Novel bounds on the probability of misclassification in majority voting: leveraging the majority size

A.T.J.R. Cobbenhagen, A. Carè, M.C. Campi, F.A. Ramponi, D.J. Antunes, W.P.M.H. Heemels

***Abstract*— Majority voting is often employed as a tool to increase the robustness of data-driven decisions and control policies, a fact which calls for rigorous, quantitative evaluations of the limits and the potentials of majority voting schemes. This work focuses on the case where the voting agents are binary classifiers and introduces novel bounds on the probability of misclassification conditioned on the size of the majority. We show that these bounds can be much smaller than the traditional upper bounds on the probability of misclassification. These bounds can be used in a 'Probably Approximately Correct' (PAC) setting, which allows for a practical implementation.**

***Index Terms*— Machine learning, Agents-based systems, Statistical learning**

I. Introduction

A. Binary classification and majority voting

THE objective in *classification* is to attach a *label* to an instance of a set of *features*. Like many machine learning techniques, classification has found a place as a standard tool in the control community. One important example of application is in the medical domain. For example, consider the case of an Automatic External Defibrillator (AED), which has to determine whether to shock a person experiencing cardiac arrest or not (the *label*) based on several properties such as blood pressure, heart rate, etc (the *features*) (see [1] for such a classifier, which is based on [2]). If the shock is effective, then the crisis is over, otherwise the shock may worsen the situation by further damaging the cardiac muscle. This implies the need to design techniques to support the control action of whether one should shock or not. Hence, we see that classification plays a role similar to a state estimator in this application.

Furthermore, classification has been used in optimal control of affine switched systems [3], policy iteration [4], and classification of the environment in order to select the appropriate controller [5], [6]. The control community has particularly been concerned about the theoretical guarantees that can be attached to classification schemes, which has motivated original investigations such as [7].

The main metric to judge such a classifier is the probability of misclassification, i.e., the probability that the classifier assigns the wrong label to a new feature vector. The optimal/safest classifier from the feature vectors to their binary labels is in many cases unknown and/or difficult to model. For this purpose, machine learning algorithms have been developed that construct approximations of these mappings from a set of example pairs of feature vectors and labels (*training set*). From an empirical estimate of the probability of misclassification in the training and/or validation set, it is possible to give an upper bound on the true probability of misclassification with a high confidence. We refer to works in *statistical learning theory* for a detailed description on this matter, see, e.g., [8], [9].

One approach to improve existing classifiers is to combine many of them. There is a vast amount of literature on this topic, see, e.g., [10]–[12] for several approaches. The main idea behind combining multiple classifiers is that the classifiers may compensate for the weaknesses of each other and thus increase the overall performance. In the present work we consider a weighted majority voting scheme. That is, the labels given by the classifiers are weighted and the label with the highest weighting is the classification of the majority vote. We consider the case where the labels can take only two values (*binary classification*).

B. Contributions: Tighter bounds on majority misclassification due to a novel perspective

Bounding the probability of error of majority voting classifiers has been a topic of interest for several years, see e.g., [10], [11], [13] for early work. Such bounds have also been used to derive new machine learning algorithms [14]. Unfortunately, although majority voting can very often lead to significant improvements, it has been shown that majority voting can, in principle, worsen the performance with respect to the individual classifiers [15].

This work is part of a project aiming at putting data-based decisions on a solid theoretical ground and, in this particular case, understanding the limits and exploiting the potentials of majority decisions. Here we adopt a novel perspective. Namely, instead of providing a single 'expected' probability of misclassification over the entire set of feature vectors, we look at the *operational* side of the majority voting scheme and provide a probability of misclassification that depends on the size of the majority for the classified feature vector. That is

This work is in part supported by "Toeslag voor Topconsortia voor Kennis en Innovatie" (TKI HTSM) from the Ministry of Economic Affairs, the Netherlands.

R. Cobbenhagen, D. Antunes and M. Heemels are with the Department of Mechanical Engineering, Eindhoven University of Technology, 5612 AZ Eindhoven, The Netherlands (e-mail: {a.t.j.r.cobbenhagen,d.antunes,m.heemels}@tue.nl).

A. Carè, M. Campi and F.A. Ramponi are with the Dipartimento di Ingegneria dell'Informazione, Università di Brescia, Brescia 25123, Italy (e-mail: {algo.care,marco.campi,federico.ramponi}@unibs.it).

Corresponding author: R. Cobbenhagen.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/LCSYS.2020.3040961, IEEE Control Systems Letters

to say, if one classifies a newly obtained feature vector using a majority voting scheme, we give tighter guarantees on the probability of misclassification, given that we know the size of the majority. Preliminary work on this perspective in the case of two classifiers has been presented by some of the authors of the present paper in [16]. The present work forms a significant extension as the main results hold for majority voting schemes of any finite number of classifiers, which requires a different approach than in [16].

The main results of this work are two novel upper bounds on the probability of misclassification conditioned on the size of the majority. These novel bounds as well as their 'unconditional' counterparts from the literature rely on unknown parameters of the ensemble of classifiers. These parameters can typically be estimated with high confidence, for example (but not necessarily) by resorting to an extra validation set. This fact was taken into account in the construction of the novel bounds in this work such that they are 'tighter' than their unconditional counterparts at the same confidence level.

C. Structure of the paper

The remainder of this paper is structured as follows. Section II presents some preliminaries on majority voting schemes. The main results are presented in Section III, along with numerical results to demonstrate their effectiveness. We demonstrate that the novel bounds can be used in order to design and analyse abstaining classifiers in Section IV.

II. Majority voting preliminaries

A. Mathematical notation

Let $\Delta = X \times Y$ denote the set of all possible data points $\delta = (x, y) \in \Delta$, where $X \subseteq \mathbb{R}^n$ is the set of possible *feature vectors* (for some $n \in \mathbb{N}$) and $Y = \{0, 1\}$ is the set of *labels*. We assume that Δ is equipped with a probability distribution $\mathbb{P}_\Delta$ that is *not known to the user*.

A training set $\mathcal{T}^N \in \Delta^N$ is a set of N random points $\{\delta^{(1)}, ..., \delta^{(N)}\}$ drawn according to the product probability $\mathbb{P}_\Delta^N$ (hence, i.i.d.). A classifier $\widehat{y}$ is a mapping $X \to Y$ and a classification algorithm is a map from training sets $\mathcal{T}^N$ to classifiers. We consider a *pool* of $M \in \mathbb{N}$ base classifiers $\widehat{y}_c \in \mathcal{P}$, where $\mathcal{P} := \{\widehat{y}_1, \widehat{y}_2, ..., \widehat{y}_M\}$. The power set of $\mathcal{P}$ is denoted by $2^{\mathcal{P}}$.

We use the notation $\mathbb{1}\{\cdot\}$ to denote the indicator function such that $\mathbb{1}\{A\} = 1$ if A is true and $\mathbb{1}\{A\} = 0$ if A is false.

B. Majority voting classifier

The majority voting classifier in this paper is denoted by $\widehat{y}_*$ and works as follows. Each base classifier $\widehat{y}_c$ is assigned a weight w_c. For ease of exposition, we represent these weights of the base classifiers through a probability distribution function $\mathbb{P}_Q : 2^{\mathcal{P}} \to [0, 1]$. In particular, assuming that the weights are positive and sum up to 1, for any $\mathcal{P}' \subseteq \mathcal{P}$ the weight is given by

$$\mathbb{P}_Q\{\widehat{y} \in \mathcal{P}'\} = \textstyle\sum_{c: \widehat{y}_c \in \mathcal{P}'} w_c.$$

This distribution $\mathbb{P}_Q$ may depend on the training set. Note that, contrarily to $\mathbb{P}_\Delta$, the distribution $\mathbb{P}_Q$ is known to the user and can be chosen.

When classifying an instance $x \in X$, the classifiers are partitioned into two sets $\mathcal{P}_0(x)$ and $\mathcal{P}_1(x)$ based on the label they return. The weights of the classifiers within each set are summed and the set with the largest weight is denoted by $\mathcal{P}_*(x) \subseteq \mathcal{P}$. Formally, $\mathcal{P}_*(x)$ is defined as[1]

$$\mathcal{P}_*(x) = \begin{cases} \mathcal{P}_0(x), & \text{if } \mathbb{P}_Q\{\widehat{y} \in \mathcal{P}_0(x)\} \geq \mathbb{P}_Q\{\widehat{y} \in \mathcal{P}_1(x)\}\,, \\ \mathcal{P}_1(x), & \text{if } \mathbb{P}_Q\{\widehat{y} \in \mathcal{P}_1(x)\} > \mathbb{P}_Q\{\widehat{y} \in \mathcal{P}_0(x)\}\,. \end{cases}$$

The majority classifier $\widehat{y}_*(x)$ assigns to x the label of the classifiers in $\mathcal{P}_*(x)$.

Furthermore, let $\mathcal{E}$ denote the error function such that for any $\delta = (x, y) \in \Delta$ and any classifier $\widehat{y}$ we have

$$\mathcal{E}(\delta, \widehat{y}) := \mathbb{1}\{\widehat{y}(x) \neq y\}.$$

The probability of error for a classifier $\widehat{y}$ is then defined as

$$\text{PE}(\widehat{y}) := \mathbb{E}_\Delta[\mathcal{E}(\delta, \widehat{y})] = \mathbb{P}_\Delta\{\widehat{y}(x) \neq y\}\,.$$

The probability of misclassification of the majority voting classifier $\widehat{y}_*$ is denoted by

$$B_Q = \text{PE}(\widehat{y}_*).$$

C. Gibbs classifier

In order to provide bounds on the probability of error of the majority voting scheme, we will make use of a stochastic classifier $\widetilde{y}_Q$ that works as follows. For an input feature vector $x \in X$, a classifier $\widehat{y} \in \mathcal{P}$ is randomly selected according to $\mathbb{P}_Q$ and the feature vector is classified according to the output of that classifier. This type of classifier is known as the *Gibbs classifier* [17, Chapter 4]. The probability of misclassification of the Gibbs classifier is equal to $G_Q := \text{PE}(\widetilde{y}_Q)$. Denoting by $\mathbb{E}_Q$ the expectation taken with respect to $\mathbb{P}_Q$, by Fubini's theorem, we have

$$G_Q = \mathbb{E}_\Delta[\mathbb{E}_Q[\mathcal{E}(\delta, \widehat{y})]] = \mathbb{E}_Q[\mathbb{E}_\Delta[\mathcal{E}(\delta, \widehat{y})]] = \mathbb{E}_Q[\text{PE}(\widehat{y})]\,, \tag{1}$$

and thus G_Q is the weighted average of the probability of misclassification of the base classifiers.

D. Agreement index and expected disagreement

Throughout this paper we will make use of the notion of an *agreement index* (as defined in [18]),

$$A_Q(x) = \mathbb{P}_Q\{\widehat{y} \in \mathcal{P}_*(x)\}\,,$$

which is the 'weight' or 'size' of the majority for a feature vector $x \in X$. Note that the value of $A_Q(x)$ is known to the user as it can be computed based on the user-chosen $\mathbb{P}_Q$, and, by the definition of the majority voting, it holds that $A_Q(x) \in [\frac{1}{2}, 1]$ for all $x \in X$.

Furthermore, we make use of the *expected disagreement*, denoted by d_Q and defined as the probability that two base classifiers $\widehat{y}'$ and $\widehat{y}''$ do not assign the same label to a feature vector if one were to select these two base classifiers independently according to $\mathbb{P}_Q$ with replacement. It follows that d_Q can be expressed as

$$\begin{aligned} d_Q :=& \mathbb{E}_\Delta \mathbb{E}_Q \mathbb{E}_Q[\mathbb{1}\{\widehat{y}'(x) \neq \widehat{y}''(x)\}] \\ =& 2\mathbb{E}_\Delta \mathbb{E}_Q \mathbb{E}_Q[\mathbb{1}\{\widehat{y}'(x) \in \mathcal{P}_*(x)\}\mathbb{1}\{\widehat{y}''(x) \notin \mathcal{P}_*(x)\}] \\ =& 2\mathbb{E}_\Delta[A_Q(x)(1 - A_Q(x))]\,. \end{aligned} \tag{2}$$

[1] In case of $\mathbb{P}_Q\{\widehat{y} \in \mathcal{P}_0(x)\} = \mathbb{P}_Q\{\widehat{y} \in \mathcal{P}_1(x)\}$, we have arbitrarily chosen to assign the label 0 as the majority. Other rules are possible.

E. Margin

The notion of *margin* M_Q is defined in [10] as the mapping $M_Q : \Delta \to [-1, 1]$ such that, in the case of binary classifiers, for any $\delta = (x, y) \in \Delta$,

$$\begin{aligned} M_Q(\delta) &:= \mathbb{P}_Q\{\widehat{y}(x) = y\} - \mathbb{P}_Q\{\widehat{y}(x) \neq y\} \\ &= \mathbb{E}_Q[1 - \mathcal{E}(\delta, \widehat{y})] - \mathbb{E}_Q[\mathcal{E}(\delta, \widehat{y})] = 1 - 2\mathbb{E}_Q[\mathcal{E}(\delta, \widehat{y})]\,. \end{aligned}$$

Hence, the expected value of the margin (also known as the *strength* of the pool of classifiers [11]) is equal to

$$\mathbb{E}_\Delta[M_Q(\delta)] = 1 - 2\,G_Q. \tag{3}$$

For any $\delta \in \Delta$, the margin $M_Q(\delta)$ captures both the size of the majority as well as whether the majority correctly classifies the feature vector. This is evident from the following derivation. For any $\widehat{y}' \in \mathcal{P}_*(x)$, it is true that $\mathcal{E}(\delta, \widehat{y}') = \mathcal{E}(\delta, \widehat{y}_*)$. Furthermore, any $\widehat{y}'' \notin \mathcal{P}_*(x)$ must assign the opposite label to x than any $\widehat{y}' \in \mathcal{P}_*(x)$. Hence, the error is also the opposite and therefore $\mathcal{E}(\delta, \widehat{y}'') = 1 - \mathcal{E}(\delta, \widehat{y}_*)$ for any $\widehat{y}'' \notin \mathcal{P}_*(x)$. Combining these observations yields

$$\begin{aligned} M_Q(\delta) &= 1 - 2\mathbb{E}_Q\big[\mathcal{E}(\delta, \widehat{y})\big] \\ &= 1 - 2\mathbb{E}_Q\big[\mathcal{E}(\delta, \widehat{y}_*)\mathbb{1}\{\widehat{y} \in \mathcal{P}_*(x)\} \\ &\qquad + (1 - \mathcal{E}(\delta, \widehat{y}_*))(1 - \mathbb{1}\{\widehat{y} \in \mathcal{P}_*(x)\})\big] \\ &= (2A_Q(x) - 1)\,(1 - 2\mathcal{E}(\delta, \widehat{y}_*)). \end{aligned} \tag{4}$$

Since $(1 - 2\mathcal{E}(\delta, \widehat{y}_*)) \in \{-1, +1\}$ and $A_Q(x) \geq \frac{1}{2}$, it is true that $|M_Q(\delta)| = 2A_Q(x) - 1$. Clearly, the sign of the margin shows whether the majority vote is correct about x (the margin is positive) or not (the margin is negative). The second moment of the margin is thus completely determined by $A_Q(x)$. In fact, using (2), it holds that

$$\mathbb{E}_\Delta\big[M_Q(\delta)^2\big] = \mathbb{E}_\Delta\big[(2A_Q(x) - 1)^2\big] = 1 - 2d_Q. \tag{5}$$

We can combine (5) with (3) to conclude that the variance of the margin is given by

$$\mathbb{V}_\Delta[M_Q(\delta)] = 4G_Q(1 - G_Q) - 2d_Q. \tag{6}$$

F. Bounds on the majority voting error

Using the margin, several bounds on B_Q have been proven in the literature. A well-known result is the following (see, e.g., [14], [15]).

Lemma 1 (2-bound): It holds that

$$B_Q \leq 2G_Q.$$

The factor of 2 in the bound in Lemma 1 was shown to be tight [15], which exposes a possible limitation of majority voting. Namely, majority voting can result in a worse performance than the average performance. This is one of the main motivations behind the analysis of the majority voting misclassification probability.

By resorting to additional information on the distribution of $A_Q(x)$ in the form of d_Q, it is possible to provide a more general result. The following lemma was introduced in [19] (see [14] for a more detailed analysis) and is known as the $\mathcal{C}$-bound.

Lemma 2 ($\mathcal{C}$-bound, [14], [19]): If $G_Q < \frac{1}{2}$, then it holds that

$$B_Q \leq 1 - \frac{(1 - 2G_Q)^2}{1 - 2d_Q}. \tag{7}$$

The right-hand side (RHS) of (7) can be smaller than G_Q.

III. Main results

A. Main theorems

As mentioned in Section I, the key new viewpoint in this work is that we look at the probability of misclassification, conditioned on the size of the majority. Led by the idea that a larger majority in many cases is associated with a smaller error of the majority vote, we define the 'probability of error conditional to the majority size' as

$$\begin{aligned} \mathcal{C}_a &:= \mathbb{P}_\Delta\{\,\widehat{y}_*(x) \neq y \,|\, A_Q(x) \geq a\,\} \\ &= \frac{\mathbb{P}_\Delta\{\widehat{y}_*(x) \neq y \wedge A_Q(x) \geq a\}}{\mathbb{P}_\Delta\{A_Q(x) \geq a\}}. \end{aligned} \tag{8}$$

The main results of this paper are the following two theorems that provide upper bounds to (8), the first one depending only on G_Q.

Theorem 1: For any $a \in [\frac{1}{2}, 1 - G_Q)$, it holds that

$$\mathcal{C}_a \leq \frac{1 - a}{a}\,\frac{G_Q}{1 - a - G_Q}. \tag{9}$$

The proof can be found in Appendix I. The reason to investigate a bound that only requires G_Q is that G_Q is often one of the easiest properties to estimate in majority classification schemes. An example is the case where one has several classifiers at his disposal with individual guarantees on the probabilities of misclassification. One can then estimate G_Q as the weighted average of the estimates of the individual probabilities of misclassification, see, e.g., [18].

By conditioning on the fact that the majority has a certain size, Theorem 1 is able to provide a better guarantee than the '2-bound' of Lemma 1. It is easily derived that $G_Q \leq \frac{3}{2} - \sqrt{2} \approx 0.0858$ is a necessary condition for the RHS of (9) to be no larger than $2G_Q$ for at least one value of a. Furthermore, the minimal value of the RHS of (9) is attained at $a = \alpha := 1 - \sqrt{G_Q}$ at which the value is $\frac{G_Q}{\alpha^2}$. Note that for $a \geq \frac{1}{2} + \frac{1}{2}\sqrt{1 - 4\,G_Q}$, the RHS of (9) is larger or equal to 1 and hence provides a trivial upper bound. Theorem 1 can thus be conservative. However, if we assume that we only have G_Q as information, then Theorem 1 is not conservative in the sense that it can drastically improve upon the bounds in the literature, i.e., Lemma 1.

As with Lemma 1, the RHS of (9) is larger or equal to G_Q. The added value of Theorem 1 with respect to Lemma 1 is that it demonstrates that even though majority voting can worsen the performance, it does not worsen by much. Theorem 1 modulates the performance guarantee based on the observed majority.

In practice, a bound on G_Q can be computed from available bounds on the probability of error of the individual classifiers, or by resorting to more sophisticated 'Probably Approximately Correct' (PAC) bounds such as in the PAC-Bayes framework, which is of particular interest when there are many base classifiers and $\mathbb{P}_Q$ is user-chosen. See also Section III-D for an example.

The second main result is an upper bound on $\mathcal{C}_a$ that depends on G_Q and d_Q.

Theorem 2: If $G_Q \leq \frac{1}{2}$, then for any $a \in [\frac{1}{2}, \frac{1}{2} + \frac{1}{2}\sqrt{1 - 2d_Q})$ it holds that

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/LCSYS.2020.3040961, IEEE Control Systems Letters

$$\mathcal{C}_a \leq \frac{G_Q(1-G_Q)-\frac{1}{2}d_Q}{G_Q(1-a)+a(a-G_Q)-\frac{1}{2}d_Q}\cdot\frac{2a(1-a)}{2a(1-a)-d_Q}. \tag{10}$$

The proof can be found in Appendix II. Similar to how Theorem 1 was able to provide better guarantees than Lemma 1, Theorem 2 can yield siginificantly better guarantees than Lemma 2. In fact, the RHS of (10) can be much smaller than G_Q.

In conclusion, Theorem 1 requires little information (only G_Q) and is more oriented towards a 'worst-case guarantee'. Theorem 2 provides a significant step towards detecting the actual benefits of a majority voting classifier.

B. Four useful lemmas

The proofs of Theorems 1 and 2 make use of several lemmas that provide bounds on either the numerator or denominator of (8). These lemmas are interesting in their own right. The proofs can be found in the Appendices.

Firstly, we provide the following two upper bounds on $\mathbb{P}_\Delta\{\widehat{y}_*(x)\neq y \wedge A_Q(x)\geq a\}$. The first of them appeared (without proof) in [18] and only requires knowledge on G_Q.

Lemma 3: For any $a\in[\frac{1}{2},1]$ it holds that

$$\mathbb{P}_\Delta\{\widehat{y}_*(x)\neq y \wedge A_Q(x)\geq a\}\leq\frac{G_Q}{a}. \tag{11}$$

The second bound requires knowledge on G_Q and d_Q. It is a generalization of the $\mathcal{C}$-bound (Lemma 2).

Lemma 4 (General $\mathcal{C}$-bound): If $G_Q<\frac{1}{2}$, then for any $a\in[\frac{1}{2},1]$ it holds that

$$\mathbb{P}_\Delta\{\widehat{y}_*(x)\neq y \wedge A_Q(x)\geq a\}\leq\frac{G_Q(1-G_Q)-\frac{1}{2}d_Q}{G_Q(1-a)+a(a-G_Q)-\frac{1}{2}d_Q} \tag{12}$$

Remark 1: Since $A_Q(x)\geq\frac{1}{2}$ by construction, Lemma 3 provides an alternative proof of Lemma 1.

Remark 2: For $G_Q<\frac{1}{2}$ and $a=\frac{1}{2}$, Lemma 4 reduces to the $\mathcal{C}$-bound of Lemma 2.

Secondly, we provide two novel lower bounds on $\mathbb{P}_\Delta\{A_Q(x)\geq a\}$. The first of which is a function of G_Q only and the second is a function of d_Q only.

Lemma 5: For any $a\in[\frac{1}{2},1-G_Q)$, it holds that

$$\mathbb{P}_\Delta\{A_Q(x)\geq a\}\geq\frac{1-a-G_Q}{1-a}. \tag{13}$$

Lemma 6: For $a\in[\frac{1}{2},\frac{1}{2}+\frac{1}{2}\sqrt{1-2d_Q}]$ it holds that

$$\mathbb{P}_\Delta\{A_Q(x)\geq a\}\geq 1-\frac{d_Q}{2a(1-a)}. \tag{14}$$

Remark 3: Lemma 6 provides a tighter bound than Lemma 5 if and only if $a\geq\frac{d_Q}{2G_Q}$.

Remark 4: Lemma 6 has a wider range of applicability than Lemma 5 because $\frac{1}{2}+\frac{1}{2}\sqrt{1-2d_Q}\geq 1-G_Q$. This is easily proven as follows: by Jensen's inequality, $\mathbb{E}_\Delta\left[M_Q(\delta)^2\right]\geq\mathbb{E}_\Delta\left[M_Q(\delta)\right]^2$; substituting (3) and (5) in this inequality yields the result.

C. Comparison of bounds

In order to provide a numerical illustration of our bounds and compare them with traditional ones, we now provide an example on detection of counterfeit banknotes. We used the 'Banknote Authentication' data set obtained from [20]. This data set contains 1372 data points consisting of $n=4$ features and the corresponding labels (762 genuine and 610 counterfeit). The data set was randomly divided into two sets of equal sizes (for other studies, see, e.g., [21]). One set, the *training set*, was used to train $M=10$ classifiers using GEM [22]. The other set was used as a *validation set* to select the weighting distribution $\mathbb{P}_Q$ (see Section III-D for details). We remark that these choices are just made for the sake of numerical illustration: the results of Theorem 1 and 2 are of general applicability, and we do not aim here at suggesting any specific training, validation or classification scheme.

The empirical estimates of d_Q and G_Q obtained from the validation set were $\hat{d}_Q=0.0169$ and $\hat{G}_Q=0.0107$, respectively. The empirical estimates of the probability of misclassification of the base classifiers ranged between 0.0029 and 0.0235. In Figure 1, where we used the empirical estimates as if these were the true values, we provide a numerical instance of the bound of this paper. It is evident that Theorem 1 can provide better guarantees for a wide range of a than the 2-bound of Lemma 1. Likewise, Theorem 2 is able to provide better guarantees than the $\mathcal{C}$-bound for a wide range of a.

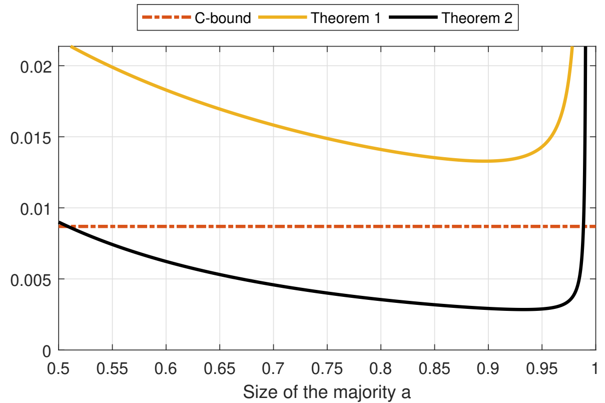

Fig. 1. Comparison of the bounds presented in this work for the simulation example. For illustrative purposes, it is assumed that the empirical estimates are equal to the true values. The dashed line ($\boldsymbol{\mathcal{C}}$-bound) is not a bound on $\boldsymbol{\mathcal{C}_a}$ but on the probability of misclassification of the majority voting classifier $\boldsymbol{B_Q}$. This line is displayed in order to show the comparison of performance increase due to the conditional perspective. Note that the vertical axis is displayed from zero to $\boldsymbol{2\hat{G}_Q}$.

D. Extension to PAC-bounds

It is evident that the empirical estimates of d_Q and G_Q do not have to be equal to the true values. However, by making use of these estimates and statistical learning theory, it is possible to provide bounds on d_Q and G_Q that hold with a high confidence. This allows for a practical use of the main results of this paper. For the numerical example described above, we resort to PAC-Bayesian theorems (for an overview, see [23]). The *prior distribution* $\mathbb{P}_P$ over the classifiers was chosen to be the uniform distribution and the *posterior distribution* $\mathbb{P}_Q$ was obtained by minimisation of the empirical $\mathcal{C}$-bound, which resulted in $\mathrm{KL}(\mathbb{P}_Q\|\mathbb{P}_P)=0.0291$. By Corollary 21 from [14, p. 809] it was found that $G_Q\leq 0.0453$ with a confidence at least $1-10^{-5}$. This upper bound was used in order to obtain a PAC version of the 2-bound and Theorem 1 as shown in Figure 2. For the $\mathcal{C}$-bound and Theorem 2 we used 'PAC-Bound 2' from [14, p. 820] *mutatis mutandis*.

All curves in Figure 2 hold with a probability at least $1-10^{-5}$. Several facts stand out. Firstly, Theorem 1 provides better guarantees than Theorem 2 for $a < 0.68$ in this example. This is due to the fact that Theorem 1 only depends on G_Q as opposed to Theorem 2, which also depends on d_Q. Although this dependency on both G_Q and d_Q resulted in tighter deterministic results (see Section III-C), it degrades more easily in the PAC setting. This demonstrates the relevance of Theorem 1 in addition to Theorem 2. Secondly, for a large range of a, both Theorems 1 and 2 can be used to provide better guarantees on the classification error than either the 2-bound or the $\mathcal{C}$-bound.

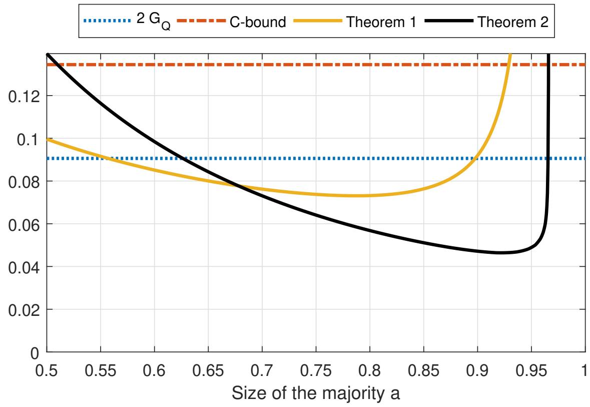

Fig. 2. Comparison of the bounds presented in this work for the simulation example. All bounds hold with a probability at least $\mathbf{1-10^{-5}}$. The dashed lines ($\mathbf{2G_Q}$ and $\mathcal{C}$-bound) are not bounds on $\mathcal{C}_a$, but they are displayed in order to show the comparison of performance increase due to the conditional perspective.

IV. Majority voting with abstention

If the size of the majority is small, it might be of interest to abstain from providing the classification, since, heuristically, a small majority indicates a higher probability of error. We call such a classifier an *abstaining classifier* (due to similarities with the notion of abstaining classifiers presented in, e.g., [24]) and it works as follows. Let $a^* \in [\frac{1}{2}, 1]$ denote the *quorum* or threshold value for the majority size. It is assumed that the base classifiers always provide an output in Y. Then, for any feature vector $x \in X$, such an abstaining classifier returns the label $\widehat{y}_*(x)$ if $A_Q(x) \geq a^*$ and it abstains from returning a label if $A_Q(x) < a^*$. In the latter case, the abstaining classifier is assumed to not be making an error. The expected probability of error of the abstaining classifier is thus

$$\mathbb{P}_\Delta\{\widehat{y}_*(x) \neq y \wedge A_Q(x) \geq a^*\}\,, \tag{15}$$

where $\widehat{y}_*(x)$ is the majority voting classifier as discussed throughout this work. Given that the abstaining classifier does provide a label, the probability of error is $\mathcal{C}_{a^*}$ (see (8)). We would like to stress that the results in this work provide upper bounds for both (15) (see Lemmas 3 and 4) and $\mathcal{C}_{a^*}$ (Theorems 1 and 2). Furthermore, the probability that the quorum is reached and the abstaining classifier thus provides a label is equal to $\mathbb{P}_\Delta\{A_Q(x) \geq a^*\}$. Lemmas 5 and 6 provide lower bounds for this probability. Hence, the results in this work can be used in the design of abstaining classifiers in order to analyse their properties.

A reasonable choice for a^* would be a minimizer of either one of the bounds on $\mathcal{C}_a$ presented in this work (Theorems 1 and 2). As an illustrative example, the following approach uses Theorem 1 to create such an abstaining classifier. Assume that there is a constant G_Q^β available for which it holds that $G_Q \leq G_Q^\beta$ with a probability larger than $1-\beta$, for some $\beta \in [0,1)$. The RHS of (9) is convex in a with the minimum at $a = 1 - \sqrt{G_Q}$. Let us now choose $a^* = \alpha := 1 - \sqrt{G_Q^\beta}$. Using Lemma 3, Lemma 5, and Theorem 1, respectively, it can be shown that

$$\begin{aligned}\mathbb{P}_\Delta\{\widehat{y}_*(x) \neq y \wedge A_Q(x) \geq \alpha\} &\leq_{1-\beta} \frac{G_Q^\beta}{\alpha},\\ \mathbb{P}_\Delta\{A_Q(x) \geq \alpha\} &\geq_{1-\beta} \alpha,\\ \mathbb{P}_\Delta\{\,\widehat{y}_*(x) \neq y \,|\, A_Q(x) \geq \alpha\,\} &\leq_{1-\beta} \frac{G_Q^\beta}{\alpha^2},\end{aligned}$$

where we used the notation '$\leq_{1-\beta}$' and '$\geq_{1-\beta}$' to denote that the particular inequality holds with a confidence of at least $1-\beta$.

V. Conclusion

This work was part of a project aiming at putting data-based decision making on solid theoretical ground, and introduced a novel perspective on the analysis of majority voting schemes in binary classification. Namely, by conditioning on the fact that the majority has a given size, it was possible to provide better guarantees on the probability that the majority voting classifier is correct. The main results of this work were two novel bounds on this conditional probability. It was shown that these bounds can be used in a PAC setting that allows them to be used practically, as illustrated in a numerical example, showing significantly better guarantees than the traditional bounds. Furthermore, we have shown that these results can be used in the analysis of abstaining classifiers.

Appendix I
Proof of Theorem 1

Theorem 1 is a consequence of Lemmas 3 and 5, proven below.

Proof of Lemma 3: From (4) it is clear that $\{\widehat{y}_*(x) \neq y \wedge A_Q(x) \geq a\}$ is equivalent to $\{M_Q(\delta) \leq -(2a-1)\}$ for any $a > \frac{1}{2}$. For $a = \frac{1}{2}$ it holds that

$$\begin{aligned}\mathbb{P}_\Delta\{M_Q(\delta) \leq 0\} = \mathbb{P}_\Delta\{\widehat{y}_*(x) \neq y \wedge A_Q(x) \geq \tfrac{1}{2}\}\\ + \mathbb{P}_\Delta\{\widehat{y}_*(x) = y \wedge A_Q(x) = \tfrac{1}{2}\}\,.\end{aligned}$$

Hence, for $a \in [\frac{1}{2}, 1]$, it holds that

$$\mathbb{P}_\Delta\{\widehat{y}_*(x) \neq y \wedge A_Q(x) \geq a\} \leq \mathbb{P}_\Delta\{M_Q(\delta) \leq -(2a-1)\}\,, \tag{16}$$

which can be rewritten as

$$\mathbb{P}_\Delta\{\widehat{y}_*(x) \neq y \wedge A_Q(x) \geq a\} \leq \mathbb{P}_\Delta\{1 - M_Q(\delta) \geq 2a\}\,.$$

Since $1 - M_Q(\delta)$ is non-negative, we can use Markov's inequality to obtain

$$\mathbb{P}_\Delta\{\widehat{y}_*(x) \neq y \wedge A_Q(x) \geq a\} \leq \tfrac{1-\mathbb{E}_\Delta[M_Q(\delta)]}{2a} = \tfrac{G_Q}{a}\,,$$

where the latter equality is due to (3). ■

We require the following lemma in order to prove Lemma 5.

Lemma 7: It holds that

$$\mathbb{E}_\Delta[A_Q(x)] \geq 1 - G_Q. \tag{17}$$

Proof: It holds that

$$\begin{aligned} G_Q &= \mathbb{E}_\Delta\left[\mathbb{E}_Q[\mathcal{E}(\delta,\widehat{y})]\right] \\ &= \mathbb{E}_\Delta\left[\mathbb{E}_Q[\mathcal{E}(\delta,\widehat{y})(\mathbb{1}\{\widehat{y}\in\mathcal{P}_*(x)\}+\mathbb{1}\{\widehat{y}\notin\mathcal{P}_*(x)\})]\right]. \end{aligned}$$

As discussed in Section II-E, it holds that

$$\begin{aligned} G_Q &= \mathbb{E}_\Delta\left[\mathbb{E}_Q\left[\mathcal{E}(\delta,\widehat{y}_*)\mathbb{1}\{\widehat{y}\in\mathcal{P}_*(x)\}\right.\right. \\ &\qquad\qquad \left.\left.+(1-\mathcal{E}(\delta,\widehat{y}_*))\mathbb{1}\{\widehat{y}\notin\mathcal{P}_*(x)\}\right]\right] \\ &= \mathbb{E}_\Delta[\mathcal{E}(\delta,\widehat{y}_*)A_Q(x)+(1-\mathcal{E}(\delta,\widehat{y}_*))(1-A_Q(x))]. \end{aligned}$$

Since $A_Q(x)\geq\frac{1}{2}$, it is true that $A_Q(x)\geq 1-A_Q(x)$ for all $x\in X$ and thus

$$\begin{aligned} G_Q &\geq \mathbb{E}_\Delta[\mathcal{E}(\delta,\widehat{y}_*)(1-A_Q(x))+(1-\mathcal{E}(\delta,\widehat{y}_*))(1-A_Q(x))] \\ &= \mathbb{E}_\Delta[1-A_Q(x)] = 1-\mathbb{E}_\Delta[A_Q(x)]. \qquad\blacksquare \end{aligned}$$

This lower bound on the expected value of the agreement index allows us to prove Lemma 5.

Proof of Lemma 5: The complementary event satisfies

$$\begin{aligned} \mathbb{P}_\Delta\{A_Q(x)<a\} &\leq \mathbb{P}_\Delta\{A_Q(x)\leq a\} \\ &= \mathbb{P}_\Delta\{1-A_Q(x)\geq 1-a\} \\ &\leq \tfrac{1-\mathbb{E}_\Delta[A_Q(x)]}{1-a} \leq \tfrac{G_Q}{1-a}, \end{aligned}$$

where the second inequality is Markov's inequality ($1-A_Q(x)\geq 0$) and the third inequality is due to (17). ■

The proof of Lemma 5 makes use of (17) to lower bound $\mathbb{E}_\Delta[A_Q(x)]$. Although this is more conservative, in this way it is possible to provide the bounds of Lemma 5 completely in terms of G_Q.

Proof of Theorem 1: The substitution of (11) and (13) in (8) completes the proof of Theorem 1. ■

Appendix II
Proof of Theorem 2

Theorem 2 is a consequence of Lemmas 4 and 6, proven below.

Proof of Lemma 4: For any $a\in[\frac{1}{2},1]$, it holds that

$$\begin{aligned} &\mathbb{P}_\Delta\{\widehat{y}_*(x)\neq y\wedge A_Q(x)\geq a\} \\ &\quad\leq \mathbb{P}_\Delta\{M_Q(\delta)\leq-(2a-1)\} = \mathbb{P}_\Delta\{-M_Q(\delta)\geq 2a-1\} \\ &\quad= \mathbb{P}_\Delta\{-M_Q(\delta)+\mathbb{E}_\Delta[M_Q(\delta)]\geq 2a-1+\mathbb{E}_\Delta[M_Q(\delta)]\} \\ &\quad= \mathbb{P}_\Delta\{-M_Q(\delta)+\mathbb{E}_\Delta[M_Q(\delta)]\geq 2(a-G_Q)\}, \end{aligned}$$

where the first inequality is due to (16) and the third equality is due to (3). Note that $a-G_Q\geq 0$ by assumption. We now use Cantelli's inequality [25] on $-M_Q(\delta)$ to obtain

$$\mathbb{P}_\Delta\{\widehat{y}_*(x)\neq y\wedge A_Q(x)\geq a\}\leq\tfrac{\mathbb{V}_\Delta[M_Q(\delta)]}{\mathbb{V}_\Delta[M_Q(\delta)]+4(a-G_Q)^2}.$$

Substitution of (6) and rewriting yields the claim. ■

Proof of Lemma 6: For $x\in[\frac{1}{2},1]$, the function $x\mapsto x(1-x)$ is non-negative and decreasing. Hence,

$$\begin{aligned} \mathbb{P}_\Delta\{A_Q(x)<a\} &\leq \mathbb{P}_\Delta\{A_Q(x)\leq a\} \\ &= \mathbb{P}_\Delta\{A_Q(x)(1-A_Q(x))\geq a(1-a)\} \\ &\leq \tfrac{\mathbb{E}_\Delta[A_Q(x)(1-A_Q(x))]}{a(1-a)} = \tfrac{d_Q}{2a(1-a)}, \end{aligned}$$

where the second inequality is Markov's and the conclusion is by substitution of (2). ■

Proof of Theorem 2: The substitution of (12) and (14) in (8) completes the proof of Theorem 2. ■

References

[1] A. Care, F. A. Ramponi, and M. C. Campi, "A New Classification Algorithm With Guaranteed Sensitivity and Specificity for Medical Applications," *IEEE Control Syst. Lett.*, vol. 2, no. 3, pp. 393–398, jul 2018.

[2] F. Baronio, M. Baronio, M. C. Campi, A. Carè, S. Garatti, and G. Perone, "Ventricular defibrillation: Classification with G.E.M. and a roadmap for future investigations," in *2017 IEEE 56th Annu. Conf. Decis. Control*, Melbourne, Australia, dec 2017, pp. 2718–2723.

[3] G. Manganini, L. Piroddi, and M. Prandini, "A classification-based approach to the optimal control of affine switched systems," *Proc. 54th IEEE Conf. Decis. Control*, pp. 2963–2968, 2015.

[4] A.-m. Farahmand, D. Precup, A. M. S. Barreto, and M. Ghavamzadeh, "Classification-Based Approximate Policy Iteration," *IEEE Trans. Automat. Contr.*, vol. 60, no. 11, pp. 2989–2993, nov 2015.

[5] E. M. Aiello, C. Toffanin, M. Messori, C. Cobelli, and L. Magni, "Postprandial glucose regulation via KNN meal classification in type 1 diabetes," *IEEE Control Syst. Lett.*, vol. 3, no. 2, pp. 230–235, 2019.

[6] Y. Qin, M. Dong, F. Zhao, R. Langari, and L. Gu, "Road profile classification for vehicle semi-active suspension system based on Adaptive Neuro-Fuzzy Inference System," in *2015 54th IEEE Conf. Decis. Control*, Osaka, dec 2015, pp. 1533–1538.

[7] K. Margellos, M. Prandini, and J. Lygeros, "On the Connection Between Compression Learning and Scenario Based Single-Stage and Cascading Optimization Problems," *IEEE Trans. Automat. Contr.*, vol. 60, no. 10, pp. 2716–2721, oct 2015.

[8] V. Vapnik, "An overview of statistical learning theory," *IEEE Trans. Neural Networks*, vol. 10, no. 5, pp. 988–999, 1999.

[9] O. Bousquet, S. Boucheron, and G. Lugosi, *Introduction to Statistical Learning Theory*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2004, pp. 169–207.

[10] R. E. Schapire, Y. Freund, P. Bartlett, and W. S. Lee, "Boosting the margin: a new explanation for the effectiveness of voting methods," *Ann. Stat.*, vol. 26, no. 5, pp. 1651–1686, oct 1998.

[11] L. Breiman, "Random Forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.

[12] M. Zhu, "Use of majority votes in statistical learning," *Wiley Interdiscip. Rev. Comput. Stat.*, vol. 7, no. 6, pp. 357–371, nov 2015.

[13] D. A. McAllester, "Some PAC-Bayesian theorems," *Mach. Learn.*, vol. 37, no. 3, pp. 355–363, 1999.

[14] P. Germain, A. Lacasse, F. Laviolette, M. Marchand, and J. F. Roy, "Risk bounds for the majority vote: From a PAC-Bayesian analysis to a learning algorithm," *J. Mach. Learn. Res.*, vol. 16, pp. 787–860, 2015.

[15] J. Langford and J. Shawe-Taylor, "PAC-Bayes & Margins," *Adv. Neural Inf. Process. Syst.*, pp. 439–446, 2003.

[16] A. Cobbenhagen, A. Carè, M. Campi, F. Ramponi, and W. Heemels, "Consensus and Reliability: The Case of Two Binary Classifiers," *IFAC-PapersOnLine*, vol. 52, no. 20, pp. 73–78, 2019.

[17] O. Catoni, *Statistical Learning Theory and Stochastic Optimization*, ser. Lecture Notes in Mathematics, J. Picard, Ed. Berlin, Heidelberg: Springer Berlin Heidelberg, 2004, vol. 1851.

[18] A. Carè, M. C. Campi, F. A. Ramponi, S. Garatti, and A. T. J. R. Cobbenhagen, "A study on majority-voting classifiers with guarantees on the probabilty of error," *IFAC World Congress 2020*, 2020.

[19] A. Lacasse, F. Laviolette, M. Marchand, P. Germain, and N. Usunier, "PAC-Bayes Bounds for the Risk of the Majority Vote and the Variance of the Gibbs Classifier," in *Adv. Neural Inf. Process. Syst. 19*. The MIT Press, 2007, no. 1, pp. 769–776.

[20] D. Dua and C. Graff, "UCI Machine Learning Repository," 2017. [Online]. Available: http://archive.ics.uci.edu/ml

[21] A. Carè, S. Garatti, and M. C. Campi, "A coverage theory for least squares," *J. R. Stat. Soc. Ser. B Stat. Methodol.*, vol. 79, no. 5, pp. 1367–1389, 2017.

[22] M. C. Campi, "Classification with guaranteed probability of error," *Mach. Learn.*, vol. 80, pp. 63–84, jul 2010.

[23] B. Guedj, "A Primer on PAC-Bayesian Learning," 2019. [Online]. Available: arXiv:1901.05353[stat.ML]

[24] T. Pietraszek, "On the use of ROC analysis for the optimization of abstaining classifiers," *Mach. Learn.*, vol. 68, no. 2, pp. 137–169, 2007.

[25] B. K. Ghosh, "Probability inequalities related to Markov's theorem," *Am. Stat.*, vol. 56, no. 3, pp. 186–190, 2002.