_c) \leq \hat{c}_c \right\} \forall c \in C = 1 - P_N^T \left\{ \exists c \in C \text{ s.t. } P_E(\hat{y}_c) > \hat{c}_c \right\}$. By a simple union bound argument, $P_N^T \left\{ \exists c \in C \text{ s.t. } P_E(\hat{y}_c) > \hat{c}_c \right\} \leq \sum_{c=1}^M P_N^T \left\{ P_E(\hat{y}_c) > \hat{c}_c \right\}$ and the claim follows by Condition 1. □

However, for a large number $M$ of classifiers, the confidence value $1 - M \beta$ ensured by (6) becomes easily too low, and indeed uninformative.

The next result shows that, by enforcing a constraint on the way in which $P_C$ can vary as a function of $T$, it is still possible to keep under control the probability of error of the majority even when $C$ is too large for (6) to be useful. In particular, the following theorem shows that if $P_C$ is prevented from concentrating too much (as compared to a reference distribution $P_C$), then meaningful confidence levels can be achieved also when $C$ is infinite.

**Theorem 4.** Assume that there exists a probability measure $P_C$ that does not depend on $T$ and a $\bar{\kappa} > 0$ such that $P_C(E) \leq \bar{\kappa} P_C(E)$ for all the measurable sets $E \subseteq X$ and every $T$. Then, under Condition 1, for $\varrho > 0$ it holds that

$$P_N^T \left\{ P_E(\hat{y}_c) \leq \frac{1}{A} E_C(\hat{c}_c) + \varrho \right\} \geq 1 - \frac{\bar{\kappa}}{\bar{\varrho}} \cdot \beta. \quad (7)$$

**Proof.** By Markov’s inequality it holds that

$$P_N^T \left\{ P_E(\hat{y}_c) - E_C(\hat{c}_c) > \varrho \right\} \leq \frac{1}{\bar{\varrho}} E_N^{\Delta} \left\{ P_E(\hat{y}_c) - E_C(\hat{c}_c) \right\},$$

where $(x)^+ = \max\{x, 0\}$. By (2) and by Jensen’s inequality ($(\cdot)^+$ is convex), we get

$$\frac{1}{\bar{\varrho}} E_N^{\Delta} \left\{ P_E(\hat{y}_c) - E_C(\hat{c}_c) \right\} \leq \frac{1}{\bar{\varrho}} E_N^{\Delta} \left[ P_E(\hat{y}_c) - \hat{c}_c \right]^+ \leq \frac{1}{\bar{\varrho}} \bar{\kappa} E_N \left[ P_E(\hat{y}_c) - \hat{c}_c \right]^+ \leq \frac{1}{\bar{\varrho}} \bar{\kappa} E_N \left[ P_N^T \left\{ P_E(\hat{y}_c) - \hat{c}_c \right\} \right] \leq \frac{\bar{\kappa}}{\bar{\varrho}} \beta,$

where the last inequality is by Condition 1. Then, Theorem 1 can be invoked to claim that $P_E(\hat{y}_c) \leq \frac{1}{A} (E_C(\hat{c}_c) + \varrho)$ for all $T \in \Delta^N$ except for a subset with probability at most $\frac{\bar{\varrho}}{\bar{\kappa}}$, which was the claim to be proven. □

When $P_C$ is uniform, $\bar{\kappa}$ can be interpreted as a parameter that balances between strict democracy and oligarchy: when $C = \{1, \ldots, M\}$, $\bar{\kappa} = 1$ enforces $P_C$ to be uniform (democracy): the confidence value is maximum $(1 - \frac{\beta}{\varrho})$ but $E_C(\hat{c}_c)$ averages over both high and low values of $\hat{c}_c$; on the other hand, $\bar{\kappa} = M$ does not constrain $P_C$ at all, so that one is free to choose the measure $P_C$ that concentrates all the probability on a single classifier with the smallest $\hat{c}_c$: in this case, $E_C(\hat{c}_c)$ is minimized but the confidence becomes $1 - \frac{M}{\beta} \beta$. Choosing $\bar{\kappa} = j$, $j \in \{2, \ldots, M - 1\}$ leads to intermediate situations where the probability is distributed over a subset of $j$ classifiers (e.g., the $j$ classifiers with smallest $\hat{c}_c$). A similar reasoning applies also to the case where $P_C$ has density.

**Example 1.** (Numerical Example). Suppose that $M = 10000$ classifiers are given, each coming with a guarantee like the one in Condition 1 where $\beta = 10^{-4}$, and let $\varrho = 0.02$. If $P_C$ can concentrate on a single classifier, our results do not provide any meaningful confidence as $M \beta = 1$. However, by setting $P_C(c) = \frac{1}{10000}$ for all $c$ (uniform distribution) and $\bar{\kappa} = 2$, we are allowed to build a majority classifier based on the 5000 ($= M/\bar{\kappa}$) classifiers with the smallest $\hat{c}_c$ and discharge the vote of the remaining 5000. Indeed, a probability measure $P_C$ having uniform mass over the 5000 classifiers with the smallest $\hat{c}_c$ and 0
For short, we will denote by although the resulting bounds are often more conservative compression schemes, see Graepel et al. [2005], Margellos

$k$

probability of error and complexity of the classifier. Moreover, it is satisfied by an application of this idea. Remark 1.

We notice the subtle fact that, with GEM, iteratively increasing necessarily, by using the information carried by the values

$\epsilon$

mass over all the other classifiers satisfies $P_{C}(c) \leq 2 \cdot \tilde{P}_{C}(c)$ for all $c$; hence Theorem 4 ensures a confidence of 99%.

Example 2. (Simulations). We ran a preliminary campaign of simulations with the following setup: a pool of $M = 250$ classifiers built according to the scheme of Carè et al. [2018] was trained over the same dataset of 250 independent data points $(x_i, y_i)$, where each $x_i = (x_{i1}, x_{i2})$, $i = 1, ..., 250$, was chosen according to the uniform distribution over $[0,1]^2$, and $y_i$ was distributed according to the following function:

$y_i = y(x_i) = \begin{cases} 1, & \text{if } x_{2i} \geq \left(x_{1i} - \frac{1}{2}\right) \cos(25x_{1i}) + \frac{1}{2} \\ 0, & \text{otherwise} \end{cases}$

(the same function as in the experimental section of Cobbenhagen et al. [2019]); each classifier was trained having the $i$-th point of the dataset as the starting point and the $N = M - 1 = 249$ remaining ones as the training set. Since these classifiers come with different values of $c_c$ that are individually valid with confidence $1 - \beta$, the results of this section apply: we sorted the classifiers in order of increasing $c_c$, we chose a $\tilde{P}_{C}$ uniform over the “voting set” of the “best” $n$ classifiers (i.e. those with the smallest $c_c$), and we tested the resulting majority classifier $\hat{y}_n$ on $T = 10000$ random points extracted according to the same distribution of the training set, recording the average error $\tilde{P}(n) = \#\text{errors}/T$. We repeated the entire procedure for all the possible sizes $n = 1, 2, ..., 250$ of the voting set, and 1000 times for each size $n$ in a Monte Carlo simulation. The results are summarized in Fig. 1.

As the reader can see from this simulation, and as we have ascertained from other ones, restricting the classifiers to a very small voting set is not convenient; on the other hand, there seems to be a tendency to improvement as the size increases, that settles around half the number of classifiers. This fact is coherent with the trade-off between error and confidence expressed by Theorem 4: Fig. 2 shows the values $E[C_c + \rho] \leq 2(14.22\% + 1\%) = 30.44\%$ (where, for simplicity, we have substituted $A$ with its lower bound $A \geq \frac{1}{2}$), to which a confidence of $1 - \frac{2\beta}{\rho} = 98\%$ is attached. By comparing Figs. 3 and 4, one can note that, in 1000 runs, the upper-bounds $\frac{1}{A}(E[C_c + \rho])$ are always higher than the empirical errors. On the one hand this confirms the soundness of the approach; on the other hand it suggests that the bounds hold with a probability higher than the
that vein of Theorems 2 and 3 (in this respect, it is remarkable expected to be removed by exploiting local values $A$ in the 500-set is computed according to the theory in Baronio et al. [2018] corresponding to $\beta = 10^{-4}$.

98% probability guaranteed by Theorem 4: this reveals some conservatism that calls for further investigation. In particular, an important source of conservatism is expected to be removed by exploiting local values $A_x$ in the vein of Theorems 2 and 3 (in this respect, it is remarkable that $\mathbb{E}_\Delta[A(x)]$ was always above 90% in our Monte Carlo runs).

4. CONCLUSIONS AND FUTURE STUDIES

In this paper, we have investigated the problem of transforming individual PAC guarantees for base classifiers into guarantees for majority voting decision schemes. We have discussed weighted majority decision schemes where the weights can depend on the training set, and we have shown that, by suitably restricting the variability of the weights, guarantees can be provided even when the base classifiers are infinitely many. While the bounds in this paper are expressed in terms of a worst-case agreement index, it is possible to provide bounds that are differentiated based on the agreement level for the observed feature vector, in the line of Theorems 2 and 3. In future works, we will aim at exploiting the specific construction of the base classifiers to improve the guarantees; for example, estimators based on the number of support points, like those in Cobbenhagen et al. [2019], could be used to detect situations that are more favourable to majority voting.

While in this paper we focused on bounding the probability of error computed with respect to all the possible pairs $(x, y)$, we plan to study the probability of error conditional to the level of agreement among the base classifiers. Preliminary studies in this direction can be found in Cobbenhagen et al. [2019].

The extension of the framework of this paper to the regression problem, in the line of Garatti and Carè [2019], is also the subject of current research.

REFERENCES


