Journal of Machine Learning Research 16 (2015) 787-860 Submitted 5/13; Revised 9/14; Published 4/15

Risk Bounds for the Majority Vote: From a PAC-Bayesian Analysis to a Learning Algorithm

Pascal Germain PASCAL.GERMAIN@IFT.ULAVAL.CA
Alexandre Lacasse ALEXANDRE.LACASSE@IFT.ULAVAL.CA
François Laviolette FRANCOIS.LAVIOLETTE@IFT.ULAVAL.CA
Mario Marchand MARIO.MARCHAND@IFT.ULAVAL.CA
Jean-Francis Roy JEAN-FRANCIS.ROY@IFT.ULAVAL.CA
Département d'informatique et de génie logiciel
Université Laval
Québec, Canada, G1V 0A6
* All authors contributed equally to this work.

Editor: Koby Crammer

Abstract

We propose an extensive analysis of the behavior of majority votes in binary classification. In particular, we introduce a risk bound for majority votes, called the $\mathcal{C}$-bound, that takes into account the average quality of the voters and their average disagreement. We also propose an extensive PAC-Bayesian analysis that shows how the $\mathcal{C}$-bound can be estimated from various observations contained in the training data. The analysis intends to be self-contained and can be used as introductory material to PAC-Bayesian statistical learning theory. It starts from a general PAC-Bayesian perspective and ends with uncommon PAC-Bayesian bounds. Some of these bounds contain no Kullback-Leibler divergence and others allow kernel functions to be used as voters (via the sample compression setting). Finally, out of the analysis, we propose the MinCq learning algorithm that basically minimizes the $\mathcal{C}$-bound. MinCq reduces to a simple quadratic program. Aside from being theoretically grounded, MinCq achieves state-of-the-art performance, as shown in our extensive empirical comparison with both AdaBoost and the Support Vector Machine.

Keywords: majority vote, ensemble methods, learning theory, PAC-Bayesian theory, sample compression

1. Previous Work and Implementation

This paper can be considered as an extended version of Lacasse et al. (2006) and Laviolette et al. (2011), and also contains ideas from Laviolette and Marchand (2005, 2007) and Germain et al. (2009, 2011). We unify this previous work, revise the mathematical approach, add new results and extend empirical experiments.

The source code to compute the various PAC-Bayesian bounds presented in this paper and the implementation of the MinCq learning algorithm is available at:

`http://graal.ift.ulaval.ca/majorityvote/`

2. Introduction

In binary classification, many state-of-the-art algorithms output prediction functions that can be seen as a majority vote of "simple" classifiers. Firstly, ensemble methods such as Bagging (Breiman, 1996), Boosting (Schapire and Singer, 1999) and Random Forests (Breiman, 2001) are well-known examples of learning algorithms that output majority votes. Secondly, majority votes are also central in the Bayesian approach (see Gelman et al., 2004, for an introductory text); in this setting, the majority vote is generally called the *Bayes Classifier*. Thirdly, it is interesting to point out that classifiers produced by kernel methods, such as the Support Vector Machine (SVM) (Cortes and Vapnik, 1995), can also be viewed as majority votes. Indeed, to classify an example x, the SVM classifier computes

$$\operatorname{sgn}\left(\sum_{i=1}^{|S|} \alpha_i \, y_i \, k(x_i, x)\right), \tag{1}$$

where $k(\cdot,\cdot)$ is a kernel function, and the input-output pairs (x_i, y_i) represent the examples from the training set S. Thus, one can interpret each $y_i \, k(x_i, \cdot)$ as a voter that chooses (with confidence level $|k(x_i, x)|$) between two alternatives ("positive" or "negative"), and α_i as the respective weight of this voter in the majority vote. Then, if the total confidence-multiplied weight of each voter that votes positive is greater than the total confidence-multiplied weight of each voter that votes negative, the classifier outputs a $+1$ label (and a -1 label in the opposite case). Similarly, each *neuron* of the last layer of an artificial neural network can be interpreted as a majority vote, since it outputs a real value given by $K(\sum_i w_i g_i(x))$ for some *activation function* K.[1]

In practice, it is well known that the classifier output by each of these learning algorithms performs much better than any of its voters individually. Indeed, voting can dramatically improve performance when the "community" of classifiers tends to compensate for individual errors. In particular, this phenomenon explains the success of Boosting algorithms (*e.g.*, Schapire et al., 1998). The first aim of this paper is to explore how bounds on the generalized risk of the majority vote are not only able to theoretically justify learning algorithms but also to detect when the voted combination provably outperforms the average of its voters. We expect that this study of the behavior of a majority vote should improve the understanding of existing learning algorithms and even lead to new ones. We indeed present a learning algorithm based on these ideas at the end of the paper.

The PAC-Bayesian theory is a well-suited approach to analyze majority votes. Initiated by McAllester (1999), this theory aims to provide Probably Approximately Correct guarantees (PAC guarantees) to "Bayesian-like" learning algorithms. Within this approach, one considers a *prior*[2] *distribution* P over a space of classifiers that characterizes its prior belief about good classifiers (before the observation of the data) and a *posterior distribution* Q (over the same space of classifiers) that takes into account the additional information provided by the training data. The classical PAC-Bayesian approach indirectly bounds the risk

[1] In this case, each voter g_i has incoming weights which are also learned (often by back propagation of errors) together with the weights w_i. The analysis presented in this paper considers fixed voters. Thus, the PAC-Bayesian theory for artificial neural networks remains to be done. Note however that the recent work by McAllester (2013) provides a first step in that direction.

[2] Priors have been used for many years in statistics. The priors in this paper have only indirect links with the *Bayesian priors*. We nevertheless use this language, since it comes from previous work.

of a Q-weighted majority vote by bounding the risk of an associate (stochastic) classifier, called the *Gibbs classifier*. A remarkable result, known as the "PAC-Bayesian Theorem", provides a risk bound for the "true" risk of the Gibbs classifier, by considering the empirical risk of this Gibbs classifier on the training data and the Kullback-Leibler divergence between a posterior distribution Q and a prior distribution P. It is well known (Langford and Shawe-Taylor, 2002; McAllester, 2003b; Germain et al., 2009) that the risk of the (deterministic) majority vote classifier is upper-bounded by twice the risk of the associated (stochastic) Gibbs classifier. Unfortunately, and especially if the involved voters are weak, this indirect bound on the majority vote classifier is far from being tight, even if the PAC-Bayesian bound itself generally gives a tight bound on the risk of the Gibbs classifier. In practice, as stated before, the "community" of classifiers can act in such a way as to compensate for individual errors. When such compensation occurs, the risk of the majority vote is then much lower than the Gibbs risk itself and, a fortiori, much lower than twice the Gibbs risk. By limiting the analysis to Gibbs risk only, the commonly used PAC-Bayesian framework is unable to evaluate whether or not this compensation occurs. Consequently, this framework cannot help in producing highly accurate voted combinations of classifiers when these classifiers are individually weak.

In this paper, we tackle this problem by studying the margin of the majority vote as a random variable. The first and second moments of this random variable are respectively linked with the risk of the Gibbs classifier and the expected disagreement between the voters of the majority vote. As we will show, the well-known factor of two used to bound the risk of the majority vote is recovered by applying Markov's inequality to the first moment of the margin. Based on this observation, we show that a tighter bound, that we call the $\mathcal{C}$-bound, is obtained by considering the first two moments of the margin, together with Chebyshev's inequality.

Section 4 presents, in a more detailed way, the work on the $\mathcal{C}$-bound originally presented in Lacasse et al. (2006). We then present both theoretical and empirical studies that show that the $\mathcal{C}$-bound is an accurate indicator of the risk of the majority vote. We also show that the $\mathcal{C}$-bound can be smaller than the risk of the Gibbs classifier and can even be arbitrarily close to zero even if the risk of the Gibbs classifier is close to 1/2. This indicates that the $\mathcal{C}$-bound can effectively capture the compensation of the individual errors made by the voters.

We then develop PAC-Bayesian guarantees on the $\mathcal{C}$-bound in order to obtain an upper bound on the risk of the majority vote based on empirical observations. Section 5 presents a general approach of the PAC-Bayesian theory by which we recover the most commonly used forms of the bounds of McAllester (1999, 2003a) and Langford and Seeger (2001); Seeger (2002); Langford (2005). Thereafter, we extend the theory to obtain upper bounds on the $\mathcal{C}$-bound in two different ways. The first method is to separately bound the risk of the Gibbs classifier and the expected disagreement—which are the two fundamental ingredients that are present in the $\mathcal{C}$-bound. Since the expected disagreement does not rely on labels, this strategy is well-suited for the semi-supervised learning framework. The second method directly bounds the $\mathcal{C}$-bound and empirically improves the achievable bounds in the supervised learning framework.

Sections 6 and 7 bring together relatively new PAC-Bayesian ideas that allow us, for one part, to derive a PAC-Bayesian bound that does not rely on the Kullback-Leibler divergence between the prior and posterior distributions (as in Catoni, 2007; Germain et al., 2011; Laviolette et al., 2011) and, for the other part, to extend the bound to the case where the voters are defined using elements of the training data, *e.g.*, voters defined by kernel functions $y_i k(x_i, \cdot)$. This second approach is based on the sample compression theory (Floyd and Warmuth, 1995; Laviolette and Marchand, 2007; Germain et al., 2011). In PAC-Bayesian theory, the sample compression approach is *a priori* problematic, since a PAC-Bayesian bound makes use of a prior distribution on the set of all voters that has to be defined before observing the data. If the voters themselves are defined using a part of the data, there is an apparent contradiction that has to be overcome.

Based on the foregoing, a learning algorithm, that we call MinCq, is presented in Section 8. The algorithm basically minimizes the $\mathcal{C}$-bound, but in a particular way that is, inter alia, justified by the PAC-Bayesian analysis of Sections 6 and 7. This algorithm was originally presented in Laviolette et al. (2011). Given a set of voters (either classifiers or kernel functions), MinCq builds a majority vote classifier by finding the posterior distribution Q on the set of voters that minimizes the $\mathcal{C}$-bound. Hence, MinCq takes into account not only the overall quality of the voters, but also their respective disagreements. In this way, MinCq builds a "community" of voters that can compensate for their individual errors. Even though the $\mathcal{C}$-bound consists of a relatively complex quotient, the MinCq learning algorithm reduces to a simple quadratic program. Moreover, extensive empirical experiments confirm that MinCq is very competitive when compared with AdaBoost (Schapire and Singer, 1999) and the Support Vector Machine (Cortes and Vapnik, 1995).

In Section 9, we conclude by pointing out recent work that uses the PAC-Bayesian theory to tackle more sophisticated machine learning problems.

3. Basic Definitions

We consider classification problems where the input space $\mathcal{X}$ is an arbitrary set and the output space is a discrete set denoted $\mathcal{Y}$. An *example* (x, y) is an input-output pair where $x \in \mathcal{X}$ and $y \in \mathcal{Y}$. A *voter* is a function $\mathcal{X} \to \overline{\mathcal{Y}}$ for some output space $\overline{\mathcal{Y}}$ related to $\mathcal{Y}$. Unless otherwise specified, we consider the binary classification problem where $\mathcal{Y} = \{-1, 1\}$ and then we either consider $\overline{\mathcal{Y}}$ as $\mathcal{Y}$ itself, or its convex hull $[-1, +1]$. In this paper, we also use the following convention: f denotes a real-valued voter (*i.e.*, $\overline{\mathcal{Y}} = [-1, 1]$), and h denotes a binary-valued voter (*i.e.*, $\overline{\mathcal{Y}} = \{-1, 1\}$). Note that this notion of voters is quite general, since any uniformly bounded real-valued set of functions can be viewed as a set of voters when properly normalized.

We consider learning algorithms that construct majority votes based on a (finite) set $\mathcal{H}$ of voters. Given any $x \in \mathcal{X}$, the output $B_Q(x)$ of a Q-weighted majority vote classifier B_Q (sometimes called the *Bayes classifier*) is given by

$$B_Q(x) \overset{\text{def}}{=} \operatorname{sgn}\left[\mathop{\mathbf{E}}_{f \sim Q} f(x)\right], \tag{2}$$

where $\operatorname{sgn}(a) = 1$ if $a > 0$, $\operatorname{sgn}(a) = -1$ if $a < 0$, and $\operatorname{sgn}(0) = 0$.

Thus, in case of a tie in the majority vote – *i.e.*, $\mathbf{E}_{f\sim Q} f(x) = 0$ –, we consider that the majority vote classifier abstains – *i.e.*, $B_Q(x) = 0$. There are other possible ways to handle this particular case. In this paper, we choose to define $\operatorname{sgn}(0) = 0$ because it simplifies the forthcoming analysis.

We adopt the PAC setting where each example (x, y) is drawn i.i.d. according to a fixed, but unknown, probability distribution D on $\mathcal{X} \times \mathcal{Y}$. The *training set* of m examples is denoted by $S = \langle\, (x_1, y_1), \ldots, (x_m, y_m)\, \rangle \sim D^m$. Throughout the paper, D' generically represents either the true (and unknown) distribution D, or its empirical counterpart U_S (*i.e.*, the uniform distribution over the training set S). Moreover, for notational simplicity, we often replace U_S by S.

In order to quantify the accuracy of a voter, we use a *loss function* $\mathcal{L} : \overline{\mathcal{Y}} \times \mathcal{Y} \to [0, 1]$. The PAC-Bayesian theory traditionally considers majority votes of binary voters of the form $h : \mathcal{X} \to \{-1, 1\}$, and the *zero-one loss* $\mathcal{L}_{01}\big(h(x), y\big) \overset{\text{def}}{=} I\big(h(x) \neq y\big)$, where $I(a) = 1$ if predicate a is true and 0 otherwise.

The extension of the zero-one loss to real-valued voters (of the form $f : \mathcal{X} \to [-1, 1]$) is given by the following definition.

Definition 1 In the (more general) case where voters are functions $f : \mathcal{X} \to [-1, 1]$, the zero-one loss $\mathcal{L}_{01}$ is defined by

$$\mathcal{L}_{01}\big(f(x), y\big) \overset{\text{def}}{=} I\big(y \cdot f(x) \leq 0\big).$$

Hence, a voter abstention – *i.e.*, when $f(x)$ outputs exactly 0 – results in a loss of 1. Clearly, other choices are possible for this particular case.[3]

In this paper, we also consider the *linear loss* $\mathcal{L}_\ell$ defined as follows.

Definition 2 Given a voter $f : \mathcal{X} \to [-1, 1]$, the linear loss $\mathcal{L}_\ell$ is defined by

$$\mathcal{L}_\ell\big(f(x), y\big) \overset{\text{def}}{=} \frac{1}{2}\Big(1 - y \cdot f(x)\Big).$$

Note that the linear loss is equal to the zero-one loss when the output space is binary. That is, for any $(h(x), y) \in \{-1, 1\}^2$, we always have

$$\mathcal{L}_\ell\big(h(x), y\big) \;=\; \mathcal{L}_{01}\big(h(x), y\big)\,, \tag{3}$$

because $\mathcal{L}_\ell\big(h(x), y\big) = 1$ if $h(x) \neq y$, and $\mathcal{L}_\ell\big(h(x), y\big) = 0$ if $h(x) = y$. Hence, we generalize all definitions implying classifiers to voters using the equality of Equation (3) as an inspiration. Figure 1 illustrates the difference between the zero-one loss and the linear loss for real-valued voters. Remember that in the case $y\, f(x) = 0$, the loss is 1 (see Definition 1).

Definition 3 Given a loss function $\mathcal{L}$ and a voter f, the *expected loss* $\mathbb{E}^{\mathcal{L}}_{D'}(f)$ of f relative to distribution D' is defined as

$$\mathbb{E}^{\mathcal{L}}_{D'}(f) \overset{\text{def}}{=} \mathop{\mathbf{E}}_{(x,y)\sim D'} \mathcal{L}\big(f(x), y\big)\,.$$

[3] As an example, when $f(x)$ outputs 0, the loss may be 1/2. However, we choose for this unlikely event the worst loss value – *i.e.*, $\mathcal{L}_{01}(0, y) = 1$ – because it simplifies the majority vote analysis.

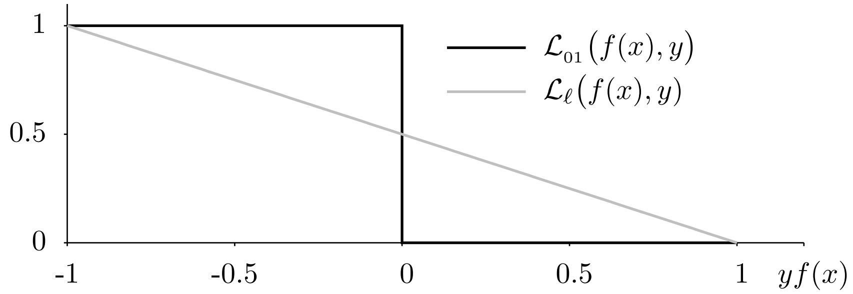


Figure 1: The zero-one loss $\mathcal{L}_{01}$ and the linear loss $\mathcal{L}_{\ell}$ as a function of $yf(x)$.

In particular, the *empirical expected loss* on a training set S is given by

$$\mathbb{E}_S^{\mathcal{L}}(f) \;=\; \frac{1}{m}\sum_{i=1}^{m}\mathcal{L}\big(f(x_i),y_i\big)\,.$$

We therefore define the risk of the majority vote $R_{D'}(B_Q)$ as follows.

Definition 4 For any probability distribution Q on a set of voters, the *Bayes risk* $R_{D'}(B_Q)$, also called *risk of the majority vote*, is defined as the expected zero-one loss of the majority vote classifier B_Q relative to D'. Hence,

$$R_{D'}(B_Q) \;\stackrel{\text{def}}{=}\; \mathbb{E}_{D'}^{\mathcal{L}_{01}}(B_Q) \;=\; \mathop{\mathbf{E}}_{(x,y)\sim D'} I\Big(B_Q(x)\neq y\Big) \;=\; \mathop{\mathbf{E}}_{(x,y)\sim D'} I\Big(\mathop{\mathbf{E}}_{f\sim Q}\, y\cdot f(x)\leq 0\Big)\,.$$

Remember from the definition of B_Q (Equation 2) that the majority vote classifier abstains in the case of a tie on an example (x,y). Therefore, the above Definition 4 implies that the Bayes risk is 1 in this case, as $R_{\langle(x,y)\rangle}(B_Q)\!=\!\mathcal{L}_{01}(0,y)\!=\!1$. In practice, a tie in the vote is a rare event, especially if there are many voters.

The output of the deterministic majority vote classifier B_Q is closely related to the output of a stochastic classifier called the *Gibbs classifier*. To classify an input example x, the Gibbs classifier G_Q randomly chooses a voter f according to Q and returns $f(x)$. Note the stochasticity of the Gibbs classifier: it can output different values when given the same input x twice. We will see later how the link between B_Q and G_Q is used in the PAC-Bayesian theory.

In the case of binary voters, the Gibbs risk corresponds to the probability that G_Q misclassifies an example of distribution D'. Hence,

$$R_{D'}(G_Q) \;=\; \mathop{\Pr}_{\substack{(x,y)\sim D'\\ h\sim Q}}\big(h(x)\neq y\big) \;=\; \mathop{\mathbf{E}}_{h\sim Q}\,\mathbb{E}_{D'}^{\mathcal{L}_{01}}(h) \;=\; \mathop{\mathbf{E}}_{(x,y)\sim D'}\;\mathop{\mathbf{E}}_{h\sim Q} I\big(h(x)\neq y\big)\,.$$

In order to handle real-valued voters, we generalize the Gibbs risk as follows.

Definition 5 For any probability distribution Q on a set of voters, the *Gibbs risk* $R_{D'}(G_Q)$ is defined as the expected linear loss of the Gibbs classifier G_Q relative to D'. Hence,

$$R_{D'}(G_Q) \;\stackrel{\text{def}}{=}\; \mathop{\mathbf{E}}_{f\sim Q}\,\mathbb{E}_{D'}^{\mathcal{L}_\ell}(f) \;=\; \frac{1}{2}\left(1-\mathop{\mathbf{E}}_{(x,y)\sim D'}\;\mathop{\mathbf{E}}_{f\sim Q}\, y\cdot f(x)\right)\,.$$

Remark 6 It is well known in the PAC-Bayesian literature (*e.g.*, Langford and Shawe-Taylor, 2002; McAllester, 2003b; Germain et al., 2009) that the Bayes risk $R_{D'}(B_Q)$ is bounded by twice the Gibbs risk $R_{D'}(G_Q)$. This statement extends to our more general definition of the Gibbs risk (Definition 5).

Proof Let $(x, y) \in \mathcal{X} \times \{-1, 1\}$ be any example. We claim that

$$R_{\langle(x,y)\rangle}(B_Q) \;\leq\; 2\, R_{\langle(x,y)\rangle}(G_Q)\,. \tag{4}$$

Notice that $R_{\langle(x,y)\rangle}(B_Q)$ is either 0 or 1 depending of the fact that B_Q errs or not on (x, y). In the case where $R_{\langle(x,y)\rangle}(B_Q) = 0$, Equation (4) is trivially true. If $R_{\langle(x,y)\rangle}(B_Q) = 1$, we know by the last equality of Definition 4 that $\mathop{\mathbf{E}}_{f\sim Q} y\cdot f(x) \leq 0$. Therefore, Definition 5 gives

$$2\cdot R_{\langle(x,y)\rangle}(G_Q) \;=\; 2\cdot\frac{1}{2}\left(1 - \mathop{\mathbf{E}}_{f\sim Q} y\cdot f(x)\right) \;\geq\; 1 = R_{\langle(x,y)\rangle}(B_Q)\,,$$

which proves the claim.

Now, by taking the expectation according to $(x, y) \sim D'$ on each side of Equation (4), we obtain

$$R_{D'}(B_Q) \;=\; \mathop{\mathbf{E}}_{(x,y)\sim D'} R_{\langle(x,y)\rangle}(B_Q) \;\leq\; \mathop{\mathbf{E}}_{(x,y)\sim D'} 2\, R_{\langle(x,y)\rangle}(G_Q) \;=\; 2\, R_{D'}(G_Q)\,,$$

as wanted. ■

Thus, PAC-Bayesian bounds on the risk of the majority vote are usually bounds on the Gibbs risk, multiplied by a factor of two. Even if this type of bound can be tight in some situations, the factor two can also be misleading. Langford and Shawe-Taylor (2002) have shown that under some circumstances, the factor of two can be reduced to $(1 + \epsilon)$. Nevertheless, distributions Q on voters giving $R_{D'}(G_Q) \gg R_{D'}(B_Q)$ are common. The extreme case happens when the expected linear loss on each example is just below one half – *i.e.*, for all (x, y), $\mathbf{E}_{f\sim Q}\, y\, f(x) = \frac{1}{2}{-}\epsilon$ –, leading to a perfect majority vote classifier but an almost inaccurate Gibbs classifier. Indeed, we have $R_{D'}(G_Q) = \frac{1}{2}{-}\epsilon$ and $R_{D'}(B_Q) = 0$. Therefore, in this circumstance, the bound $R_{D'}(B_Q) \leq 1{-}2\epsilon$, given by Remark 6, fails to represent the perfect accuracy of the majority vote. This problem is due to the fact that the Gibbs risk only considers the loss of the average output of the population of voters. Hence, the bound of Remark 6 states that the majority vote is weak whenever every individual voter is weak. The bound cannot capture the fact that it might happen that the "community" of voters compensates for individual errors. To overcome this lacuna, we need a bound that compares the output of voters between them, not only the average quality of each voter taken individually.

We can compare the output of binary voters by considering the probability of disagreement between them:

$$\begin{aligned}
\Pr_{\substack{x\sim D'_{\mathcal{X}}\\ h_1,h_2\sim Q}}\Big(h_1(x) \neq h_2(x)\Big) \;&=\; \mathop{\mathbf{E}}_{x\sim D'_{\mathcal{X}}}\,\mathop{\mathbf{E}}_{h_1\sim Q}\,\mathop{\mathbf{E}}_{h_2\sim Q}\, I\Big(h_1(x) \neq h_2(x)\Big)\\
&=\; \mathop{\mathbf{E}}_{x\sim D'_{\mathcal{X}}}\,\mathop{\mathbf{E}}_{h_1\sim Q}\,\mathop{\mathbf{E}}_{h_2\sim Q}\, I\Big(h_1(x)\cdot h_2(x) \neq 1\Big)\\
&=\; \mathop{\mathbf{E}}_{x\sim D'_{\mathcal{X}}}\,\mathop{\mathbf{E}}_{h_1\sim Q}\,\mathop{\mathbf{E}}_{h_2\sim Q}\, \mathcal{L}_{01}\big(\,h_1(x){\cdot}h_2(x)\,,\,1\,\big)\,,
\end{aligned}$$

where $D'_{\mathcal{X}}$ denotes the marginal on $\mathcal{X}$ of distribution D'. Definition 7 extends this notion of disagreement to real-valued voters.

Definition 7 For any probability distribution Q on a set of voters, the *expected disagreement* $d_Q^{D'}$ relative to D' is defined as

$$\begin{aligned} d_Q^{D'} &\overset{\text{def}}{=} \mathop{\mathbf{E}}_{x\sim D'_{\mathcal{X}}} \mathop{\mathbf{E}}_{f_1\sim Q} \mathop{\mathbf{E}}_{f_2\sim Q} \mathcal{L}_\ell\left(f_1(x)\cdot f_2(x)\,,\,1\right) \\ &= \frac{1}{2}\left(1-\mathop{\mathbf{E}}_{x\sim D'_{\mathcal{X}}} \mathop{\mathbf{E}}_{f_1\sim Q} \mathop{\mathbf{E}}_{f_2\sim Q} 1\cdot f_1(x)\cdot f_2(x)\right) \\ &= \frac{1}{2}\left(1-\mathop{\mathbf{E}}_{x\sim D'_{\mathcal{X}}} \left[\mathop{\mathbf{E}}_{f\sim Q} f(x)\right]^2\right). \end{aligned}$$

Notice that the value of $d_Q^{D'}$ does not depend on the labels y of the examples $(x, y) \sim D'$. Therefore, we can estimate the expected disagreement with unlabeled data.

4. Bounds on the Risk of the Majority Vote

The aim of this section is to introduce the $\mathcal{C}$-bound, which upper-bounds the risk of the majority vote (Definition 4) based on the Gibbs risk (Definition 5) and the expected disagreement (Definition 7). We start by studying the margin of a majority vote as a random variable (Section 4.1). From the first moment of the margin, we easily recover the well-known bound of twice the Gibbs risk presented by Remark 6 (Section 4.2). We therefore suggest extending this analysis to the second moment of the margin to obtain the $\mathcal{C}$-bound (Section 4.3). Finally, we present some statistical properties of the $\mathcal{C}$-bound (Section 4.4) and an empirical study of its predictive power (Section 4.5).

4.1 The Margin of the Majority Vote and its Moments

The bounds on the risk of a majority vote classifier proposed in this section result from the study of the weighted margin of the majority vote as a random variable.

Definition 8 Let $M_Q^{D'}$ be the random variable that, given any example (x, y) drawn according to D', outputs the *margin* of the majority vote B_Q on that example, which is

$$M_Q(x,y) \overset{\text{def}}{=} \mathop{\mathbf{E}}_{f\sim Q} y\cdot f(x)\,.$$

From Definitions 4 and 8, we have the following nice property:[4]

$$R_{D'}(B_Q) \;=\; \Pr_{(x,y)\sim D'}\Big(M_Q(x,y)\leq 0\Big)\,. \tag{5}$$

[4] Note that for another choice of the zero-one loss definition (Definition 1), the tie in the majority vote – *i.e.*, when $M_Q(x, y) = 0$ – would have been more complicated to handle, and the statement should have been relaxed to

$$\Pr_{(x,y)\sim D'}\Big(M_Q(x,y)<0\Big) \;\leq\; R_{D'}(B_Q) \;\leq\; \Pr_{(x,y)\sim D'}\Big(M_Q(x,y)\leq 0\Big)\,.$$

The margin is not only related to the risk of the majority vote, but also to Gibbs risk. For that purpose, let us consider the first moment $\mu_1(M_Q^{D'})$ of the random variable $M_Q^{D'}$ which is defined as

$$\mu_1(M_Q^{D'}) \overset{\text{def}}{=} \mathop{\mathbf{E}}_{(x,y)\sim D'} M_Q(x,y)\,. \tag{6}$$

We can now rewrite the Gibbs risk (Definition 5) as a function of $\mu_1(M_Q^{D'})$, since

$$\begin{aligned} R_{D'}(G_Q) &= \mathop{\mathbf{E}}_{f\sim Q} \mathbb{E}_{D'}^{\mathcal{L}_\ell}(f) \;=\; \frac{1}{2}\left(1 - \mathop{\mathbf{E}}_{(x,y)\sim D'}\,\mathop{\mathbf{E}}_{f\sim Q}\, y\cdot f(x)\right)\\ &= \frac{1}{2}\Big(1 - \mathop{\mathbf{E}}_{(x,y)\sim D'} M_Q(x,y)\Big)\\ &= \frac{1}{2}\Big(1-\mu_1(M_Q^{D'})\Big)\,. \end{aligned} \tag{7}$$

Similarly, we can rewrite the expected disagreement as a function of the second moment of the margin. We use $\mu_2(M_Q^{D'})$ to denote the second moment. Since $y \in \{-1,1\}$ and, therefore, $y^2 = 1$, the second moment of the margin does not rely on labels. Indeed, we have

$$\begin{aligned} \mu_2(M_Q^{D'}) &\overset{\text{def}}{=} \mathop{\mathbf{E}}_{(x,y)\sim D'} \Big[M_Q(x,y)\Big]^2 \qquad (8)\\ &= \mathop{\mathbf{E}}_{(x,y)\sim D'} y^2\cdot\Big[\mathop{\mathbf{E}}_{f\sim Q} f(x)\Big]^2\\ &= \mathop{\mathbf{E}}_{x\sim D'_{\mathcal{X}}} \Big[\mathop{\mathbf{E}}_{f\sim Q} f(x)\Big]^2\,. \end{aligned}$$

Hence, from the last equality and Definition 7, the expected disagreement can be expressed as

$$\begin{aligned} d_Q^{D'} &= \frac{1}{2}\left(1 - \mathop{\mathbf{E}}_{x\sim D'_{\mathcal{X}}} \Big[\mathop{\mathbf{E}}_{f\sim Q} f(x)\Big]^2\right)\\ &= \frac{1}{2}\Big(1-\mu_2(M_Q^{D'})\Big)\,. \end{aligned} \tag{9}$$

Equation (9) shows that $0 \le d_Q^{D'} \le 1/2$, since $0 \le \mu_2(M_Q^{D'}) \le 1$. Furthermore, we can upper-bound the disagreement more tightly than simply saying it is at most 1/2 by making use of the value of the Gibbs risk. To do so, let us write the variance of the margin as

$$\begin{aligned} \mathrm{Var}(M_Q^{D'}) &\overset{\text{def}}{=} \mathop{\mathbf{Var}}_{(x,y)\sim D'} \big(M_Q(x,y)\big)\\ &= \mu_2(M_Q^{D'}) - \big(\mu_1(M_Q^{D'})\big)^2\,. \end{aligned} \tag{10}$$

Therefore, as the variance cannot be negative, it follows that

$$\mu_2(M_Q^{D'}) \;\ge\; \big(\mu_1(M_Q^{D'})\big)^2\,,$$

which implies that

$$1-2\cdot d_Q^{D'} \quad \geq \quad (1-2\cdot R_{D'}(G_Q))^2\,. \tag{11}$$

Easy calculation then gives the desired bound of $d_Q^{D'}$ (that is based on the Gibbs risk):

$$d_Q^{D'} \quad \leq \quad 2\cdot R_{D'}(G_Q)\cdot\big(1-R_{D'}(G_Q)\big)\,. \tag{12}$$

We therefore have the following proposition.

Proposition 9 *For any distribution Q on a set of voters and any distribution D' on $\mathcal{X}\times\{-1,1\}$, we have*

$$d_Q^{D'} \;\leq\; 2\cdot R_{D'}(G_Q)\cdot\big(1-R_{D'}(G_Q)\big) \;\leq\; \frac{1}{2}\,.$$

Moreover, if $d_Q^{D'}=\frac{1}{2}$ then $R_{D'}(G_Q)=\frac{1}{2}$.

Proof Equation (12) gives the first inequality. The rest of the proposition directly follows from the fact that $f(x)=2x(1-x)$ is a parabola whose (unique) maximum is at the point $(\frac{1}{2},\frac{1}{2})$. ■

4.2 Rediscovering the bound $R_{D'}(B_Q)\leq 2\cdot R_{D'}(G_Q)$

The well-known factor of two with which one can transform a bound on the Gibbs risk $R_{D'}(G_Q)$ into a bound on the risk $R_{D'}(B_Q)$ of the majority vote is usually justified by an argument similar to the one given in Remark 6. However, as shown by the proof of Proposition 10, the result can also be obtained by considering that the risk of the majority vote is the probability that the margin $M_Q^{D'}$ is lesser than or equal to zero (Equation 5) and by simply applying Markov's inequality (Lemma 46, provided in Appendix A).

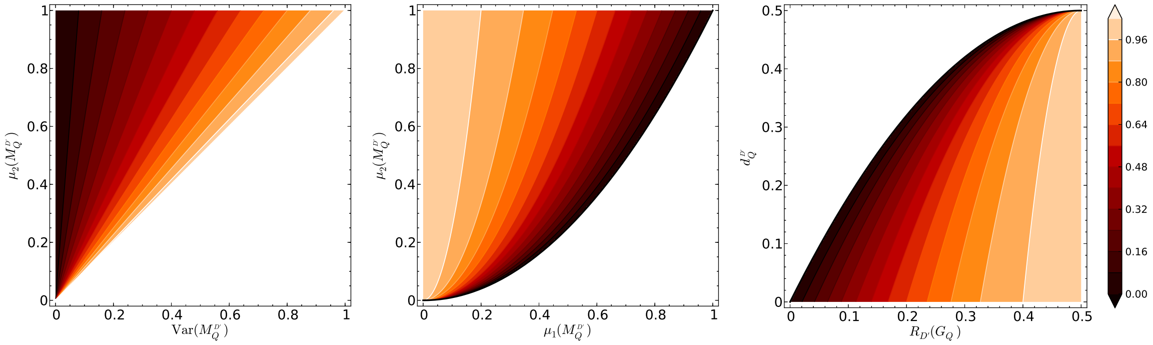
Proposition 10 *For any distribution Q on a set of voters and any distribution D' on $\mathcal{X}\times\{-1,1\}$, we have*

$$R_{D'}(B_Q) \;\leq\; 2\cdot R_{D'}(G_Q)\,.$$

Proof Starting from Equation (5) and using Markov's inequality (Lemma 46), we have

$$\begin{aligned}
R_{D'}(B_Q) &= \Pr_{(x,y)\sim D'}\big(M_Q(x,y)\leq 0\big)\\
&= \Pr_{(x,y)\sim D'}\big(1-M_Q(x,y)\geq 1\big)\\
&\leq \mathop{\mathbf{E}}_{(x,y)\sim D'}\big(1-M_Q(x,y)\big) && \text{(Markov's inequality)}\\
&= 1-\mathop{\mathbf{E}}_{(x,y)\sim D'}M_Q(x,y)\\
&= 1-\mu_1(M_Q^{D'})\\
&= 2\cdot R_{D'}(G_Q)\,.
\end{aligned}$$

The last equality is directly obtained from Equation (7). ■



(a) First form.

(b) Second form.

(c) Third form.

Figure 2: Contour plots of the $\mathcal{C}$-bound.

This proof highlights that we can upper-bound $R_{D'}(B_Q)$ by considering solely the first moment of the margin $\mu_1(M_Q^{D'})$. Once we realize this fact, it becomes natural to extend this result to higher moments. We do so in the following subsection where we make use of Chebyshev's inequality (instead of Markov's inequality), which uses not only the first, but also the second moment of the margin. This gives rise to the $\mathcal{C}$-bound of Theorem 11.

4.3 The $\mathcal{C}$-bound: a Bound on $R_{D'}(B_Q)$ That Can Be Much Smaller Than $R_{D'}(G_Q)$

Here is the bound on which most of the results of this paper are based. We refer to it as the $\mathcal{C}$-bound. It was first introduced (but in a different form) in Lacasse et al. (2006).[5] We give here three different (but equivalent) forms of the $\mathcal{C}$-bound. Each one highlights a different property or behavior of the bound. Figure 2 illustrates these behaviors.

It is interesting to note that the proof of Theorem 11 below has the same starting point as the proof of Proposition 10, but uses Chebyshev's inequality instead of Markov's inequality (respectively Lemmas 48 and 46, both provided in Appendix A). Therefore, Theorem 11 is based on the variance of the margin in addition of its mean.

Theorem 11 (The $\mathcal{C}$-bound) *For any distribution Q on a set of voters and any distribution D' on $\mathcal{X}\times\{-1,1\}$, if $\mu_1(M_Q^{D'})>0$ (i.e., $R_{D'}(G_Q)<1/2$), we have*

$$R_{D'}(B_Q) \;\le\; \mathcal{C}_Q^{D'}\,,$$

where

$$\mathcal{C}_Q^{D'} \;\overset{\text{def}}{=}\; \underbrace{\frac{\mathrm{Var}(M_Q^{D'})}{\mu_2(M_Q^{D'})}}_{\textit{First form}} \;=\; \underbrace{1-\frac{\Big(\mu_1(M_Q^{D'})\Big)^2}{\mu_2(M_Q^{D'})}}_{\textit{Second form}} \;=\; \underbrace{1-\frac{\Big(1-2\cdot R_{D'}(G_Q)\Big)^2}{1-2\cdot d_Q^{D'}}}_{\textit{Third form}}\,.$$

[5] We present the form used by Lacasse et al. (2006) in Remark 12 at the end of the present subsection.

Proof Starting from Equation (5) and using the one-sided Chebyshev inequality (Lemma 48), with $X = -M_Q(x,y)$, $\mu = \underset{(x,y)\sim D'}{\mathbf{E}}\big(-M_Q(x,y)\big)$ and $a = \underset{(x,y)\sim D'}{\mathbf{E}} M_Q(x,y)$, we obtain

$$\begin{aligned}
R_{D'}(B_Q) &= \underset{(x,y)\sim D'}{\Pr}\Big(M_Q(x,y) \leq 0\Big) \\
&= \underset{(x,y)\sim D'}{\Pr}\left(-M_Q(x,y) + \underset{(x,y)\sim D'}{\mathbf{E}} M_Q(x,y) \geq \underset{(x,y)\sim D'}{\mathbf{E}} M_Q(x,y)\right) \\
&\leq \frac{\underset{(x,y)\sim D'}{\mathbf{Var}}\,(M_Q(x,y))}{\underset{(x,y)\sim D'}{\mathbf{Var}}\,(M_Q(x,y)) + \left(\underset{(x,y)\sim D'}{\mathbf{E}}\, M_Q(x,y)\right)^2} \qquad \text{(Chebyshev's inequality)} \\
&= \frac{\mathrm{Var}(M_Q^{D'})}{\mu_2(M_Q^{D'}) \;\; - \Big(\mu_1(M_Q^{D'})\Big)^2 + \Big(\mu_1(M_Q^{D'})\Big)^2} \quad = \quad \frac{\mathrm{Var}(M_Q^{D'})}{\mu_2(M_Q^{D'})} \qquad (13) \\
&= \frac{\mu_2(M_Q^{D'}) \;\; - \Big(\mu_1(M_Q^{D'})\Big)^2}{\mu_2(M_Q^{D'})} \\
&= 1 - \frac{\Big(\mu_1(M_Q^{D'})\Big)^2}{\mu_2(M_Q^{D'})} \qquad (14) \\
&= 1 - \frac{\Big(1 - 2 \cdot R_{D'}(G_Q)\Big)^2}{1 - 2 \cdot d_Q^{D'}}\,. \qquad (15)
\end{aligned}$$

Lines (13) and (14) respectively present the first and the second forms of $\mathcal{C}_Q^{D'}$, and follow from the definitions of $\mu_1(M_Q^{D'})$, $\mu_2(M_Q^{D'})$, and $\mathrm{Var}(M_Q^{D'})$ (see Equations 6, 8 and 10). The third form of $\mathcal{C}_Q^{D'}$ is obtained at Line (15) using $\mu_1(M_Q^{D'}) = 1 - 2 \cdot R_{D'}(G_Q)$ and $\mu_2(M_Q^{D'}) = 1 - 2 \cdot d_Q^{D'}$, which can be derived directly from Equations (7) and (9). ■

The third form of the $\mathcal{C}$-bound shows that the bound decreases when the Gibbs risk $R_{D'}(G_Q)$ decreases or when the disagreement $d_Q^{D'}$ increases. This new bound therefore suggests that a majority vote should perform a trade-off between the Gibbs risk and the disagreement in order to achieve a low Bayes risk. This is more informative than the usual bound of Proposition 10, which focuses solely on the minimization of the Gibbs risk.

The first form of the $\mathcal{C}$-bound highlights that its value is always positive (since the variance and the second moment of the margin are positive), whereas the second form of the $\mathcal{C}$-bound highlights that it cannot exceed one. Finally, the fact that $d_Q^{D'} = \frac{1}{2} \Rightarrow R_{D'}(G_Q) = \frac{1}{2}$ (Proposition 9) implies that the bound is always defined, since $R_{D'}(G_Q)$ is here assumed to be strictly less than $\frac{1}{2}$.

Remark 12 As explained before, the $\mathcal{C}$-bound was originally stated in Lacasse et al. (2006), but in a different form. It was presented as a function of $W_Q(x,y)$, the Q-weight of voters making an error on example (x,y). More precisely, the $\mathcal{C}$-bound was presented as follows:

$$\mathcal{C}_Q^D \;=\; \frac{\underset{(x,y)\sim D'}{\mathbf{Var}}\,\big(W_Q(x,y)\big)}{\underset{(x,y)\sim D'}{\mathbf{Var}}\,\big(W_Q(x,y)\big) + (1/2 - R_{D'}(G_Q))^2}\,.$$

It is easy to show that this form is equivalent to the three forms stated in Theorem 11, and that $W_Q(x,y)$ and $M_Q(x,y)$ are related by

$$W_Q(x,y) \ \stackrel{\text{def}}{=} \ \mathop{\mathbf{E}}_{f\sim Q} \mathcal{L}_\ell\big(f(x),y\big) \ = \ \frac{1}{2}\left(1 - y\cdot \mathop{\mathbf{E}}_{f\sim Q} f(x)\right) \ = \ \frac{1}{2}\Big(1 - M_Q(x,y)\Big)\,.$$

However, we do not discuss further this form of the $\mathcal{C}$-bound here, since we now consider that the margin $M_Q(x,y)$ is a more natural notion than $W_Q(x,y)$.

4.4 Statistical Analysis of the $\mathcal{C}$-bound's Behavior

This section presents some properties of the $\mathcal{C}$-bound. In the first place, we discuss the conditions under which the $\mathcal{C}$-bound is optimal, in the sense that if the only information that one has about a majority vote is the first two moments of its margin distribution, it is possible that the value given by the $\mathcal{C}$-bound *is* the Bayes risk, *i.e.*, $\mathcal{C}_Q^{D'} = R_{D'}(B_Q)$.[6] In the second place, we show that the $\mathcal{C}$-bound can be arbitrarily small, especially in the presence of "non-correlated" voters, even if the Gibbs risk is large, *i.e.*, $\mathcal{C}_Q^{D'} \ll R_{D'}(G_Q)$.

4.4.1 Conditions of Optimality

For the sake of simplicity, let us focus on a random variable M that represents a margin distribution (here, we ignore underlying distributions Q on $\mathcal{H}$ and D' on $\mathcal{X}\times\{-1,1\}$) of first moment $\mu_1(M)$ and second moment $\mu_2(M)$. By Equation (5), we have

$$R(B_M) \ \stackrel{\text{def}}{=} \ \Pr\left(M \le 0\right). \tag{16}$$

Moreover, $R(B_M)$ is upper-bounded by $\mathcal{C}_M$, the $\mathcal{C}$-bound given by the second form of Theorem 11,

$$\mathcal{C}_M \ \stackrel{\text{def}}{=} \ 1 - \frac{\big(\mu_1(M)\big)^2}{\mu_2(M)}\,. \tag{17}$$

The next proposition shows when the $\mathcal{C}$-bound can be achieved.

Proposition 13 (Optimality of the $\mathcal{C}$-bound) *Let M be any random variable that represents the margin of a majority vote. Then there exists a random variable $\widetilde{M}$ such that*

$$\mu_1(\widetilde{M}) = \mu_1(M)\,, \quad \mu_2(\widetilde{M}) = \mu_2(M)\,, \quad \textit{and} \quad \mathcal{C}_{\widetilde{M}} = \mathcal{C}_M = R(B_{\widetilde{M}}) \tag{18}$$

if and only if

$$0 < \mu_2(M) \le \mu_1(M)\,. \tag{19}$$

Proof First, let us show that (19) implies (18). Given $0 < \mu_2(M) \le \mu_1(M)$, we consider a distribution $\widetilde{M}$ concentrated in two points defined as

$$\widetilde{M} \ = \ \begin{cases} 0 & \text{with probability } \mathcal{C}_M = 1 - \dfrac{\big(\mu_1(M)\big)^2}{\mu_2(M)}\,, \\[2ex] \dfrac{\mu_2(M)}{\mu_1(M)} & \text{with probability } 1 - \mathcal{C}_M = \dfrac{\big(\mu_1(M)\big)^2}{\mu_2(M)}\,. \end{cases}$$

[6] In other words, the *optimality of the $\mathcal{C}$-bound* means here that there exists a random variable with the same first moments as the margin distribution, such that Chebyshev's inequality of Lemma 48 is reached.

This distribution has the required moments, as

$$\mu_1(\widetilde{M}) = \frac{\big(\mu_1(M)\big)^2}{\mu_2(M)}\left[\frac{\mu_2(M)}{\mu_1(M)}\right] = \mu_1(M)\,, \text{ and } \mu_2(\widetilde{M}) = \frac{\big(\mu_1(M)\big)^2}{\mu_2(M)}\left[\frac{\mu_2(M)}{\mu_1(M)}\right]^2 = \mu_2(M)\,.$$

It follows directly from Equation (17) that $\mathcal{C}_{\widetilde{M}} = \mathcal{C}_M$. Moreover, by Equation (16) and because $\frac{\mu_2(M)}{\mu_1(M)} > 0$, we obtain as desired

$$R(B_{\widetilde{M}}) \;=\; \Pr\big(\widetilde{M} \le 0\big) \;=\; \mathcal{C}_M\,.$$

Now, let us show that (18) implies (19). Consider a distribution $\widetilde{M}$ such that the equalities of Line (18) are satisfied. By Proposition 10 and Equation (7), we obtain the inequality

$$\mathcal{C}_M \;=\; R(B_{\widetilde{M}}) \;\le\; 1-\mu_1(\widetilde{M}) \;=\; 1-\mu_1(M)\,.$$

Hence, by the definition of $\mathcal{C}_M$, we have

$$1-\frac{\big(\,\mu_1(M)\,\big)^2}{\mu_2(M)} \;\le\; 1-\mu_1(M)\,,$$

which, by straightforward calculations, implies $0 \;<\; \mu_2(M) \;\le\; \mu_1(M)$, and we are done. ■

We discussed in Section 4.1 the multiple connections between the moments of the margin, the Gibbs risk and the expected disagreement of a majority vote. In the next proposition, we exploit these connections to derive expressions equivalent to Line (19) of Proposition 13. Thus, this shows three (equivalent) necessary conditions under which the $\mathcal{C}$-bound is optimal.

Proposition 14 *For any distribution Q on a set of voters and any distribution D' on $\mathcal{X}\times\{-1,1\}$, if $\mu_1(M_Q^{D'}) > 0$ (i.e., $R_{D'}(G_Q) < 1/2$), then the three following statements are equivalent:*

(i) $\mu_2(M_Q^{D'}) \;\le\; \mu_1(M_Q^{D'})$ *;*

(ii) $R_{D'}(G_Q) \;\le\; d_Q^{D'}$ *;*

(iii) $\mathcal{C}_Q^{D'} \;\le\; 2\,R_{D'}(G_Q)$ *.*

Proof The truth of $(i) \Leftrightarrow (ii)$ is a direct consequence of Equations (7) and (9). To prove $(ii) \Leftrightarrow (iii)$, we express $\mathcal{C}_Q^{D'}$ in its third form. Straightforward calculations give

$$\mathcal{C}_Q^{D'} \;=\; 1-\frac{\big(1-2\,R_{D'}(G_Q)\big)^2}{1-2\,d_Q^{D'}} \;\le\; 2R_{D'}(G_Q) \quad\Longleftrightarrow\quad R_{D'}(G_Q) \;\le\; d_Q^{D'}\,.$$

■

Propositions 13 and 14 illustrate an interesting result: the $\mathcal{C}$-bound is optimal if and only if its value is lower than twice the Gibbs risk, the classical bound on the risk of the majority vote (see Proposition 10).

4.4.2 The $\mathcal{C}$-bound Can Be Arbitrarily Small, Even for Large Gibbs Risks

The next result shows that, when the number of voters tends to infinity (and the weight of each voter tends to zero), the variance of M_Q will tend to 0 provided that the average of the covariance of the outputs of all pairs of distinct voters is ≤ 0. In particular, the variance will always tend to 0 if the risk of the voters is pairwise independent. To quantify the independence between voters, we use the concept of covariance of a pair of voters (f_1, f_2):

$$\begin{aligned}\mathrm{Cov}_{D'}(f_1,f_2) &\overset{\text{def}}{=} \mathop{\mathbf{Cov}}_{(x,y)\sim D'}\Big(y\cdot f_1(x),\, y\cdot f_2(x)\Big)\\ &= \mathop{\mathbf{E}}_{(x,y)\sim D'} f_1(x)f_2(x) - \left(\mathop{\mathbf{E}}_{(x,y)\sim D'} f_1(x)\right)\left(\mathop{\mathbf{E}}_{(x,y)\sim D'} f_2(x)\right).\end{aligned}$$

Note that the covariance $\mathrm{Cov}_{D'}(f_1, f_2)$ is zero when f_1 and f_2 are independent (uncorrelated).

Proposition 15 *For any countable set of voters $\mathcal{H}$, any distribution Q on $\mathcal{H}$, and any distribution D' on $\mathcal{X}\times\{-1,1\}$, we have*

$$\mathrm{Var}(M_Q^{D'}) \;\leq\; \sum_{f\in\mathcal{H}} Q^2(f) + \sum_{f_1\in\mathcal{H}}\sum_{\substack{f_2\in\mathcal{H}:\\ f_2\neq f_1}} Q(f_1)Q(f_2)\cdot\mathrm{Cov}_{D'}(f_1,f_2)\,.$$

Proof By the definition of the margin (Definition 8), we rewrite $M_Q(x,y)$ as a sum of random variables:

$$\begin{aligned}&\mathop{\mathbf{Var}}_{(x,y)\sim D'}\Big(M_Q(x,y)\Big)\\ &\quad= \mathop{\mathbf{Var}}_{(x,y)\sim D'}\left(\sum_{f\in\mathcal{H}} Q(f)\cdot y\cdot f(x)\right)\\ &\quad= \sum_{f\in\mathcal{H}} Q^2(f)\mathop{\mathbf{Var}}_{(x,y)\sim D'}\Big(y\cdot f(x)\Big) + \sum_{f_1\in\mathcal{H}}\sum_{\substack{f_2\in\mathcal{H}:\\ f_2\neq f_1}} Q(f_1)Q(f_2)\mathop{\mathbf{Cov}}_{(x,y)\sim D'}\Big(y\cdot f_1(x), y\cdot f_2(x)\Big).\end{aligned}$$

The inequality is a consequence of the fact that $\forall f\in\mathcal{H}: \mathop{\mathbf{Var}}_{(x,y)\sim D'}\Big(y\cdot f(x)\Big)\leq 1$. ■

The key observation that comes out of this result is that $\sum_{f\in\mathcal{H}} Q^2(f)$ is usually much smaller than one. Consider, for example, the case where Q is uniform on $\mathcal{H}$ with $|\mathcal{H}| = n$. Then $\sum_{f\in\mathcal{H}} Q^2(f) = 1/n$. Moreover, if $\mathrm{Cov}_{D'}(f_1, f_2)\leq 0$ for each pair of distinct classifiers in $\mathcal{H}$, then $\mathrm{Var}(M_Q^{D'})\leq 1/n$. Hence, in these cases, we have that $\mathcal{C}_Q^{D'}\in\mathcal{O}(1/n)$ whenever $1-2\,R_{D'}(G_Q)$ and $1-2\,d_Q^{D'}$ are larger than some positive constants independent of n. Thus, even when $R_{D'}(G_Q)$ is large, we see that the $\mathcal{C}$-bound can be arbitrarily close to 0 as we increase the number of classifiers having non-positive pairwise covariance of their risk. More precisely, we have

Corollary 16 *Given n independent voters under a uniform distribution Q, we have*

$$R_{D'}(B_Q)\;\leq\;\mathcal{C}_Q^{D'}\;\leq\;\frac{1}{n\cdot\Big(1-2\,d_Q^{D'}\Big)}\;\leq\;\frac{1}{n\cdot\Big(1-2\,R_{D'}(G_Q)\Big)^2}\,.$$

Proof The first inequality directly comes from the $\mathcal{C}$-bound (Theorem 11). The second inequality is a consequence of Proposition 15, considering that in the case of a uniform distribution of independent voters, we have $\mathrm{Cov}_{D'}(f_1, f_2) = 0$, and then $\mathrm{Var}(M_Q^{D'}) \leq 1/n$. Applying this to the first form of the $\mathcal{C}$-bound, we obtain

$$\mathcal{C}_Q^{D'} \;=\; \frac{\mathrm{Var}(M_Q^{D'})}{\mu_2(M_Q^{D'})} \;=\; \frac{\mathrm{Var}(M_Q^{D'})}{1-2\,d_Q^{D'}} \;\leq\; \frac{\frac{1}{n}}{1-2\,d_Q^{D'}} \;=\; \frac{1}{n\cdot\left(1-2\,d_Q^{D'}\right)}\,.$$

To obtain the third inequality, we simply apply Equation (11), and we are done. ■

4.5 Empirical Study of The Predictive Power of the $\mathcal{C}$-bound

To further motivate the use of the $\mathcal{C}$-bound, we investigate how its empirical value relates to the risk of the majority vote by conducting two experiments. The first experiment shows that the $\mathcal{C}$-bound clearly outperforms the individual capacity of the other quantities of Theorem 11 in the task of predicting the risk of the majority vote. The second experiment shows that the $\mathcal{C}$-bound is a great stopping criterion for Boosting algorithms.

4.5.1 Comparison with Other Indicators

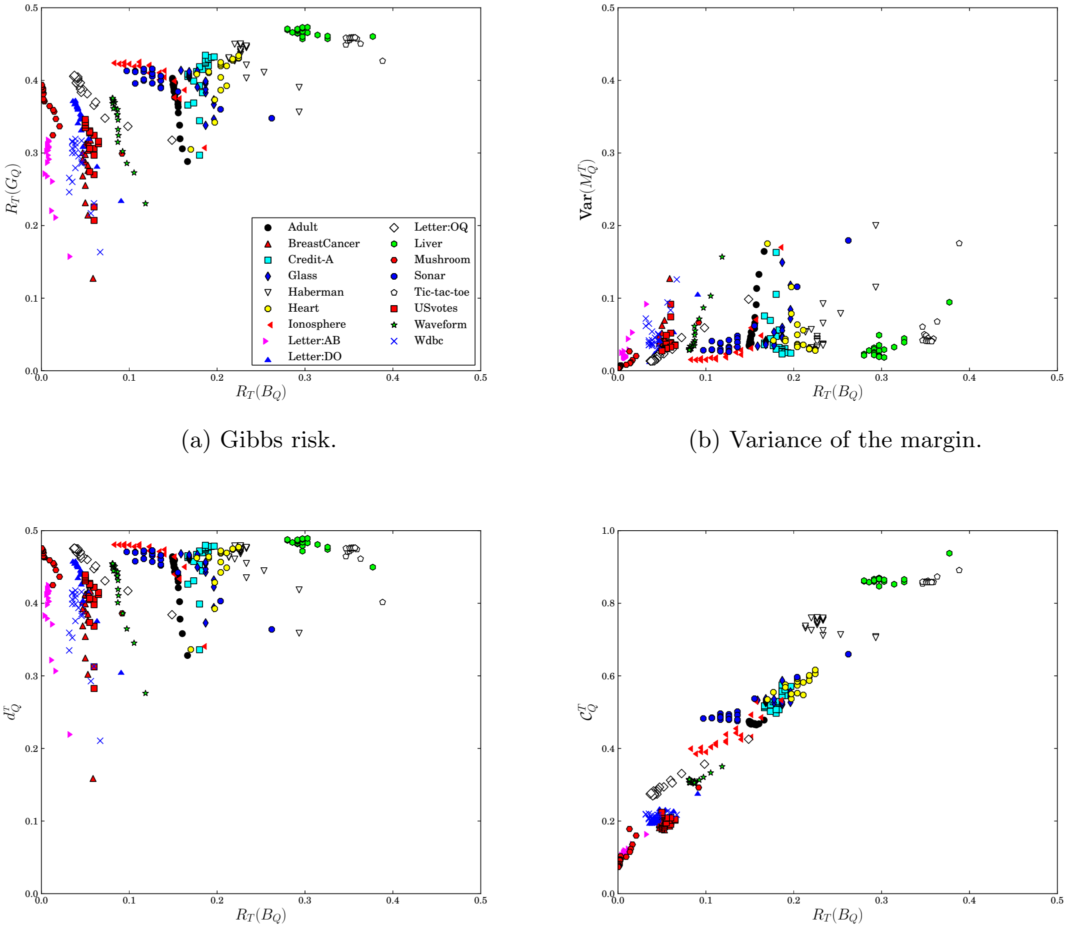
We study how $R_{D'}(G_Q)$, $\mathrm{Var}(M_Q^{D'})$, $d_Q^{D'}$ and $\mathcal{C}_Q^{D'}$ are respectively related to $R_{D'}(B_Q)$. Note that these four quantities appear in the first form or the third form of the $\mathcal{C}$-bound (Theorem 11). We omit here the moments $\mu_1(M_Q^{D'})$ and $\mu_2(M_Q^{D'})$ required by the second form of the $\mathcal{C}$-bound, as there is a linear relation between $\mu_1(M_Q^{D'})$ and $R_{D'}(G_Q)$, as well as between $\mu_2(M_Q^{D'})$ and $d_Q^{D'}$.

The results of Figure 3 are obtained with the AdaBoost algorithm of Schapire and Singer (1999), used with "decision stumps" as weak learners, on several UCI binary classification data sets (Blake and Merz, 1998). Each data set is split into two halves: a training set S and a testing set T. We run AdaBoost on set S for 100 rounds and compute the quantities $R_T(G_Q)$, $\mathrm{Var}(M_Q^T)$, d_Q^T and $\mathcal{C}_Q^T$ on set T at every 5 rounds of boosting. That is, we study 20 different majority vote classifiers per data set.

In Figure 3a, we see that we almost always have $R_T(B_Q) < R_T(G_Q)$. There is, however, no clear correlation between $R_T(B_Q)$ and $R_T(G_Q)$. We also see no clear correlation between $R_T(B_Q)$ and $\mathrm{Var}(M_Q^T)$ or between $R_T(B_Q)$ and d_Q^T in Figures 3b and 3c respectively, except that generally $R_T(B_Q) > \mathrm{Var}(M_Q^T)$ and $R_T(B_Q) < d_Q^T$. In contrast, Figure 3d shows a strong correlation between $\mathcal{C}_Q^T$ and $R_T(B_Q)$. Indeed, it is almost a linear relation! Therefore, the $\mathcal{C}$-bound seems well-suited to characterize the behavior of the Bayes risk, whereas each of the individual quantities contained in the $\mathcal{C}$-bound is insufficient to do so.

4.5.2 The $\mathcal{C}$-bound as a Stopping Criterion for Boosting

We now evaluate the accuracy of the empirical value of the $\mathcal{C}$-bound as a model selection tool. More specifically, we compare its ability to act as a stopping criterion for the AdaBoost algorithm.



(a) Gibbs risk.

(b) Variance of the margin.

(c) Expected disagreement.

(d) $\mathcal{C}$-bound.

Figure 3: $R_T(B_Q)$ versus $R_T(G_Q)$, $\mathrm{Var}(M_Q^T)$, d_Q^T and $\mathcal{C}_Q^T$ respectively.

We use the same version of the algorithm and the same data sets as in the previous experiment. However, for this experiment, each data set is split into a training set S of at most 400 examples and a testing set T containing the remaining examples. We run AdaBoost on set S for 1000 rounds. At each round, we compute the empirical $\mathcal{C}$-bound $\mathcal{C}_Q^S$ (on the training set). Afterwards, we select the majority vote classifier with the lowest value of $\mathcal{C}_Q^S$ and compute its Bayes risk $R_T(B_Q)$ (on the test set). We compare this stopping criterion with three other methods. For the first method, we compute the empirical Bayes risk $R_S(B_Q)$ at each round of boosting and, after that, we select the one having the lowest such risk.[7] The second method consists in performing 5-fold cross-validation and selecting the number of boosting rounds having the lowest cross-validation risk. Finally, the third method is to reserve 10% of S as a validation set, train AdaBoost on the remaining 90%,

[7] When several iterations have the same value of $R_S(B_Q)$, we select the earlier one.

Data Set Information			Risk $R_T(B_Q)$ by Stopping Criterion *(and number of rounds performed)*								
Name	$\|S\|$	$\|T\|$	$\mathcal{C}$-bound $\mathcal{C}_Q^S$		Risk $R_S(B_Q)$		Validation Set		Cross-Validation		1000 rounds
Adult	400	11409	0.166	*(149)*	0.169	*(314)*	**0.165**	*(13)*	0.166	*(97)*	0.172
BreastCancer	341	342	0.050	*(127)*	0.047	*(48)*	**0.041**	*(57)*	0.047	*(108)*	0.058
Credit-A	326	327	0.187	*(346)*	0.199	*(854)*	**0.156**	*(9)*	0.174	*(47)*	0.199
Glass	107	107	0.252	*(72)*	**0.196**	*(299)*	0.346	*(6)*	0.290	*(35)*	**0.196**
Haberman	147	147	0.320	*(27)*	0.320	*(45)*	**0.279**	*(1)*	0.320	*(38)*	0.340
Heart	148	149	0.215	*(124)*	0.289	*(950)*	**0.181**	*(31)*	0.195	*(14)*	0.289
Ionosphere	175	176	**0.085**	*(210)*	0.120	*(56)*	0.142	*(2)*	0.114	*(67)*	**0.085**
Letter:AB	400	1155	**0.005**	*(42)*	0.014	*(17)*	0.061	*(2)*	**0.005**	*(60)*	0.010
Letter:DO	400	1158	**0.041**	*(179)*	**0.041**	*(44)*	0.143	*(1)*	0.044	*(83)*	0.043
Letter:OQ	400	1136	0.050	*(65)*	0.050	*(138)*	0.063	*(26)*	**0.044**	*(118)*	0.049
Liver	172	173	**0.289**	*(541)*	**0.289**	*(743)*	0.335	*(5)*	**0.289**	*(603)*	0.295
Mushroom	400	7724	**0.010**	*(612)*	0.024	*(38)*	0.079	*(6)*	0.024	*(51)*	**0.010**
Sonar	104	104	0.192	*(688)*	0.250	*(20)*	0.317	*(2)*	**0.163**	*(34)*	0.202
Tic-tac-toe	400	558	0.389	*(59)*	0.364	*(2)*	**0.358**	*(5)*	0.403	*(9)*	0.389
USvotes	217	218	0.032	*(11)*	0.041	*(598)*	0.032	*(16)*	**0.028**	*(1)*	0.046
Waveform	400	7600	**0.101**	*(145)*	0.102	*(178)*	0.106	*(13)*	0.103	*(22)*	0.115
Wdbc	284	285	0.049	*(40)*	0.060	*(19)*	0.091	*(2)*	**0.046**	*(10)*	0.060

	Statistical Comparison Tests			
	$\mathcal{C}_Q^S$ vs $R_S(B_Q)$	$\mathcal{C}_Q^S$ vs Validation Set	$\mathcal{C}_Q^S$ vs Cross-Validation	$\mathcal{C}_Q^S$ vs 1000 rounds
Poisson binomial test	91%	86%	57%	90%
Sign test (p-value)	0.05	0.23	0.60	0.02

Table 1: Comparison of various stopping criteria over 1000 rounds of boosting. The Poisson binomial test gives the probability that $\mathcal{C}_Q^S$ is a better stopping criterion than every other approach. The sign test gives a p-value representing the probability that the null hypothesis is true (*i.e.*, the $\mathcal{C}_Q^S$ stopping criterion has the same performance as every other approach).

and keep the majority vote with the lowest Bayes risk on the validation set. Note that this last method differs from the others because AdaBoost sees 10% fewer examples during the learning process, but this is the price to pay for using a validation set.

Table 1 compares the Bayes risks on the test set $R_T(B_Q)$ of the majority vote classifiers selected by the different stopping criteria. We compute the probability of $\mathcal{C}$-bound being a better stopping criteria than every other methods with two statistical tests: the Poisson binomial test (Lacoste et al., 2012) and the sign test (Mendenhall, 1983). Both statistical tests suggest that the empirical $\mathcal{C}$-bound is a better model selection tool than the empirical Bayes risk (as usual in machine learning tasks, this method is prone to overfitting) and the validation set (although this method performs very well sometimes, it suffers from the small quantity of training examples on several tasks). The empirical $\mathcal{C}$-bound and the cross-validation methods obtain a similar accuracy. However, the cross-validation procedure needs more running time. We conclude that the empirical $\mathcal{C}$-bound is a surprisingly good stopping criterion for Boosting.

5. A PAC-Bayesian Story: From Zero to a PAC-Bayesian $\mathcal{C}$-bound

In this section, we present a PAC-Bayesian theory that allows one to estimate the $\mathcal{C}$-bound value $\mathcal{C}_Q^D$ from its empirical estimate $\mathcal{C}_Q^S$. From there, we derive bounds on the risk of the majority vote $R_D(B_Q)$ based on empirical observations. We first recall the classical PAC-Bayesian bound (here called the PAC-Bound 0) that bounds the true Gibbs risk by its empirical counterpart. We then present two different PAC-Bayesian bounds on the majority vote classifier (respectively called PAC-Bounds 1 and 2). A third bound, PAC-Bound 3, will be presented in Section 6. This analysis intends to be self-contained, and can act as an introduction to PAC-Bayesian theory.[8]

The first PAC-Bayesian theorem was proposed by McAllester (1999). Given a set of voters $\mathcal{H}$, a *prior* distribution P on $\mathcal{H}$ chosen before observing the data, and a *posterior* distribution Q on $\mathcal{H}$ chosen after observing a training set $S \sim D^m$ (Q is typically chosen by running a learning algorithm on S), PAC-Bayesian theorems give tight risk bounds for the Gibbs classifier G_Q. These bounds on $R_D(G_Q)$ usually rely on two quantities:

a) The empirical Gibbs risk $R_S(G_Q)$, that is computed on the m examples of S,

$$R_S(G_Q) \;=\; \frac{1}{m}\sum_{i=1}^{m} \mathop{\mathbf{E}}_{f\sim Q} \mathcal{L}_\ell(f(x_i), y_i)\,.$$

b) The Kullback-Leibler divergence between distributions Q and P, that measures "how far" the chosen posterior Q is from the prior P,

$$\mathrm{KL}(Q\|P) \;\overset{\text{def}}{=}\; \mathop{\mathbf{E}}_{f\sim Q} \ln\frac{Q(f)}{P(f)}\,. \tag{20}$$

Note that the obtained PAC-Bayesian bounds are uniformly valid for all possible posteriors Q.

In the following, we present a very general PAC-Bayesian theorem (Section 5.1), and we specialize it to obtain a bound on the Gibbs risk $R_D(G_Q)$ that is converted in a bound on the risk of the majority vote $R_D(B_Q)$ by the factor 2 of Proposition 10 (Section 5.2). Then, we define new losses that rely on a pair of voters (Section 5.3). These new losses allow us to extend the PAC-Bayesian theory to directly bound $R_D(B_Q)$ through the $\mathcal{C}$-bound (Sections 5.4 and 5.5). For each proposed bound, we explain the algorithmic procedure required to compute its value.

5.1 General PAC-Bayesian Theory for Real-Valued Losses

A key step of most PAC-Bayesian proofs is summarized by the following *Change of measure inequality* (Lemma 17).

We present here the same proof as in Seldin and Tishby (2010) and McAllester (2013). Note that the same result is derived from Fenchel's inequality in Banerjee (2006) and Donsker-Varadhan's variational formula for relative entropy in Seldin et al. (2012); Tolstikhin and Seldin (2013).

[8] We also recommend the "practical prediction tutorial" of Langford (2005), that contains an insightful PAC-Bayesian introduction.

Lemma 17 (Change of measure inequality) *For any set $\mathcal{H}$, for any distributions P and Q on $\mathcal{H}$, and for any measurable function $\phi : \mathcal{H} \to \mathbb{R}$, we have*

$$\mathop{\mathbf{E}}_{f\sim Q} \phi(f) \;\le\; \mathrm{KL}(Q\|P) + \ln\left(\mathop{\mathbf{E}}_{f\sim P} e^{\phi(f)}\right) .$$

Proof The result is obtained by simple calculations, exploiting the definition of the KL-divergence given by Equation (20), and then Jensen's inequality (Lemma 47, in Appendix A) on concave function $\ln(\cdot)$:

$$\begin{aligned}
\mathop{\mathbf{E}}_{f\sim Q} \phi(f) &= \mathop{\mathbf{E}}_{f\sim Q} \ln e^{\phi(f)} = \mathop{\mathbf{E}}_{f\sim Q} \ln\left(\frac{Q(f)}{P(f)}\cdot\frac{P(f)}{Q(f)}\cdot e^{\phi(f)}\right)\\
&= \mathrm{KL}(Q\|P) + \mathop{\mathbf{E}}_{f\sim Q} \ln\left(\frac{P(f)}{Q(f)}\cdot e^{\phi(f)}\right)\\
&\le \mathrm{KL}(Q\|P) + \ln\left(\mathop{\mathbf{E}}_{f\sim Q} \frac{P(f)}{Q(f)}\cdot e^{\phi(f)}\right) \qquad \text{(Jensen's inequality)}\\
&\le \mathrm{KL}(Q\|P) + \ln\left(\mathop{\mathbf{E}}_{f\sim P} e^{\phi(f)}\right) .
\end{aligned}$$

Note that the last inequality becomes an equality if Q and P share the same support. ■

Let us now present a general PAC-Bayesian theorem which bounds the expectation of any real-valued loss function $\mathcal{L} : \overline{\mathcal{Y}} \times \mathcal{Y} \to [0,1]$. This theorem is slightly more general than the PAC-Bayesian theorem of Germain et al. (2009, Theorem 2.1), that is specialized to the expected linear loss, and therefore gives rise to a bound of the "generalized" Gibbs risk of Definition 5. A similar result is presented in Tolstikhin and Seldin (2013, Lemma 1).

Theorem 18 (General PAC-Bayesian theorem for real-valued losses) *For any distribution D on $\mathcal{X} \times \mathcal{Y}$, for any set $\mathcal{H}$ of voters $\mathcal{X} \to \overline{\mathcal{Y}}$, for any loss $\mathcal{L} : \overline{\mathcal{Y}} \times \mathcal{Y} \to [0,1]$, for any prior distribution P on $\mathcal{H}$, for any $\delta \in (0,1]$, for any $m' > 0$, and for any convex function $\mathcal{D} : [0,1]\times[0,1] \to \mathbb{R}$, we have*

$$\Pr_{S\sim D^m}\left(\begin{array}{l}\textit{For all posteriors } Q \textit{ on } \mathcal{H} :\\ \mathcal{D}(\mathop{\mathbf{E}}_{f\sim Q} \mathbb{E}_S^{\mathcal{L}}(f), \mathop{\mathbf{E}}_{f\sim Q} \mathbb{E}_D^{\mathcal{L}}(f)) \le \dfrac{1}{m'}\left[\mathrm{KL}(Q\|P) + \ln\left(\dfrac{1}{\delta}\mathop{\mathbf{E}}_{S\sim D^m}\mathop{\mathbf{E}}_{f\sim P} e^{m'\cdot\mathcal{D}(\mathbb{E}_S^{\mathcal{L}}(f),\mathbb{E}_D^{\mathcal{L}}(f))}\right)\right]\end{array}\right) \ge 1-\delta\,,$$

where $\mathrm{KL}(Q\|P)$ *is the Kullback-Leibler divergence between Q and P of Equation* (20).

Most of the time, this theorem is used with $m' = m$, the size of the training set. However, as pointed out by Lever et al. (2010), m' does not have to be so. One can easily show that different values of m' affect the relative weighting between the terms $\mathrm{KL}(Q\|P)$ and $\ln\left(\frac{1}{\delta}\mathbf{E}_{S\sim D^m}\mathbf{E}_{f\sim P} e^{m'\cdot\mathcal{D}(\mathbb{E}_S^{\mathcal{L}}(f),\mathbb{E}_D^{\mathcal{L}}(f))}\right)$ in the bound. Hence, especially in situations where these two terms have very different values, a "good" choice for the value of m' can tighten the bound.

Proof Note that $\mathop{\mathbf{E}}_{f\sim P} e^{m'\cdot\mathcal{D}(\mathbb{E}_S^{\mathcal{L}}(f),\mathbb{E}_D^{\mathcal{L}}(f))}$ is a non-negative random variable. By Markov's inequality (Lemma 46, in Appendix A), we have

$$\Pr_{S\sim D^m}\left(\mathop{\mathbf{E}}_{f\sim P} e^{m'\cdot\mathcal{D}(\mathbb{E}_S^{\mathcal{L}}(f),\mathbb{E}_D^{\mathcal{L}}(f))} \le \frac{1}{\delta}\mathop{\mathbf{E}}_{S\sim D^m}\mathop{\mathbf{E}}_{f\sim P} e^{m'\cdot\mathcal{D}(\mathbb{E}_S^{\mathcal{L}}(f),\mathbb{E}_D^{\mathcal{L}}(f))}\right) \ge 1-\delta\,.$$

Hence, by taking the logarithm on each side of the innermost inequality, we obtain

$$\Pr_{S\sim D^m}\left(\ln\left[\mathop{\mathbf{E}}_{f\sim P} e^{m'\cdot\mathcal{D}(\mathbb{E}_S^{\mathcal{L}}(f),\mathbb{E}_D^{\mathcal{L}}(f))}\right] \le \ln\left[\frac{1}{\delta}\mathop{\mathbf{E}}_{S\sim D^m}\mathop{\mathbf{E}}_{f\sim P} e^{m'\cdot\mathcal{D}(\mathbb{E}_S^{\mathcal{L}}(f),\mathbb{E}_D^{\mathcal{L}}(f))}\right]\right) \ge 1-\delta\,.$$

We apply the change of measure inequality (Lemma 17) on the left side of innermost inequality, with $\phi(f) = m'\cdot\mathcal{D}(\mathbb{E}_S^{\mathcal{L}}(f),\mathbb{E}_D^{\mathcal{L}}(f))$. We then use Jensen's inequality (Lemma 47, in Appendix A), exploiting the convexity of $\mathcal{D}$:

$$\begin{aligned}\forall\, Q \text{ on } \mathcal{H}: \quad \ln\left[\mathop{\mathbf{E}}_{f\sim P} e^{m'\cdot\mathcal{D}(\mathbb{E}_S^{\mathcal{L}}(f),\mathbb{E}_D^{\mathcal{L}}(f))}\right] \;&\ge\; m'\cdot\mathop{\mathbf{E}}_{f\sim Q}\mathcal{D}(\mathbb{E}_S^{\mathcal{L}}(f),\mathbb{E}_D^{\mathcal{L}}(f)) - \mathrm{KL}(Q\|P)\\ &\ge\; m'\cdot\mathcal{D}(\mathop{\mathbf{E}}_{f\sim Q}\mathbb{E}_S^{\mathcal{L}}(f),\mathop{\mathbf{E}}_{f\sim Q}\mathbb{E}_D^{\mathcal{L}}(f)) - \mathrm{KL}(Q\|P)\,.\end{aligned}$$

We therefore have

$$\Pr_{S\sim D^m}\left(\begin{array}{l}\text{For all posteriors } Q:\\ m'\cdot\mathcal{D}(\mathop{\mathbf{E}}_{f\sim Q}\mathbb{E}_S^{\mathcal{L}}(f),\mathop{\mathbf{E}}_{f\sim Q}\mathbb{E}_D^{\mathcal{L}}(f)) - \mathrm{KL}(Q\|P) \le \ln\left[\frac{1}{\delta}\mathop{\mathbf{E}}_{S\sim D^m}\mathop{\mathbf{E}}_{f\sim P} e^{m'\cdot\mathcal{D}(\mathbb{E}_S^{\mathcal{L}}(f),\mathbb{E}_D^{\mathcal{L}}(f))}\right]\end{array}\right)\ge 1-\delta\,.$$

The result then follows from easy calculations. ■

As shown in Germain et al. (2009), the general PAC-Bayesian theorem can be used to recover many common variants of the PAC-Bayesian theorem, simply by selecting a well-suited function $\mathcal{D}$. Among these, we obtain a similar bound as the one proposed by Langford and Seeger (2001); Seeger (2002); Langford (2005) by using the Kullback-Leibler divergence between the Bernoulli distributions with probability of success q and probability of success p:

$$\mathrm{kl}(q\,\|\,p) \;\overset{\text{def}}{=}\; q\ln\frac{q}{p} + (1-q)\ln\frac{1-q}{1-p}\,. \tag{21}$$

Note that $\mathrm{kl}(q\,\|\,p)$ is a shorthand notation for $\mathrm{KL}(Q\|P)$ of Equation (20), with $Q = (q, 1{-}q)$ and $P = (p, 1{-}p)$. Corollary 50 (in Appendix A) shows that $\mathrm{kl}(q\,\|\,p)$ is a convex function.

In order to apply Theorem 18 with $\mathcal{D}(q,p) = \mathrm{kl}(q\|p)$ and $m' = m$, we need the next lemma.

Lemma 19 *For any distribution D on $\mathcal{X}\times\mathcal{Y}$, for any voter $f:\mathcal{X}\to\overline{\mathcal{Y}}$, for any loss $\mathcal{L}:\overline{\mathcal{Y}}\times\mathcal{Y}\to[0,1]$, and any positive integer m, we have*

$$\mathop{\mathbf{E}}_{S\sim D^m}\exp\left[m\cdot\mathrm{kl}\Big(\mathbb{E}_S^{\mathcal{L}}(f)\,\|\,\mathbb{E}_D^{\mathcal{L}}(f)\Big)\right] \;\le\; \xi(m)\,,$$

where

$$\xi(m) \;\overset{\text{def}}{=}\; \sum_{k=0}^{m}\binom{m}{k}\left(\frac{k}{m}\right)^k\left(1-\frac{k}{m}\right)^{m-k}. \tag{22}$$

Moreover, $\sqrt{m} \,\le\, \xi(m) \,\le\, 2\sqrt{m}\,.$

Proof Let us introduce a random variable X_f that follows a binomial distribution of m trials with a probability of success $\mathbb{E}_D^{\mathcal{L}}(f)$. Hence, $X_f \sim B(m, \mathbb{E}_D^{\mathcal{L}}(f))\,.$

As $e^{m\cdot \mathrm{kl}\left(\,\cdot\,\|\,\mathbb{E}_D^{\mathcal{L}}(f)\right)}$ is a convex function, Lemma 51 (due to Maurer, 2004, and provided in Appendix A), shows that

$$\mathop{\mathbf{E}}_{S\sim D^m}\ \exp\left[m\cdot \mathrm{kl}\Big(\mathbb{E}_S^{\mathcal{L}}(f)\,\|\,\mathbb{E}_D^{\mathcal{L}}(f)\Big)\right]\ \le \mathop{\mathbf{E}}_{X_f\sim B(m,\mathbb{E}_D^{\mathcal{L}}(f))}\ \exp\left[m\cdot \mathrm{kl}\Big(\tfrac{1}{m}X_f\,\|\,\mathbb{E}_D^{\mathcal{L}}(f)\Big)\right].$$

We then have

$$\begin{aligned}
&\mathop{\mathbf{E}}_{X_f\sim B(m,\mathbb{E}_D^{\mathcal{L}}(f))}\ e^{m\mathrm{kl}(\frac{1}{m}X_f\|\mathbb{E}_D^{\mathcal{L}}(f))}\\
&\quad= \mathop{\mathbf{E}}_{X_f\sim B(m,\mathbb{E}_D^{\mathcal{L}}(f))}\left(\frac{\frac{1}{m}X_f}{\mathbb{E}_D^{\mathcal{L}}(f)}\right)^{X_f}\left(\frac{1-\frac{1}{m}X_f}{1-\mathbb{E}_D^{\mathcal{L}}(f)}\right)^{m-X_f}\\
&\quad= \sum_{k=0}^{m}\ \mathop{\Pr}_{X_f\sim B(m,\mathbb{E}_D^{\mathcal{L}}(f))}\Big(X_f=k\Big)\cdot\left(\frac{\frac{k}{m}}{\mathbb{E}_D^{\mathcal{L}}(f)}\right)^{k}\left(\frac{1-\frac{k}{m}}{1-\mathbb{E}_D^{\mathcal{L}}(f)}\right)^{m-k}\\
&\quad= \sum_{k=0}^{m}\binom{m}{k}\Big(\mathbb{E}_D^{\mathcal{L}}(f)\Big)^{k}\Big(1-\mathbb{E}_D^{\mathcal{L}}(f)\Big)^{m-k}\cdot\left(\frac{\frac{k}{m}}{\mathbb{E}_D^{\mathcal{L}}(f)}\right)^{k}\left(\frac{1-\frac{k}{m}}{1-\mathbb{E}_D^{\mathcal{L}}(f)}\right)^{m-k}\\
&\quad= \sum_{k=0}^{m}\binom{m}{k}\left(\frac{k}{m}\right)^{k}\left(1-\frac{k}{m}\right)^{m-k}\ =\ \xi(m)\,.
\end{aligned}$$

Maurer (2004) shows that $\xi(m)\le 2\sqrt{m}$ for $m\ge 8$, and $\xi(m)\ge\sqrt{m}$ for $m\ge 2$. However, the cases for $m\in\{1,2,3,4,5,6,7\}$ are easy to verify computationally. ■

Theorem 20 below specializes the general PAC-Bayesian theorem to $\mathcal{D}(q,p)=\mathrm{kl}(q\|p)$, but still applies to any real-valued loss functions. This theorem can be seen as an intermediate step to obtain Corollary 21 of the next section, which uses the linear loss to bound the Gibbs risk. However, Theorem 20 below is reused afterwards in Section 5.3 to derive PAC-Bayesian theorems for other loss functions.

Theorem 20 *For any distribution D on $\mathcal{X}\times\mathcal{Y}$, for any set $\mathcal{H}$ of voters $\mathcal{X}\to\overline{\mathcal{Y}}$, for any loss $\mathcal{L}:\overline{\mathcal{Y}}\times\mathcal{Y}\to[0,1]$, for any prior distribution P on $\mathcal{H}$, for any $\delta\in(0,1]$, we have*

$$\mathop{\Pr}_{S\sim D^m}\left(\begin{array}{l}\textit{For all posteriors } Q \textit{ on } \mathcal{H}:\\ \mathrm{kl}\Big(\mathop{\mathbf{E}}_{f\sim Q}\mathbb{E}_S^{\mathcal{L}}(f)\,\Big\|\,\mathop{\mathbf{E}}_{f\sim Q}\mathbb{E}_D^{\mathcal{L}}(f)\Big)\ \le\ \dfrac{1}{m}\left[\mathrm{KL}(Q\|P)+\ln\dfrac{\xi(m)}{\delta}\right]\end{array}\right)\ \ge\ 1-\delta\,.$$

Proof By Theorem 18, with $\mathcal{D}(q,p)=\mathrm{kl}(q\|p)$ and $m'=m$, we have

$$\mathop{\Pr}_{S\sim D^m}\left(\begin{array}{l}\forall Q \text{ on } \mathcal{H}:\\ \mathrm{kl}(\mathop{\mathbf{E}}_{f\sim Q}\mathbb{E}_S^{\mathcal{L}}(f)\,\|\,\mathop{\mathbf{E}}_{f\sim Q}\mathbb{E}_D^{\mathcal{L}}(f))\le\dfrac{1}{m}\left[\mathrm{KL}(Q\|P)+\ln\left(\dfrac{1}{\delta}\mathop{\mathbf{E}}_{S\sim D^m}\mathop{\mathbf{E}}_{f\sim P}e^{m\cdot\mathrm{kl}(\mathbb{E}_S^{\mathcal{L}}(f)\,\|\,\mathbb{E}_D^{\mathcal{L}}(f))}\right)\right]\end{array}\right)\ge 1-\delta\,.$$

As the prior P is independent of S, we can swap the two expectations in $\mathop{\mathbf{E}}_{S\sim D^m}\mathop{\mathbf{E}}_{f\sim P}e^{m\cdot\mathrm{kl}(\cdot\|\cdot)}$. This observation, together with Lemma 19, gives

$$\mathop{\mathbf{E}}_{S\sim D^m}\ \mathop{\mathbf{E}}_{f\sim P}e^{m\cdot\mathrm{kl}(\mathbb{E}_S^{\mathcal{L}}(f)\,\|\,\mathbb{E}_D^{\mathcal{L}}(f))}\ =\ \mathop{\mathbf{E}}_{f\sim P}\ \mathop{\mathbf{E}}_{S\sim D^m}e^{m\cdot\mathrm{kl}(\mathbb{E}_S^{\mathcal{L}}(f)\,\|\,\mathbb{E}_D^{\mathcal{L}}(f))}\ \le\ \mathop{\mathbf{E}}_{f\sim P}\xi(m)\ =\ \xi(m)\,.$$

■

5.2 PAC-Bayesian Theory for the Gibbs Classifier

This section presents two classical PAC-Bayesian results that bound the risk of the Gibbs classifier. One of these bounds is used to express a first PAC-Bayesian bound on the risk of the majority vote classifier. Then, we explain how to compute the empirical value of this bound by a root-finding method.

5.2.1 PAC-Bayesian Theorems for the Gibbs Risk

We interpret the two following results as straightforward corollaries of Theorem 20. Indeed, from Definition 5, the expected linear loss of a Gibbs classifier G_Q on a distribution D' *is* $R_{D'}(G_Q)$. These two Corollaries are very similar to well-known PAC-Bayesian theorems. At first, Corollary 21 is similar to the PAC-Bayesian theorem of Langford and Seeger (2001); Seeger (2002); Langford (2005), with the exception that $\ln \frac{m+1}{\delta}$ is replaced by $\ln \frac{\xi(m)}{\delta}$. Since $\xi(m) \leq 2\sqrt{m} \leq m+1$, this result gives slightly better bounds. Similarly, Corollary 22 provides a slight improvement of the PAC-Bayesian bound of McAllester (1999, 2003a).

Corollary 21 (Langford and Seeger, 2001; Seeger, 2002; Langford, 2005) *For any distribution D on $\mathcal{X}\times\{-1,1\}$, for any set $\mathcal{H}$ of voters $\mathcal{X} \to [-1,1]$, for any prior distribution P on $\mathcal{H}$, and any $\delta \in (0,1]$, we have*

$$\Pr_{S\sim D^m}\left(\begin{array}{l}\textit{For all posteriors } Q \textit{ on } \mathcal{H}: \\ \mathrm{kl}\big(R_S(G_Q)\big\|R_D(G_Q)\big) \;\leq\; \dfrac{1}{m}\left[\mathrm{KL}(Q\|P) + \ln\dfrac{\xi(m)}{\delta}\right]\end{array}\right) \geq 1-\delta\,.$$

Proof The result is directly obtained from Theorem 20 using the linear loss $\mathcal{L} = \mathcal{L}_\ell$ to recover the Gibbs risk of Definition 5. ■

Corollary 22 (McAllester, 1999, 2003a) *For any distribution D on $\mathcal{X}\times\{-1,1\}$, for any set $\mathcal{H}$ of voters $\mathcal{X} \to [-1,1]$, for any prior distribution P on $\mathcal{H}$, and any $\delta \in (0,1]$, we have*

$$\Pr_{S\sim D^m}\left(\begin{array}{l}\textit{For all posteriors } Q \textit{ on } \mathcal{H}: \\ R_D(G_Q) \;\leq\; R_S(G_Q) + \sqrt{\dfrac{1}{2m}\left[\mathrm{KL}(Q\|P) + \ln\dfrac{\xi(m)}{\delta}\right]}\end{array}\right) \geq 1-\delta\,.$$

Proof The result is obtained from Corollary 21 together with Pinsker's inequality

$$2(q-p)^2 \;\leq\; \mathrm{kl}(q\|p)\,.$$

We then have

$$\Pr_{S\sim D^m}\left(\begin{array}{l}\text{For all posteriors } Q \text{ on } \mathcal{H}: \\ 2\cdot\Big(R_S(G_Q) - R_D(G_Q)\Big)^2 \;\leq\; \dfrac{1}{m}\left[\mathrm{KL}(Q\|P) + \ln\dfrac{\xi(m)}{\delta}\right]\end{array}\right) \geq 1-\delta\,.$$

The result is obtained by isolating $R_D(G_Q)$ in the inequality, omitting the lower bound of $R_D(G_Q)$. Recall that the probability is "$\geq 1{-}\delta$", hence if we omit an event, the probability may just increase, continuing to be greater than $1{-}\delta$. ■

5.2.2 A First Bound for the Risk of the Majority Vote

Let assume that the Gibbs risk $R_D(G_Q)$ of a classifier is lower than or equal to $\frac{1}{2}$. Given an empirical Gibbs risk $R_S(G_Q)$ computed on a training set of m examples, the Kullback-Leibler divergence $\mathrm{KL}(Q\|P)$, and a confidence parameter δ, Corollary 21 says that the Gibbs risk $R_D(G_Q)$ is included (with confidence $1-\delta$) in the continuous set $\mathcal{R}^{\delta}_{Q,S}$ defined as

$$\mathcal{R}^{\delta}_{Q,S} \overset{\text{def}}{=} \left\{ r \,:\, \mathrm{kl}\big(R_S(G_Q)\,\|\,r\big) \,\le\, \frac{1}{m}\left[\mathrm{KL}(Q\|P)+\ln\frac{\xi(m)}{\delta}\right] \quad \text{and} \quad r\le\tfrac{1}{2}\right\}. \tag{23}$$

Thus, an upper bound on $R_D(G_Q)$ is obtained by seeking the maximum value of $\mathcal{R}^{\delta}_{Q,S}$. As explained by Proposition 10, we need to multiply the obtained value by a factor 2 to have an upper bound on $R_D(B_Q)$. This methodology is summarized by PAC-Bound 0.

Note that PAC-Bound 0 is also valid when $R_D(G_Q)$ is greater than $\frac{1}{2}$, because in this case, $2\cdot\sup\mathcal{R}^{\delta}_{Q,S} = 1$ (with confidence at least $1-\delta$), which is a trivial upper bound of $R_D(B_Q)$.

PAC-Bound 0 *For any distribution D on $\mathcal{X}\times\{-1,1\}$, for any set $\mathcal{H}$ of voters $\mathcal{X}\to[-1,1]$, for any prior distribution P on $\mathcal{H}$, and any $\delta\in(0,1]$, we have*

$$\Pr_{S\sim D^m}\left(\forall\, Q \textit{ on } \mathcal{H} \,:\, R_D(B_Q) \,\le\, 2\cdot\sup\mathcal{R}^{\delta}_{Q,S}\right) \,\ge\, 1-\delta\,.$$

Proof If $\sup\mathcal{R}^{\delta}_{Q,S}=\frac{1}{2}$, the bound is trivially valid because $R_D(B_Q)\le 1$. Otherwise, the bound is a direct consequence of Proposition 10 and Corollary 21. ■

As we see, the proposed bound cannot be obtained by a closed-form expression. Thus, we need to use a strategy as the one suggested in the following.

5.2.3 Computation of PAC-Bound 0

One can compute the value $r=\sup\mathcal{R}^{\delta}_{Q,S}$ of PAC-Bound 0 by solving

$$\mathrm{kl}\big(R_S(G_Q)\,\|\,r\big) \;=\; \tfrac{1}{m}\big[\mathrm{KL}(Q\|P)+\ln\tfrac{\xi(m)}{\delta}\big]\,, \qquad \text{with } R_S(G_Q)\le r\le\tfrac{1}{2}\,,$$

by a root-finding method. This turns out to be an easy task since the left-hand side of the equality is a convex function of r and the right-hand side is a constant value. Note that solving the same equation with the constraint $r\le R_S(G_Q)$ gives a lower bound of $R_D(G_Q)$, but not a lower bound on $R_D(B_Q)$. Figure 4 shows an application example of PAC-Bound 0.

5.3 Joint Error, Joint Success, and Paired-voters

We now introduce a few notions that are necessary to obtain new PAC-Bayesian theorems for the $\mathcal{C}$-bound in Sections 5.4 and 5.5.

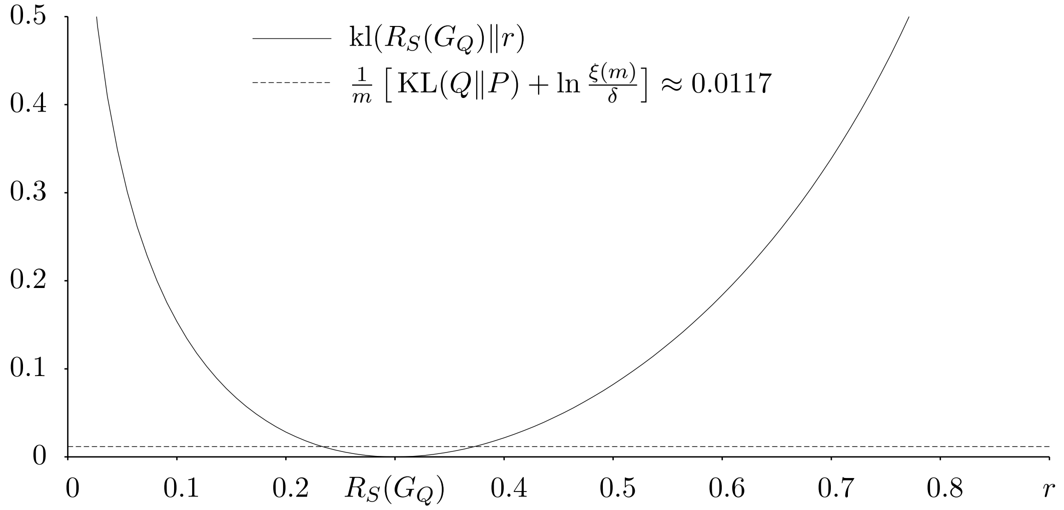


Figure 4: Example of application of PAC-Bound 0. We suppose that $\mathrm{KL}(Q\|P) = 5$, $m = 1000$ and $\delta = 0.05$. If we observe an empirical Gibbs risk $R_S(G_Q) = 0.30$, then $R_D(G_Q) \in \mathcal{R}^{\delta}_{Q,S} \approx [0.233, 0.373]$ with a confidence of 95%. On the figure, the intersections between the two curves correspond to the limits of the interval $\mathcal{R}^{\delta}_{Q,S}$. Then, with these values, PAC-bound 0 gives $R_D(B_Q) \lesssim 2 \cdot 0.373 = 0.746$.

5.3.1 The Joint Error and the Joint Success

We have already defined the expected disagreement $d_Q^{D'}$ of a distribution Q of voters (Definition 7). In the case of binary voters, the expected disagreement corresponds to

$$d_Q^{D'} \;=\; \mathop{\mathbf{E}}_{h_1\sim Q}\mathop{\mathbf{E}}_{h_2\sim Q}\Big(\mathop{\mathbf{E}}_{(x,y)\sim D'} I(h_1(x)\neq h_2(x))\Big)\,.$$

Let us now define two closely related notions, the expected joint success $s_Q^{D'}$ and the expected joint error $e_Q^{D'}$. In the case of binary voters, these two concepts are expressed naturally by

$$\begin{aligned} e_Q^{D'} &= \mathop{\mathbf{E}}_{h_1\sim Q}\mathop{\mathbf{E}}_{h_2\sim Q}\Big(\mathop{\mathbf{E}}_{(x,y)\sim D'} I(h_1(x)\neq y)I(h_2(x)\neq y)\Big)\,,\\ s_Q^{D'} &= \mathop{\mathbf{E}}_{h_1\sim Q}\mathop{\mathbf{E}}_{h_2\sim Q}\Big(\mathop{\mathbf{E}}_{(x,y)\sim D'} I(h_1(x)= y)I(h_2(x)= y)\Big)\,. \end{aligned}$$

Let us now extend in the usual way these equations to the case of real-valued voters.

Definition 23 For any probability distribution Q on a set of voters, we define the *expected joint error* $e_Q^{D'}$ relative to D' and the *expected joint success* $s_Q^{D'}$ relative to D' as

$$\begin{aligned} e_Q^{D'} &\overset{\mathrm{def}}{=} \mathop{\mathbf{E}}_{f_1\sim Q}\mathop{\mathbf{E}}_{f_2\sim Q}\Big(\mathop{\mathbf{E}}_{(x,y)\sim D'} \mathcal{L}_\ell(f_1(x),y)\cdot\mathcal{L}_\ell(f_2(x),y)\Big)\,,\\ s_Q^{D'} &\overset{\mathrm{def}}{=} \mathop{\mathbf{E}}_{f_1\sim Q}\mathop{\mathbf{E}}_{f_2\sim Q}\Big(\mathop{\mathbf{E}}_{(x,y)\sim D'} \Big[1-\mathcal{L}_\ell(f_1(x),y)\Big]\cdot\Big[1-\mathcal{L}_\ell(f_2(x),y)\Big]\Big)\,. \end{aligned}$$

From the definitions of the linear loss (Definition 2) and the margin (Definition 8), we can easily see that

$$\begin{aligned}
e_Q^{D'} &= \mathop{\mathbf{E}}_{(x,y)\sim D'} \left(\frac{1-M_Q(x,y)}{2}\right)^2 &=& \frac{1}{4}\Big(1-2\cdot\mu_1(M_Q^{D'})+\mu_2(M_Q^{D'})\Big)\,,\\
s_Q^{D'} &= \mathop{\mathbf{E}}_{(x,y)\sim D'} \left(\frac{1+M_Q(x,y)}{2}\right)^2 &=& \frac{1}{4}\Big(1+2\cdot\mu_1(M_Q^{D'})+\mu_2(M_Q^{D'})\Big)\,.
\end{aligned}$$

Remembering from Equation (9) that $d_Q^{D'} = \frac{1}{2}\Big(1-\mu_2(M_Q^{D'})\Big)$, we can conclude that $e_Q^{D'}$, $s_Q^{D'}$ and $d_Q^{D'}$ always sum to one:[9]

$$e_Q^{D'} + s_Q^{D'} + d_Q^{D'} = 1\,.$$

We can now rewrite the first moment of the margin and the Gibbs risk as

$$\begin{aligned}
\mu_1(M_Q^{D'}) &= s_Q^{D'} - e_Q^{D'} = 1-(2e_Q^{D'}+d_Q^{D'})\,,\\
R_{D'}(G_Q) &= \tfrac{1}{2}\,(1-s_Q^{D'}+e_Q^{D'}) = \tfrac{1}{2}\,(2e_Q^{D'}+d_Q^{D'})\,. \qquad (24)
\end{aligned}$$

Therefore, the third form of $\mathcal{C}$-bound of Theorem 11 can be rewritten as

$$\mathcal{C}_Q^{D'} = 1 - \frac{\Big(1-(2e_Q^{D'}+d_Q^{D'})\Big)^2}{1-2d_Q^{D'}}\,. \qquad (25)$$

5.3.2 Paired-Voters and Their Losses

This first generalization of the PAC-Bayesian theorem allows us to bound *separately* either d_Q^D, e_Q^D or s_Q^D, and therefore to bound $\mathcal{C}_Q^D$. To prove this result, we need to define a new kind of voter that we call a paired-voter.

Definition 24 Given two voters $f_i : \mathcal{X} \to [-1,1]$ and $f_j : \mathcal{X} \to [-1,1]$, the *paired-voter* $f_{ij} : \mathcal{X} \to [-1,1]^2$ outputs a tuple:

$$f_{ij}(x) \stackrel{\text{def}}{=} \langle\, f_i(x), f_j(x)\,\rangle\ .$$

Given a set of voters $\mathcal{H}$ weighted by a distribution Q on $\mathcal{H}$, we define a set of paired-voters $\mathcal{H}^2$ weighted by a distribution Q^2 as

$$\mathcal{H}^2 \stackrel{\text{def}}{=} \{f_{ij} : f_i, f_j \in \mathcal{H}\}\,, \quad \text{and} \quad Q^2(f_{ij}) \stackrel{\text{def}}{=} Q(f_i)\cdot Q(f_j)\,. \qquad (26)$$

We now present three losses for paired-voters. Remember that a loss function has the form $\overline{\mathcal{Y}}\times\mathcal{Y} \to [0,1]$, where $\overline{\mathcal{Y}}$ is the voter's output space. As a paired-voter output is a

[9] This is fairly intuitive in the case of binary voters. Indeed, given any example (x,y) and any two binary voters h_1, h_2, we have either: both voters misclassify the example – *i.e.*, $h_1(x) = h_2(x) \neq y$ –, both voters correctly classify the example – *i.e.*, $h_1(x) = h_2(x) = y$ –, or both voters disagree – *i.e.*, $h_1(x) \neq h_2(x)$.

tuple, our new loss functions map $[-1,1]^2 \times \{-1,1\}$ to $[0,1]$. Thus,

$$\begin{aligned}
\mathcal{L}_e\big(f_{ij}(x),\, y\big) &\overset{\text{def}}{=} \mathcal{L}_\ell(f_i(x), y) \cdot \mathcal{L}_\ell(f_j(x), y)\,,\\
\mathcal{L}_s\big(f_{ij}(x),\, y\big) &\overset{\text{def}}{=} \Big[1 - \mathcal{L}_\ell(f_i(x), y)\Big] \cdot \Big[1 - \mathcal{L}_\ell(f_j(x), y)\Big]\,,\\
\mathcal{L}_d\big(f_{ij}(x),\, y\big) &\overset{\text{def}}{=} \mathcal{L}_\ell(f_i(x){\cdot}f_j(x)\,,\, 1)\,. \qquad (27)
\end{aligned}$$

The key observation to understand the next theorems is that the expected losses of paired-voters $\mathcal{H}^2$ defined by Equation (26) allow one to recover the values of $e_Q^{D'}$, $s_Q^{D'}$ and $d_Q^{D'}$. Indeed, it directly follows from Definitions 3, 7 and 23, that

$$e_Q^{D'} = \mathop{\mathbf{E}}_{f_{ij}\sim Q^2} \mathbb{E}_{D'}^{\mathcal{L}_e}\Big(f_{ij}\Big)\,; \quad s_Q^{D'} = \mathop{\mathbf{E}}_{f_{ij}\sim Q^2} \mathbb{E}_{D'}^{\mathcal{L}_s}\Big(f_{ij}\Big)\,; \quad d_Q^{D'} = \mathop{\mathbf{E}}_{f_{ij}\sim Q^2} \mathbb{E}_{D'}^{\mathcal{L}_d}\Big(f_{ij}\Big)\,. \qquad (28)$$

5.4 PAC-Bayesian Theory For Losses of Paired-voters

As explained in Section 5.2, classical PAC-Bayesian theorems, like Corollaries 21 and 22, provide an upper bound on $R_D(G_Q)$ that holds uniformly for all posteriors Q. A bound on $R_D(B_Q)$ is typically obtained by multiplying the former bound by the usual factor of 2, as in PAC-Bound 0.

In this subsection, we present a first bound of $R_D(B_Q)$ relying on the $\mathcal{C}$-bound of Theorem 11. A uniform bound on $\mathcal{C}_Q^D$ is obtained using the third form of the $\mathcal{C}$-bound, through a bound on the Gibbs risk $R_D(G_Q)$ and another bound on the disagreement d_Q^D. The desired bound on $R_D(G_Q)$ is obtained by Corollary 21 as in PAC-Bound 0. To obtain a bound on d_Q^D, we capitalize on the notion of paired-voters presented in the previous section. This allows us to express two new PAC-Bayesian bounds on the risk of a majority vote, one for the supervised case and another for the semi-supervised case.

5.4.1 A PAC-Bayesian Theorem for e_Q^D, s_Q^D, or d_Q^D

The following PAC-Bayesian theorem can either bound the expected disagreement d_Q^D, the expected joint success s_Q^D or the expected joint error e_Q^D of a majority vote (see Definitions 7 and 23).

Theorem 25 *For any distribution D on $\mathcal{X}\times\{-1,1\}$, for any set $\mathcal{H}$ of voters $\mathcal{X}\to[-1,1]$, for any prior distribution P on $\mathcal{H}$, and any $\delta\in(0,1]$, we have*

$$\Pr_{S\sim D^m}\left(\begin{array}{l}\textit{For all posteriors } Q \textit{ on } \mathcal{H} :\\ \mathrm{kl}\big(\alpha_Q^S \,\big\|\, \alpha_Q^D\big) \;\le\; \dfrac{1}{m}\left[2{\cdot}\mathrm{KL}(Q\|P) + \ln\dfrac{\xi(m)}{\delta}\right]\end{array}\right) \ge 1-\delta\,,$$

where $\alpha_Q^{D'}$ can be either $e_Q^{D'}$, $s_Q^{D'}$ or $d_Q^{D'}$.

Proof Theorem 25 is deduced from Theorem 20. We present here the proof for $\alpha_Q^{D'} = e_Q^{D'}$. The two other cases are very similar.

Consider the set of paired-voters $\mathcal{H}^2$ and the posterior distribution Q^2 of Equation (26).

Also consider the prior distribution P^2 on $\mathcal{H}^2$ such that $P^2(f_{ij}) \stackrel{\text{def}}{=} P(f_i) \cdot P(f_j)$. Then we have,

$$\begin{aligned}
\mathrm{KL}(Q^2\|P^2) &= \mathop{\mathbf{E}}_{f_{ij}\sim Q^2} \ln \frac{Q^2(f_{ij})}{P^2(f_{ij})} = \mathop{\mathbf{E}}_{f_{ij}\sim Q^2} \ln \frac{Q(f_i)\cdot Q(f_j)}{P(f_i)\cdot P(f_j)} \\
&= \mathop{\mathbf{E}}_{f_{ij}\sim Q^2} \left[\ln \frac{Q(f_i)}{P(f_i)} + \ln \frac{Q(f_j)}{P(f_j)} \right] \\
&= 2\cdot \mathrm{KL}(Q\|P)\,.
\end{aligned}$$

Finally, from Equation (28), we have $\mathop{\mathbf{E}}_{f_{ij}\sim Q^2} \mathbb{E}_D^{\mathcal{L}_e}\Big(f_{ij}\Big) = e_Q^D$ and $\mathop{\mathbf{E}}_{f_{ij}\sim Q^2} \mathbb{E}_S^{\mathcal{L}_e}\Big(f_{ij}\Big) = e_Q^S$. Hence, by applying Theorem 20, we are done. ■

5.4.2 A New Bound for the Risk of the Majority Vote

Based on the fact that Theorem 25 gives a lower bound on the expected disagreement d_Q^D, we now derive PAC-Bound 1, which is a PAC-Bayesian bound for the $\mathcal{C}$-bound, and therefore, for the risk of the majority vote.

Given any prior distribution P on $\mathcal{H}$, we need the interval $\mathcal{R}_{Q,S}^{\delta}$ of Equation (23), together with

$$\mathcal{D}_{Q,S}^{\delta} \stackrel{\text{def}}{=} \left\{ d \,:\, \mathrm{kl}(d_Q^S\|\, d) \,\leq\, \frac{1}{m}\left[2\cdot\mathrm{KL}(Q\|P) + \ln\frac{\xi(m)}{\delta} \right] \right\}. \tag{29}$$

We then express the following bound on the Bayes risk.

PAC-Bound 1 *For any distribution D on $\mathcal{X}\times\{-1,1\}$, for any set $\mathcal{H}$ of voters $\mathcal{X}\to[-1,1]$, for any prior distribution P on $\mathcal{H}$, and any $\delta\in(0,1]$, we have*

$$\Pr_{S\sim D^m}\left(\forall\, Q \text{ on } \mathcal{H} \,:\, R_D(B_Q) \,\leq\, 1 - \frac{\left(1-2\cdot\sup\mathcal{R}_{Q,S}^{\delta/2}\right)^2}{1-2\cdot\inf\mathcal{D}_{Q,S}^{\delta/2}} \right) \,\geq\, 1-\delta\,,$$

where $\mathcal{R}_{Q,S}^{\delta/2}$ and $\mathcal{D}_{Q,S}^{\delta/2}$ are respectively defined by Equations (23) and (29).

Proof By Proposition 9, we have that $d_Q^S \leq \frac{1}{2}$. This, together with the facts that m is finite and $d_Q^S \in \mathcal{D}_{Q,S}^{\delta}$, implies that $\inf\mathcal{D}_{Q,S}^{\delta/2} < \frac{1}{2}$, and therefore that the denominator of the fraction in the statement of PAC-Bound 1 is always strictly positive.

Necessarily, $\sup\mathcal{R}_{Q,S}^{\delta/2} \leq \frac{1}{2}$. Let us consider the two following cases.

Case 1: $\sup\mathcal{R}_{Q,S}^{\delta/2} = \frac{1}{2}$. Then, $1-2\cdot\sup\mathcal{R}_{Q,S}^{\delta/2} = 0$, and the bound on $R_D(B_Q)$ is 1, which is trivially valid.

Case 2: $\sup\mathcal{R}_{Q,S}^{\delta/2} < \frac{1}{2}$. Then, we can apply the third form of Theorem 11 to obtain the upper bound on $R_D(B_Q)$. The desired bound is obtained by replacing d_Q^D by its lower bound

$\inf \mathcal{D}^{\delta/2}_{Q,S}$, and $R_D(G_Q)$, by its upper bound $\sup \mathcal{R}^{\delta/2}_{Q,S}$. The two bounds can therefore be deduced by suitably applying Corollary 21 (replacing δ by $\delta/2$) and Theorem 25 (replacing α^S_Q by d^S_Q, α^D_Q by d^D_Q and δ by $\delta/2$). ■

This bound has a major inconvenience: it degrades rapidly if the bounds on the numerator and the denominator are not tight. Note however that in the semi-supervised framework, we can achieve tighter results because the labels of the examples do not affect the value of $d^{D'}_Q$ (see Definition 7). Indeed, it is generally assumed in this framework that the learner has access to a huge amount m' of unlabeled data (*i.e.*, $m' \gg m$). One can then obtain a tighter bound of the disagreement. In this context, PAC-Bound 1' stated below is tighter than PAC-Bound 1.

PAC-Bound 1' (Semi-supervised bound) *For any distribution D on $\mathcal{X}\times\{-1,1\}$, for any set $\mathcal{H}$ of voters $\mathcal{X}\to[-1,1]$, for any prior distribution P on $\mathcal{H}$, and any $\delta\in(0,1]$, we have*

$$\Pr_{\substack{S\sim D^m\\ S_{\mathcal{U}}\sim D^{m'}_{unlabeled}}}\left(\forall\, Q \text{ on } \mathcal{H} \,:\, R_D(B_Q) \;\le\; 1-\frac{\left(1-2\cdot\sup\mathcal{R}^{\delta/2}_{Q,S}\right)^2}{1-2\cdot\inf\mathcal{D}^{\delta/2}_{Q,S_{\mathcal{U}}}}\right) \;\ge\; 1-\delta\,.$$

Proof In the presence of a large amount of unlabeled data (denoted by the set $S_{\mathcal{U}}$), one can use Corollary 25 to obtain an accurate lower bound of d^D_Q. An upper bound of $R_D(G_Q)$ can also be obtained via Corollary 21 but, this time, on the labeled data S. Thus, similarly as in the proof of PAC-Bound 1, the result follows from Theorem 11. ■

5.4.3 Computation of PAC-Bounds 1 and 1'

To compute PAC-Bound 1, we obtain the values of $r = \sup\mathcal{R}^{\delta/2}_{Q,S}$ and $d = \inf\mathcal{D}^{\delta/2}_{Q,S}$ by solving

$$\begin{aligned}
\mathrm{kl}\big(R_S(G_Q)\,\big\|\,r\big) &= \tfrac{1}{m}\big[\mathrm{KL}(Q\|P)+\ln\tfrac{\xi(m)}{\delta/2}\big]\,, && \text{with } R_S(G_Q)\le r\le\tfrac{1}{2}\,,\\
\text{and }\ \mathrm{kl}\big(d^S_Q\,\big\|\,d\big) &= \tfrac{1}{m}\big[2\cdot\mathrm{KL}(Q\|P)+\ln\tfrac{\xi(m)}{\delta/2}\big]\,, && \text{with } d\le d^S_Q\,.
\end{aligned}$$

These equations are very similar to the one we solved to compute PAC-Bound 0, as described in Section 5.2.2. Once r and d are computed, the bound on $R_D(B_Q)$ is given by $1-\frac{(1-2\cdot r)^2}{1-2\cdot d}$.

The same methodology can be used to compute PAC-Bound 1', except that in the semi-supervised setting, the disagreement is computed on the unlabeled data $S_{\mathcal{U}}$.

5.5 PAC-Bayesian Theory to Directly Bound the $\mathcal{C}$-bound

PAC-Bounds 1 and 1' of the last section require two approximations to upper bound $\mathcal{C}^D_Q$: one on $R_D(G_Q)$ and another on d^D_Q. We introduce below an extension to the PAC-Bayesian theory (Theorem 28) that enables us to directly bound $\mathcal{C}^D_Q$. To do so, we directly bound any pair of expectations among e^D_Q, s^D_Q and d^D_Q. For this reason, the new PAC-Bayesian theorem is based on a trivalent random variable instead of a Bernoulli one (which is bivalent). Note that Seeger (2003) and Seldin and Tishby (2010) have presented more general PAC-Bayesian theorems valid for k-valent random variables, for any positive integer k. However, our result leads to tighter bounds for the $k = 3$ case.

Before we get to this new PAC-Bayesian theorem (Theorem 28), we need some preliminary results.

5.5.1 A General PAC-Bayesian Theorem for Two Losses of Paired-Voters

Theorem 26 below allows us to simultaneously bound two losses of paired-voters. This result is inspired by the general PAC-Bayesian theorem for real-valued losses (Theorem 18).

Theorem 26 *For any distribution D on $\mathcal{X}\times\{-1,1\}$, for any set $\mathcal{H}$ of voters $\mathcal{X}\to[-1,1]$, for any two losses $\mathcal{L}_\alpha,\mathcal{L}_\beta : [-1,1]\times\{-1,1\}\to[0,1]$ with $\alpha,\beta\in\{e,s,d\}$, for any prior distribution P on $\mathcal{H}$, for any $\delta\in(0,1]$, for any $m'>0$, and for any convex function $\mathcal{D}(q_1,q_2\,\|\,p_1,p_2)$, we have*

$$\Pr_{S\sim D^m}\left(\begin{array}{l} \textit{For all posteriors } Q \textit{ on } \mathcal{H}: \\ \mathcal{D}\left(\mathop{\mathbf{E}}_{f_{ij}\sim Q^2}\mathbb{E}_S^{\mathcal{L}_\alpha}\big(f_{ij}\big), \mathop{\mathbf{E}}_{f_{ij}\sim Q^2}\mathbb{E}_S^{\mathcal{L}_\beta}\big(f_{ij}\big) \,\middle\|\, \mathop{\mathbf{E}}_{f_{ij}\sim Q^2}\mathbb{E}_D^{\mathcal{L}_\alpha}\big(f_{ij}\big), \mathop{\mathbf{E}}_{f_{ij}\sim Q^2}\mathbb{E}_D^{\mathcal{L}_\beta}\big(f_{ij}\big)\right) \\ \qquad\qquad\qquad\qquad\qquad\qquad \leq \quad \dfrac{1}{m'}\left[2\cdot\mathrm{KL}(Q\|P)+\ln\left(\dfrac{\Omega}{\delta}\right)\right] \end{array}\right) \geq 1-\delta\,,$$

where $\Omega \overset{\text{def}}{=} \mathop{\mathbf{E}}_{S\sim D^m}\ \mathop{\mathbf{E}}_{f_{ij}\sim P^2}\ e^{m'\cdot\mathcal{D}\left(\mathbb{E}_S^{\mathcal{L}_\alpha}\left(f_{ij}\right),\, \mathbb{E}_S^{\mathcal{L}_\beta}\left(f_{ij}\right) \,\middle\|\, \mathbb{E}_D^{\mathcal{L}_\alpha}\left(f_{ij}\right),\, \mathbb{E}_D^{\mathcal{L}_\beta}\left(f_{ij}\right)\right)}$.

Proof To simplify the notation, first let $\alpha_{ij}^{D'} \overset{\text{def}}{=} \mathbb{E}_{D'}^{\mathcal{L}_\alpha}\big(f_{ij}\big)$ and $\beta_{ij}^{D'} \overset{\text{def}}{=} \mathbb{E}_{D'}^{\mathcal{L}_\beta}\big(f_{ij}\big)$.

Now, since $\mathbf{E}_{f_{ij}\sim P^2}\, e^{m'\cdot\mathcal{D}\left(\alpha_{ij}^S,\beta_{ij}^S \,\middle\|\, \alpha_{ij}^D,\beta_{ij}^D\right)}$ is a positive random variable, Markov's inequality (Lemma 46, in Appendix A) can be applied to give

$$\Pr_{S\sim D^m}\left(\mathop{\mathbf{E}}_{f_{ij}\sim P^2} e^{m'\cdot\mathcal{D}\left(\alpha_{ij}^S,\beta_{ij}^S \,\middle\|\, \alpha_{ij}^D,\beta_{ij}^D\right)} \leq \frac{1}{\delta}\mathop{\mathbf{E}}_{S\sim D^m}\ \mathop{\mathbf{E}}_{f_{ij}\sim P^2} e^{m'\cdot\mathcal{D}\left(\alpha_{ij}^S,\beta_{ij}^S \,\middle\|\, \alpha_{ij}^D,\beta_{ij}^D\right)}\right) \geq 1-\delta\,.$$

By exploiting the fact that $\ln(\cdot)$ is an increasing function, and by the definition of Ω, we obtain

$$\Pr_{S\sim D^m}\left(\ln\left[\mathop{\mathbf{E}}_{f_{ij}\sim P^2} e^{m'\cdot\mathcal{D}\left(\alpha_{ij}^S,\beta_{ij}^S \,\middle\|\, \alpha_{ij}^D,\beta_{ij}^D\right)}\right] \ \leq\ \ln\left[\frac{\Omega}{\delta}\right]\right) \geq 1-\delta\,. \qquad (30)$$

We apply the change of measure inequality (Lemma 17) on the left side of innermost inequality, with $\phi(f) = m'\cdot\mathcal{D}\big(\alpha_{ij}^S,\beta_{ij}^S \,\big\|\, \alpha_{ij}^D,\beta_{ij}^D\big)$, $P = P^2$ and $Q = Q^2$. We then use Jensen's inequality (Lemma 47, in Appendix A), exploiting the convexity of $\mathcal{D}$:

$$
\begin{aligned}
&\ln\left[\mathop{\mathbf{E}}_{f_{ij}\sim P^2} e^{m'\cdot\mathcal{D}\left(\alpha^S_{ij},\beta^S_{ij}\,\middle\|\,\alpha^D_{ij},\beta^D_{ij}\right)}\right]\\
&\qquad\geq\; m' \mathop{\mathbf{E}}_{f_{ij}\sim Q^2} \mathcal{D}\big(\alpha^S_{ij},\beta^S_{ij}\,\big\|\,\alpha^D_{ij},\beta^D_{ij}\big) - \mathrm{KL}\big(Q^2\big\|P^2\big)\\
&\qquad\geq\; m'\cdot\mathcal{D}\left(\mathop{\mathbf{E}}_{f_{ij}\sim Q^2}\alpha^S_{ij}, \mathop{\mathbf{E}}_{f_{ij}\sim Q^2}\beta^S_{ij}\,\middle\|\,\mathop{\mathbf{E}}_{f_{ij}\sim Q^2}\alpha^D_{ij}, \mathop{\mathbf{E}}_{f_{ij}\sim Q^2}\beta^D_{ij}\right) - \mathrm{KL}\big(Q^2\big\|P^2\big)\\
&\qquad=\; m'\cdot\mathcal{D}\left(\mathop{\mathbf{E}}_{f_{ij}\sim Q^2}\alpha^S_{ij}, \mathop{\mathbf{E}}_{f_{ij}\sim Q^2}\beta^S_{ij}\,\middle\|\,\mathop{\mathbf{E}}_{f_{ij}\sim Q^2}\alpha^D_{ij}, \mathop{\mathbf{E}}_{f_{ij}\sim Q^2}\beta^D_{ij}\right) - 2\cdot\mathrm{KL}\big(Q\big\|P\big)\,.
\end{aligned}
$$

The last equality $\mathrm{KL}\big(Q^2\big\|P^2\big) = 2\cdot\mathrm{KL}\big(Q\big\|P\big)$ has been shown in the proof of Theorem 25. The result can then be straightforwardly obtained by inserting the last inequality into Equation (30). ■

5.5.2 A PAC-Bayesian Theorem for Any Pair Among e^D_Q, s^D_Q, and d^D_Q

In Section 5.1, Theorem 20 was obtained from Theorem 18. Similarly, the main theorem of this subsection (Theorem 28) is deduced from Theorem 26. However, a notable difference between Theorems 20 and 28 is that the former uses of the KL-divergence $\mathrm{kl}(\cdot\|\cdot)$ between distributions of two Bernoulli (*i.e.*, *bivalent*) random variables, and the latter uses the KL-divergence $\mathrm{kl}(\cdot,\cdot\|\cdot,\cdot)$ between distributions of two *trivalent* random variables.

Given two trivalent random variables Y_q and Y_p with $P(Y_q\!=\!a) = q_1$, $P(Y_q\!=\!b) = q_2$, $P(Y_q\!=\!c) = 1\!-\!q_1\!-\!q_2$, and $P(Y_p\!=\!a) = p_1$, $P(Y_p\!=\!b) = p_2$, $P(Y_p\!=\!c) = 1\!-\!p_1\!-\!p_2$, we denote by $\mathrm{kl}(q_1,q_2\,\|\,p_1,p_2)$ the Kullback-Leibler divergence between Y_q and Y_p. Thus, we have

$$
\mathrm{kl}(q_1,q_2\,\|\,p_1,p_2) \;\stackrel{\text{def}}{=}\; q_1\ln\frac{q_1}{p_1} + q_2\ln\frac{q_2}{p_2} + (1-q_1-q_2)\ln\frac{1-q_1-q_2}{1-p_1-p_2}\,. \tag{31}
$$

Note that $\mathrm{kl}\big(q_1,q_2\,\|\,p_1,p_2\big)$ is a shorthand notation for $\mathrm{KL}(Q\|P)$ of Equation (20), with $Q=(q_1,q_2,1\!-\!q_1\!-\!q_2)$ and $P=(p_1,p_2,1\!-\!p_1\!-\!p_2)$. Corollary 50 (in Appendix A) shows that $\mathrm{kl}\big(q_1,q_2\,\|\,p_1,p_2\big)$ is a convex function.

To be able to apply Theorem 26 with $\mathcal{D}(q_1,q_2\,\|\,p_1,p_2) = \mathrm{kl}(q_1,q_2\|p_1,p_2)$, we need Lemma 27 (below). This lemma is inspired by Lemma 19. However, in contrast with the latter, which is based on Maurer's lemma, Lemma 27 needs a generalization of it to trivalent random variables (instead of bivalent ones). The proof of this generalization is provided in Appendix A, listed as Lemma 52.

Lemma 27 *For any distribution D on $\mathcal{X}\times\{-1,1\}$, for any paired-voters f_{ij}, and any positive integer m, we have*

$$
\mathop{\mathbf{E}}_{S\sim D^m} e^{m\cdot\mathrm{kl}\left(\mathbb{E}_S^{\mathcal{L}_\alpha}(f_{ij}),\mathbb{E}_S^{\mathcal{L}_\beta}(f_{ij})\,\middle\|\,\mathbb{E}_D^{\mathcal{L}_\alpha}(f_{ij}),\mathbb{E}_D^{\mathcal{L}_\beta}\left(f_{ij}\right)\right)} \;\leq\; \xi(m)+m\,,
$$

where $\mathcal{L}_\alpha$ and $\mathcal{L}_\beta$ can be any two of the three losses $\mathcal{L}_s$, $\mathcal{L}_e$ or $\mathcal{L}_d$, and where $\xi(m)$ is defined at Equation (22). Therefore, $m+\sqrt{m} \;\leq\; \xi(m)+m \;\leq\; m+2\sqrt{m}\,.$

Proof Let Y_{ij} be a random variable that follows a multinomial distribution with three possible outcomes: $a \overset{\text{def}}{=} (1,0)$, $b \overset{\text{def}}{=} (0,1)$ and $c \overset{\text{def}}{=} (0,0)$. The "*Trinomial*" distribution is chosen such that $\Pr(Y_{ij}=a) = \mathbb{E}_D^{\mathcal{L}_\alpha}(f_{ij})$, $\Pr(Y_{ij}=b) = \mathbb{E}_D^{\mathcal{L}_\beta}(f_{ij})$ and $\Pr(Y_{ij}=c) = 1 - \mathbb{E}_D^{\mathcal{L}_\alpha}(f_{ij}) - \mathbb{E}_D^{\mathcal{L}_\beta}(f_{ij})$. Given m trials of Y_{ij}, we denote Y_{ij}^a, Y_{ij}^b and Y_{ij}^c the number of times each outcome is observed. Note that Y_{ij} is totally defined by (Y_{ij}^a, Y_{ij}^b), since $Y_{ij}^c = m - Y_{ij}^a - Y_{ij}^b$. We thus use the notation

$$Y_{ij} \;=\; (Y_{ij}^a, Y_{ij}^b) \;\sim\; \mathcal{T}_{ij} \;\;\overset{\text{def}}{=}\;\; \text{Trinomial}\Big(m,\, \mathbb{E}_D^{\mathcal{L}_\alpha}(f_{ij}),\, \mathbb{E}_D^{\mathcal{L}_\beta}(f_{ij})\Big).$$

Hence, we have

$$\Pr_{(Y_{ij}^a,Y_{ij}^b)\sim\mathcal{T}_{ij}}\Big(Y_{ij}^a\!=\!k_1 \wedge Y_{ij}^b\!=\!k_2\Big) = \tbinom{m}{k_1}\tbinom{m-k_1}{k_2}\big[\mathbb{E}_S^{\mathcal{L}_\alpha}(f_{ij})\big]^{k_1}\big[\mathbb{E}_S^{\mathcal{L}_\beta}(f_{ij})\big]^{k_2}\big[1-\mathbb{E}_S^{\mathcal{L}_\alpha}(f_{ij})-\mathbb{E}_S^{\mathcal{L}_\beta}(f_{ij})\big]^{m-k_1-k_2},$$

for any $k_1 \in \{0,..,m\}$ and any $k_2 \in \{0,..,m-k_1\}$.

Now, applying Lemma 52 to the convex function $e^{m\,\cdot\,\mathrm{kl}\left(\,\cdot\,,\,\cdot\,\,\|\,\mathbb{E}_D^{\mathcal{L}_\alpha}(f_{ij}),\mathbb{E}_D^{\mathcal{L}_\beta}(f_{ij})\right)}$, and by the definition of $\mathrm{kl}(\cdot,\cdot\|\cdot,\cdot)$, we have

$$\begin{aligned}
&\mathop{\mathbf{E}}_{S\sim D^m}\; e^{m\,\cdot\,\mathrm{kl}\left(\mathbb{E}_S^{\mathcal{L}_\alpha}(f_{ij}),\mathbb{E}_S^{\mathcal{L}_\beta}(f_{ij})\,\middle\|\,\mathbb{E}_D^{\mathcal{L}_\alpha}(f_{ij}),\mathbb{E}_D^{\mathcal{L}_\beta}\left(f_{ij}\right)\right)} \\
&\qquad\le\; \mathop{\mathbf{E}}_{(Y_{ij}^a,Y_{ij}^b)\sim\mathcal{T}_{ij}}\; e^{m\,\cdot\,\mathrm{kl}\left(\frac{1}{m}Y_{ij}^a,\frac{1}{m}Y_{ij}^b\,\middle\|\,\mathbb{E}_D^{\mathcal{L}_\alpha}(f_{ij}),\mathbb{E}_D^{\mathcal{L}_\beta}\left(f_{ij}\right)\right)} \\
&\qquad=\; \mathop{\mathbf{E}}_{(Y_{ij}^a,Y_{ij}^b)\sim\mathcal{T}_{ij}}\; \left(\frac{\frac{1}{m}Y_{ij}^a}{\mathbb{E}_S^{\mathcal{L}_\alpha}(f_{ij})}\right)^{Y_{ij}^a}\left(\frac{\frac{1}{m}Y_{ij}^b}{\mathbb{E}_S^{\mathcal{L}_\beta}(f_{ij})}\right)^{Y_{ij}^b}\left(\frac{1-\frac{1}{m}Y_{ij}^a-\frac{1}{m}Y_{ij}^b}{1-\mathbb{E}_S^{\mathcal{L}_\alpha}(f_{ij})-\mathbb{E}_S^{\mathcal{L}_\beta}(f_{ij})}\right)^{m-Y_{ij}^a-Y_{ij}^b}.
\end{aligned}$$

As Y_{ij} follows a trinomial law, we then have

$$\begin{aligned}
&\mathop{\mathbf{E}}_{(Y_{ij}^a,Y_{ij}^b)\sim\mathcal{T}_{ij}}\; \left(\frac{\frac{1}{m}Y_{ij}^a}{\mathbb{E}_S^{\mathcal{L}_\alpha}(f_{ij})}\right)^{Y_{ij}^a}\left(\frac{\frac{1}{m}Y_{ij}^b}{\mathbb{E}_S^{\mathcal{L}_\beta}(f_{ij})}\right)^{Y_{ij}^b}\left(\frac{1-\frac{1}{m}Y_{ij}^a-\frac{1}{m}Y_{ij}^b}{1-\mathbb{E}_S^{\mathcal{L}_\alpha}(f_{ij})-\mathbb{E}_S^{\mathcal{L}_\beta}(f_{ij})}\right)^{m-Y_{ij}^a-Y_{ij}^b} \\
&\quad=\; \sum_{k_1=0}^{m}\sum_{k_2=0}^{m-k_1}\Bigg[\Pr_{(Y_{ij}^a,Y_{ij}^b)\sim\mathcal{T}_{ij}}\Big(Y_{ij}^a = k_1 \wedge Y_{ij}^b = k_2\Big) \\
&\qquad\qquad\times\left(\frac{\frac{k_1}{m}}{\mathbb{E}_S^{\mathcal{L}_\alpha}(f_{ij})}\right)^{k_1}\left(\frac{\frac{k_2}{m}}{\mathbb{E}_S^{\mathcal{L}_\beta}(f_{ij})}\right)^{k_2}\left(\frac{1-\frac{k_1}{m}-\frac{k_2}{m}}{1-\mathbb{E}_S^{\mathcal{L}_\alpha}(f_{ij})-\mathbb{E}_S^{\mathcal{L}_\beta}(f_{ij})}\right)^{m-k_1-k_2}\Bigg] \\
&\quad=\; \sum_{k_1=0}^{m}\sum_{k_2=0}^{m-k_1}\Bigg[\binom{m}{k_1}\binom{m-k_1}{k_2}\Big(\mathbb{E}_S^{\mathcal{L}_\alpha}(f_{ij})\Big)^{k_1}\Big(\mathbb{E}_S^{\mathcal{L}_\beta}(f_{ij})\Big)^{k_2}\Big(1-\mathbb{E}_S^{\mathcal{L}_\alpha}(f_{ij})-\mathbb{E}_S^{\mathcal{L}_\beta}(f_{ij})\Big)^{m-k_1-k_2} \\
&\qquad\qquad\times\left(\frac{\frac{k_1}{m}}{\mathbb{E}_S^{\mathcal{L}_\alpha}(f_{ij})}\right)^{k_1}\left(\frac{\frac{k_2}{m}}{\mathbb{E}_S^{\mathcal{L}_\beta}(f_{ij})}\right)^{k_2}\left(\frac{1-\frac{k_1}{m}-\frac{k_2}{m}}{1-\mathbb{E}_S^{\mathcal{L}_\alpha}(f_{ij})-\mathbb{E}_S^{\mathcal{L}_\beta}(f_{ij})}\right)^{m-k_1-k_2}\Bigg] \\
&\quad=\; \sum_{k_1=0}^{m}\sum_{k_2=0}^{m-k_1}\binom{m}{k_1}\binom{m-k_1}{k_2}\left(\frac{k_1}{m}\right)^{k_1}\left(\frac{k_2}{m}\right)^{k_2}\left(1-\frac{k_1}{m}-\frac{k_2}{m}\right)^{m-k_1-k_2} \\
&\quad=\; \xi(m)+m\,.
\end{aligned}$$

The last equality has been proven by Younsi (2012). Recall that $\xi(m)$ is defined by Equation (22). ■

We are now ready to present the main result of this section. By bounding any pair of expectations among e_Q^D, s_Q^D and d_Q^D, Theorem 28 is the perfect tool to directly bound the $\mathcal{C}$-bound.

Theorem 28 *For any distribution D on $\mathcal{X}\times\{-1,1\}$, for any set $\mathcal{H}$ of voters $\mathcal{X}\to[-1,1]$, for any prior distribution P on $\mathcal{H}$, and any $\delta\in(0,1]$, we have*

$$\Pr_{S\sim D^m}\left(\begin{array}{l}\textit{For all posteriors } Q \textit{ on } \mathcal{H}:\\ \mathrm{kl}\big(\,\alpha_Q^S,\beta_Q^S\,\big\|\,\alpha_Q^D,\beta_Q^D\,\big)\;\le\;\dfrac{1}{m}\left[2{\cdot}\mathrm{KL}(Q\|P)+\ln\dfrac{\xi(m)+m}{\delta}\right]\end{array}\right)\ge 1-\delta\,,$$

where $\alpha_Q^{D'}$ and $\beta_Q^{D'}$ can be any two distinct choices among $d_Q^{D'}$, $e_Q^{D'}$ and $s_Q^{D'}$.

Proof The result follows from Theorem 26 with $\mathcal{D}(q_1,q_2\,\|\,p_1,p_2)=\mathrm{kl}(q_1,q_2\,\|\,p_1,p_2)$ and $m'=m$. Since Equation (28) shows that $\alpha_Q^{D'}=\mathop{\mathbf{E}}_{f_{ij}\sim Q^2}\alpha_{ij}^{D'}$ and $\beta_Q^{D'}=\mathop{\mathbf{E}}_{f_{ij}\sim Q^2}\beta_{ij}^{D'}$, we have

$$\Pr_{S\sim D^m}\Bigg(\forall\, Q \text{ on}\,\mathcal{H}\colon\ \mathrm{kl}\left(\alpha_Q^S,\beta_Q^S\,\|\,\alpha_Q^D,\beta_Q^D\right)\le \frac{1}{m}\left[2\cdot\mathrm{KL}(Q\|P)+\ln\left(\frac{1}{\delta}\mathop{\mathbf{E}}_{S\sim D^m}\ \mathop{\mathbf{E}}_{f_{ij}\sim P^2}\ e^{m\mathrm{kl}\left(\alpha_{ij}^S,\beta_{ij}^S\,\|\,\alpha_{ij}^D,\beta_{ij}^D\right)}\right)\right]\Bigg)\ge 1-\delta\,.$$

As the prior distribution P^2 is independent of S, we can swap the two expectations in expression $\mathop{\mathbf{E}}_{S\sim D^m}\ \mathop{\mathbf{E}}_{f_{ij}\sim P^2} e^{m\mathrm{kl}(\alpha_{ij}^S,\beta_{ij}^S\,\|\,\alpha_{ij}^D,\beta_{ij}^D)}$. This observation, together with Lemma 27, gives

$$\begin{aligned}\mathop{\mathbf{E}}_{S\sim D^m}\ \mathop{\mathbf{E}}_{f_{ij}\sim P^2}\ e^{m\mathrm{kl}\left(\alpha_{ij}^S,\beta_{ij}^S\,\|\,\alpha_{ij}^D,\beta_{ij}^D\right)} &= \mathop{\mathbf{E}}_{f_{ij}\sim P^2}\ \mathop{\mathbf{E}}_{S\sim D^m}\ e^{m\mathrm{kl}\left(\alpha_{ij}^S,\beta_{ij}^S\,\|\,\alpha_{ij}^D,\beta_{ij}^D\right)}\\ &\le \mathop{\mathbf{E}}_{f_{ij}\sim P^2}\ \xi(m)+m\\ &= \xi(m)+m\,.\end{aligned}$$

■

A first version of Theorem 28 was proposed by Lacasse et al. (2006), with the difference that $\ln\frac{(m+1)(m+2)}{2\delta}$ in the latter is now replaced by $\ln\frac{\xi(m)+m}{\delta}$ in the former. Since $\xi(m)+m<\frac{(m+1)(m+2)}{2}$, the new theorem is therefore tighter.

5.5.3 Another Bound for the Risk of the Majority Vote

First, we need the following notation that is related to Theorem 28. Given any prior distribution P on $\mathcal{H}$,

$$\mathcal{A}_{Q,S}^{\delta}\ \stackrel{\text{def}}{=}\ \left\{(d,e)\ :\ \mathrm{kl}(d_Q^S,e_Q^S\|\,d,e)\ \le\ \frac{1}{m}\left[2{\cdot}\mathrm{KL}(Q\|P)+\ln\tfrac{\xi(m)+m}{\delta}\right]\right\}.\qquad(32)$$

The bound is obtained by seeking the point of $\mathcal{A}_{Q,S}^{\delta}$ maximizing the $\mathcal{C}$-bound. Since a point (d,e) of $\mathcal{A}_{Q,S}^{\delta}$ expresses a disagreement d and a joint error e, we directly compute the bound on $\mathcal{C}_Q^D$ using Equation (25).

Note however that $\mathcal{A}^{\delta}_{Q,S}$ can contain points that are not possible in practice, *i.e.*, points that are not achievable with any data-generating distribution D. Indeed, by Proposition 9, we know that

$$d^{D}_{Q} \;\leq\; 2 \cdot R_D(G_Q) \cdot \big(1 - R_D(G_Q)\big)\,.$$

Based on this property, it is possible to significantly reduce the achievable region of $\mathcal{A}^{\delta}_{Q,S}$. To do so, we must first rewrite this property based on d^{D}_{Q} and e^{D}_{Q} only.

$$\begin{array}{rrcl} & d^{D}_{Q} & \leq & 2 \cdot R_D(G_Q) \cdot \big(1 - R_D(G_Q)\big) \;=\; 2 \cdot \big(e^{D}_{Q} + \tfrac{1}{2} d^{D}_{Q}\big) \cdot \big(1 - (e^{D}_{Q} + \tfrac{1}{2} d^{D}_{Q})\big) \\ \Leftrightarrow & 0 & \leq & -\tfrac{1}{2}(d^{D}_{Q})^2 - 2e^{D}_{Q} \cdot d^{D}_{Q} + 2e^{D}_{Q} - 2(e^{D}_{Q})^2 \\ \Leftrightarrow & d^{D}_{Q} & \leq & 2 \cdot \left(\sqrt{e^{D}_{Q}} - e^{D}_{Q}\right). \end{array} \qquad (33)$$

Note also that if $R_D(G_Q) \geq \frac{1}{2}$, there is no bound on $R_D(B_Q)$ better than the trivial one $R_D(B_Q) \leq 1$. We therefore consider only the pairs $(d, e) \in \mathcal{A}^{\delta}_{Q,S}$ that do not correspond to that situation. Since $R_D(G_Q) = \frac{1}{2}(2e^{D}_{Q} + d^{D}_{Q})$ (Equation 24), this is therefore equivalent to considering only the pairs (d, e) such that $2e + d < 1$. We later show that this still gives a valid bound. Thus, from all these ideas, we restrain $\mathcal{A}^{\delta}_{Q,S}$ (Equation 32) as follows:

$$\widetilde{\mathcal{A}}^{\delta}_{Q,S} \;\stackrel{\text{def}}{=}\; \left\{(d, e) \in \mathcal{A}^{\delta}_{Q,S} \;:\; d \;\leq\; 2(\sqrt{e} - e) \quad \text{and} \quad 2e + d < 1\right\}, \qquad (34)$$

and obtain the following bound that, in contrast with PAC-Bound 1, directly bounds $\mathcal{C}^{D}_{Q}$.

PAC-Bound 2 *For any distribution D on $\mathcal{X} \times \{-1, 1\}$, for any set $\mathcal{H}$ of voters $\mathcal{X} \to [-1, 1]$, for any prior distribution P on $\mathcal{H}$, and any $\delta \in (0, 1]$, we have*

$$\Pr_{S \sim D^m}\left(\forall Q \text{ on } \mathcal{H} \;:\; R_D(B_Q) \leq \sup_{(d,e) \in \widetilde{\mathcal{A}}^{\delta}_{Q,S}} \left[1 - \frac{\Big(1 - (2e + d)\Big)^2}{1 - 2d}\right]\right) \geq 1 - \delta\,.$$

Proof We need to show that the supremum value in the statement of PAC-Bound 2 is a valid upper bound of $R_D(B_Q)$. Note that if $\widetilde{\mathcal{A}}^{\delta}_{Q,S} = \emptyset$, then the supremum is $+\infty$, and the bound is trivially valid. Therefore, we assume below that $\widetilde{\mathcal{A}}^{\delta}_{Q,S}$ is not empty.

Let us consider $(d, e) \in \widetilde{\mathcal{A}}^{\delta}_{Q,S}$. From the conditions $d \leq 2(\sqrt{e} - e)$ and $2e + d < 1$, it follows by straightforward calculations that $d < \frac{1}{2}$. This implies that

$$1 - \frac{\big(1 - (2e + d)\big)^2}{1 - 2d} \;<\; 1\,,$$

because both the numerator and the denominator of the fraction are strictly positive (remember that $2e + d < 1$). Thus, the supremum is at most 1.

Let us consider the three following cases.

Case 1: The supremum is not attained in $\widetilde{\mathcal{A}}^{\delta}_{Q,S}$. Note that as $\widetilde{\mathcal{A}}^{\delta}_{Q,S}$ is a subset of $\mathbb{R}^2$, the supremum must be attained for a pair in the closure of $\widetilde{\mathcal{A}}^{\delta}_{Q,S}$. The latter is not a closed set only because of its $2e + d < 1$ constraint. Therefore, the supremum is achieved for a pair (d, e) in the closure for which $1 - (2e + d) = 0$, implying that the value of the supremum is in that case 1, which trivially is a valid bound for $R_D(B_Q)$.

Case 2: The supremum is attained in $\widetilde{\mathcal{A}}^{\delta}_{Q,S}$ *and has value* 1. In that case, the bound is again trivially valid.

Case 3: The supremum is attained in $\widetilde{\mathcal{A}}^{\delta}_{Q,S}$ *and has a value strictly lower than* 1. In that case, there must be an $\epsilon > 0$ such that $2e + d < 1 - \epsilon$ for all $(d, e) \in \widetilde{\mathcal{A}}^{\delta}_{Q,S}$. Hence, because of Equation (33) and Theorem 28, we have that $2e^{D}_{Q} + d^{D}_{Q} < 1 - \epsilon$ with probability $1-\delta$. Since $R_D(G_Q) = \frac{1}{2}(2e^{D}_{Q} + d^{D}_{Q})$ (Equation 24), this implies that, with probability $1-\delta$, $R_D(G_Q) < 1/2 - 1/2\epsilon$. Hence, with probability $1-\delta$, Theorem 11 is valid – *i.e.*, $\mathcal{C}^{D}_{Q}$ bounds $R_D(B_Q)$ – and $(d^{D}_{Q}, e^{D}_{Q}) \in \widetilde{\mathcal{A}}^{\delta}_{Q,S}$. Thus,

$$R_D(B_Q) \;\le\; \mathcal{C}^{D}_{Q} \;=\; 1 - \frac{\Big(1 - (2e^{D}_{Q} + d^{D}_{Q})\Big)^2}{1 - 2d^{D}_{Q}} \;\le\; \sup_{(d,e)\in\widetilde{\mathcal{A}}^{\delta}_{Q,S}} \left[1 - \frac{\Big(1 - (2e + d)\Big)^2}{1 - 2d}\right],$$

and we are done. ■

In some situations, we can slightly improve PAC-Bound 2 by bounding the joint error e^{D}_{Q} via Theorem 25 with δ replaced by $\delta/2$. This removes all pairs (d, e) such that e does not belong to the set $\mathcal{E}^{\delta/2}_{Q,S}$ defined as

$$\mathcal{E}^{\delta/2}_{Q,S} \;\stackrel{\text{def}}{=}\; \left\{e \;:\; \mathrm{kl}(e^{S}_{Q} \| \, e) \;\le\; \frac{1}{m}\left[2\cdot\mathrm{KL}(Q\|P) + \ln\frac{\xi(m)}{\delta/2}\right]\right\}.$$

Then, by applying PAC-Bound 2, with δ replaced by $\delta/2$, one can obtain the following slightly improved bound.

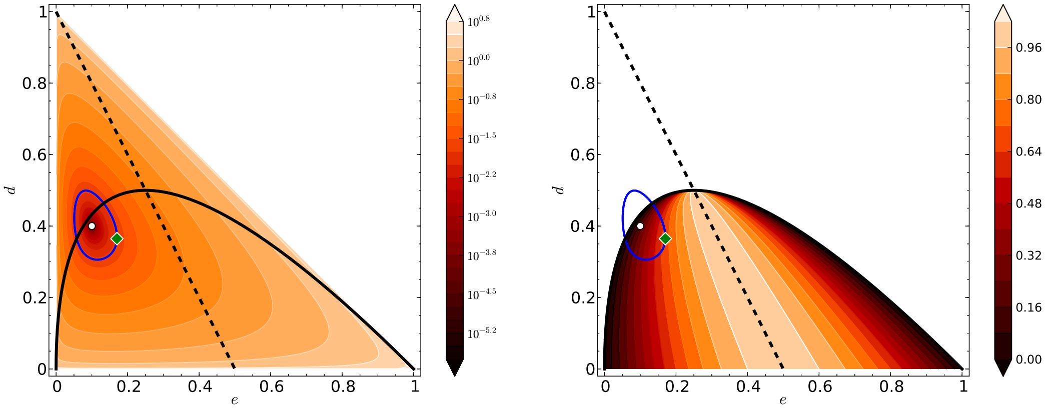
PAC-Bound 2' *For any distribution D on $\mathcal{X}\times\{-1,1\}$, for any set $\mathcal{H}$ of voters $\mathcal{X} \to [-1,1]$, for any prior distribution P on $\mathcal{H}$, and any $\delta \in (0,1]$, we have*

$$\Pr_{S\sim D^m}\left(\forall\, Q \text{ on } \mathcal{H} \;:\; R_D(B_Q) \;\le\; \sup_{(d,e)\in\widehat{\mathcal{A}}^{\delta/2}_{Q,S}} \left[1 - \frac{\Big(1 - (2e + d)\Big)^2}{1 - 2d}\right]\right) \;\ge\; 1 - \delta\,,$$

where

$$\widehat{\mathcal{A}}^{\delta/2}_{Q,S} \;\stackrel{\text{def}}{=}\; \left\{(d, e) \in \mathcal{A}^{\delta/2}_{Q,S} \;:\; d \;\le\; 2(\sqrt{e} - e)\,,\;\; 2e + d < 1 \;\text{ and }\; e \;\le\; \sup \mathcal{E}^{\delta/2}_{Q,S}\right\}. \qquad (35)$$

Proof Immediate consequence of Theorem 25, PAC-Bound 2, and the union bound. ■



(a) Contour plot of $\mathrm{kl}(0.4, 0.1\|d, e)$.

(b) Contour plot of $F_{\mathcal{C}}(d, e)$.

Figure 5: Example of application of PAC-Bound 2. We suppose that $\mathrm{KL}(Q\|P) = 5$, $m = 1000$ and $\delta = 0.05$. If we observe an empirical joint error $e_Q^S = 0.10$ and an empirical disagreement $d_Q^S = 0.40$ (thus, a Gibbs risk $R_S(G_Q) = 0.1 + \frac{1}{2} \cdot 0.4 = 0.30$), then we need to maximize the function $F_{\mathcal{C}}(d, e)$ over the domain $\widetilde{\mathcal{A}}_{Q,S}^{\delta}$ given by three constraints: $\mathrm{kl}(0.4, 0.1\|\, d, e) \leq \frac{1}{m}\big[2\mathrm{KL}(Q\|P) + \ln\frac{\xi(m)+m}{\delta}\big] \approx 0.0199$ (blue oval), $d \leq 2(\sqrt{e} - e)$ (black curve) and $2e + d < 1$ (black dashed line). Therefore, we obtain a bound $R_D(B_Q) \leq 0.679$ (corresponding to the green diamond marker).

5.5.4 Computation of PAC-Bounds 2 and 2'

Let us consider the $\mathcal{C}$-bound as a function $F_{\mathcal{C}}$ of two variables $(d, e) \in [0, \frac{1}{2}] \times [0, 1]$, instead of a function of the distribution Q.

$$F_{\mathcal{C}}(d, e) \ \stackrel{\text{def}}{=}\ 1 - \frac{\big[\,1 - (2e + d)\,\big]^2}{1 - 2d}\,. \tag{36}$$

Proposition 54 (provided in Appendix A) shows that $F_{\mathcal{C}}$ is a concave function. Therefore, PAC-Bound 2 is obtained by maximizing $F_{\mathcal{C}}(d, e)$ in the domain $\widetilde{\mathcal{A}}_{Q,S}^{\delta}$ (Equation 34), which is both bounded and convex. Several optimization methods can achieve this. In our experiments, we decompose $F_{\mathcal{C}}(d, e)$ in two nested functions of a single argument:

$$\sup_{(d,e)\in\widetilde{\mathcal{A}}_{Q,S}^{\delta}} \Big[F_{\mathcal{C}}(d, e)\Big] \;=\; \sup_{d:(d,\cdot)\in\widetilde{\mathcal{A}}_{Q,S}^{\delta}} \Big[F_{\mathcal{C}}^*(d)\Big], \quad \text{where } F_{\mathcal{C}}^*(d) \ \stackrel{\text{def}}{=} \sup_{e:(d,e)\in\widetilde{\mathcal{A}}_{Q,S}^{\delta}} \Big[F_{\mathcal{C}}(d, e)\Big]\,.$$

Thus, we implement the maximization of $F_{\mathcal{C}}$ using a one-dimensional optimization algorithm twice. Figure 5 shows an application example of PAC-Bound 2.

The computation of PAC-Bound 2' is done using the same method, but we optimize over the domain $\widehat{\mathcal{A}}_{Q,S}^{\delta/2}$ (Equation 35) instead of $\widetilde{\mathcal{A}}_{Q,S}^{\delta}$, which is also bounded and convex. Of course, this requires computing $\sup \mathcal{E}_{Q,S}^{\delta/2}$ beforehand, using the same technique as for PAC-Bounds 0, 1 and 1'. Figure 6 shows an application example of PAC-Bound 2'.

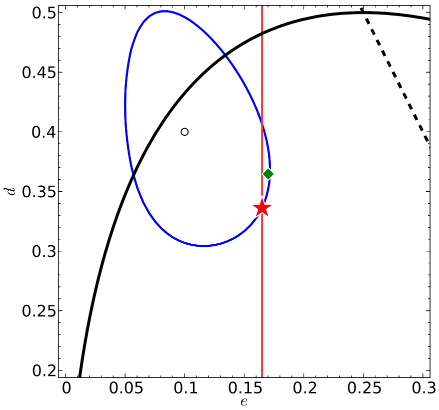


Figure 6: Example of application of PAC-Bound 2'. We use the same quantities as for Figure 5. The red vertical line corresponds to the upper bound on the joint error, resulting in an improved bound of $R_D(B_Q) \leq 0.660$ (corresponding to the red star marker). Note however that, even if the bound here is tighter, the egg-region is a bit bigger than in the case of PAC-Bound 2 because all the δ has been replaced by $\delta/2$.

5.6 Empirical Comparison Between PAC-Bounds on the Bayes Risk $R_D(B_Q)$

We now propose an empirical comparison of all PAC-Bounds we presented so far. The numerical results of Figure 7 are obtained by using AdaBoost (Schapire and Singer, 1999) with decision stumps on the Mushroom UCI data set (which contains 8124 examples). This data set is randomly split into two halves: one training set S and one testing set T. For each round of boosting, we compute the usual PAC-Bayesian bound of twice the Gibbs risk (PAC-Bound 0) of the corresponding majority vote classifier, as well as the other variants of the PAC-Bayesian bounds presented in this paper.

We can see that PAC-Bound 1 is generally tighter than PAC-Bound 0, and we obtain a substantial improvement with PAC-Bound 2. Almost no improvement is obtained with PAC-Bound 2' in that case. We can also see that using unlabeled data to estimate d_Q^D helps, as PAC-Bound 1' is the tightest.[10]

However, we see in Figure 7 that after 8 rounds of boosting, all the bounds are degrading even if the value of $\mathcal{C}_Q^S$ continues to decrease. This drawback is due to the fact that the denominator of $\mathcal{C}_Q^S$ tends to 0, that is the second moment of the margin $\mu_2(M_Q^S)$ is close to 0 (see the first or the second forms of Theorem 11). Hence, in this context, the first moment of the margin $\mu_1(M_Q^S)$ must be small as well. Thus, any slack in the bound of $\mu_1(M_Q^D)$ has a multiplicative effect on each of the three proposed PAC-bounds of $R_D(B_Q)$. Unfortunately, Boosting algorithms tend to construct majority votes with $\mu_1(M_Q^S)$ just slightly larger than 0.

[10] To obtain PAC-Bound 1', we simulate the case where we have access to a large number of unlabeled data by simply using the empirical value of d_Q^T computed on the testing set.

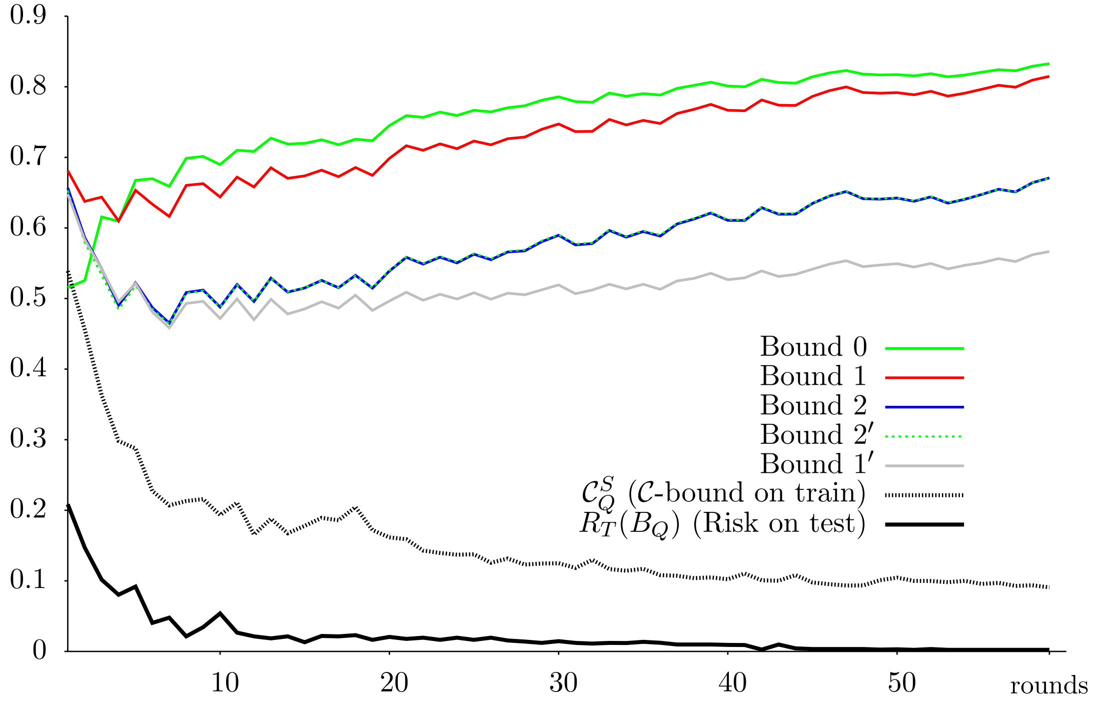


Figure 7: Comparison of bounds of $R_D(B_Q)$ during 60 rounds of Boosting.

6. PAC-Bayesian Bounds without KL

Having PAC-Bayesian theorems that bound the difference between $\mathcal{C}_Q^S$ and $\mathcal{C}_Q^D$ opens the way to structural $\mathcal{C}$-bound minimization algorithms. As for most PAC-Bayesian results, the bound on $\mathcal{C}_Q^D$ depends on an empirical estimate of it, and on the Kullback-Leibler divergence $\mathrm{KL}(Q\|P)$ between the output distribution Q and the a priori defined distribution P. In this section, we present a theoretical extension of our PAC-Bayesian approach that is mandatory to develop the $\mathcal{C}_Q^D$-minimization algorithm of Section 8.

The next theorems introduce PAC-Bayesian bounds that have the surprising property of having no KL term. This new approach is driven by the fact that our attempts to construct algorithms that minimize any of the PAC-Bounds presented in the previous section ended up being unsuccessful. Surprisingly, the KL-divergence is a poor regularizer in this case, as its empirical value tends to be overweighted in comparison with the empirical value of the $\mathcal{C}$-bound (*i.e.*, $\mathcal{C}_Q^S$).

There have already been some attempts to develop PAC-Bayesian bounds that do not rely on the KL-divergence (see the localized priors of Catoni, 2007, or the distribution-dependent priors of Lever et al., 2013). The usual idea is to bound the KL-divergence via some concentration inequality. In the following, the KL term simply vanishes from the bound, provided that we restrict ourselves to *aligned posteriors*, a notion that is properly defined later on in this section. The fact that these new PAC-Bayesian bounds do not contain any KL divergence terms indicates that the restriction to aligned posteriors has some "built in" regularization action.

The following theory is similar to the one used by Germain et al. (2011), in which two learning algorithms inspired by the PAC-Bayesian theory are compared: one regularized with the KL divergence, using a hyperparameter to control its weight, and one regularized by restricting the posterior distributions to be *aligned* on the prior distribution. Surprisingly, the latter algorithm uses one less parameter, and has been shown to have an as good accuracy.

6.1 Self-Complemented Sets of Voters and Aligned Distributions

In this section, we assume that the (possibly infinite) set of voters $\mathcal{H}$ is *self-complemented*[11].

Definition 29 A set of voters $\mathcal{H}$ is said to be *self-complemented* if there exists a bijection $c : \mathcal{H} \to \mathcal{H}$ such that for any $f \in \mathcal{H}$,

$$c(f) \;=\; -f\,.$$

Moreover, we say that a distribution Q on any self-complemented $\mathcal{H}$ is *aligned* on a prior distribution P if

$$Q(f) + Q(c(f)) \;=\; P(f) + P(c(f))\,, \quad \forall f \in \mathcal{H}\,.$$

When P is the uniform prior distribution and Q is aligned on P, we say that Q is *quasi-uniform*. Note that the uniform distribution is itself a quasi-uniform distribution.

In the finite case, we consider self-complemented sets $\mathcal{H}$ of $2n$ voters $\mathcal{X} \to \overline{\mathcal{Y}}$. In this setting, for any $x \in \mathcal{X}$ and any $i \in \{1, \ldots, n\}$, we have that $f_{i+n}(x) = -f_i(x)$. Moreover, finite quasi-uniform distributions Q is such that for any $i \in \{1, \ldots, n\}$,

$$Q(f_i) + Q(f_{i+n}) \;=\; \frac{1}{n}\,. \tag{37}$$

Equation (37) shows that when a distribution Q is restricted to being quasi-uniform, the sum of the weight given to a pair of complementary voters is equal to $\frac{1}{n}$. As Q is a distribution, this means that the weight of any voter is lower-bounded by 0 and upper-bounded by $\frac{1}{n}$, giving rise to an L_∞-norm regularization. Note that, in this context, the maximum value of $\mathrm{KL}(Q\|P)$ is reached when all voters have a weight of either 0 or $\frac{1}{n}$. Indeed, a quasi-uniform distribution Q is such that $\mathrm{KL}(Q\|P) \leq n(\frac{1}{n})\ln(\frac{1}{n}/\frac{1}{2n}) = \ln 2$. Consequently, the value of the KL term is necessarily small and plays a little role in PAC-Bayesian bounds computed with quasi-uniform distributions. The following theorems and corollaries are specializations that allow to slightly improve these PAC-Bayesian bounds by getting rid of the KL term completely. To achieve these results, the associated proofs require restrictions on the choice of convex function $\mathcal{D}$ and loss function $\mathcal{L}$.

[11] In Laviolette et al. (2011), this notion was introduced as an *auto-complemented* set of voters. However, *self-complemented* is a more suitable name. Also, note that a similar notion, called a *symmetric hypothesis class*, is introduced in Daniely et al. (2013).

6.2 PAC-Bayesian Theorems without KL for the Gibbs Risk

Let us first specialize Theorem 18 to aligned distributions and linear loss $\mathcal{L}_\ell$. We first need a new change of measure inequality, as this is the part of Theorem 18 where the KL term appears.

Lemma 30 (Change of measure inequality for aligned posteriors) *For any self-complemented set $\mathcal{H}$, for any distribution P on $\mathcal{H}$, any distribution Q aligned on P, and for any measurable function $\phi : \mathcal{H} \to \mathbb{R}$ such that $\phi(f) = \phi(c(f))$ for all $f \in \mathcal{H}$, we have*

$$\mathop{\mathbf{E}}_{f\sim Q} \phi(f) \;\le\; \ln\left(\mathop{\mathbf{E}}_{f\sim P} e^{\phi(f)}\right).$$

Proof First, note that one can change the expectation over Q to an expectation over P, using the fact that $\phi(f) = \phi(c(f))$ for any $f \in \mathcal{H}$, and that Q is aligned on P.

$$\begin{aligned}
2 \cdot \mathop{\mathbf{E}}_{f\sim Q} \phi(f) \;&=\; \int_{\mathcal{H}} df\; Q(f)\,\phi(f) + \int_{\mathcal{H}} df\; Q(c(f))\,\phi(c(f))\\
&=\; \int_{\mathcal{H}} df\; Q(f)\,\phi(f) + \int_{\mathcal{H}} df\; Q(c(f))\,\phi(f)\\
&=\; \int_{\mathcal{H}} df\; \Big(Q(f) + Q(c(f))\Big)\,\phi(f)\\
&=\; \int_{\mathcal{H}} df\; \Big(P(f) + P(c(f))\Big)\,\phi(f)\\
&=\; \int_{\mathcal{H}} df\; P(f)\,\phi(f) + \int_{\mathcal{H}} df\; P(c(f))\,\phi(f)\\
&=\; \int_{\mathcal{H}} df\; P(f)\,\phi(f) + \int_{\mathcal{H}} df\; P(c(f))\,\phi(c(f))\\
&=\; 2 \cdot \mathop{\mathbf{E}}_{f\sim P} \phi(f)\,.
\end{aligned}$$

The result is obtained by changing the expectation over Q to an expectation over P, and then by applying Jensen's inequality (Lemma 47, in Appendix A).

$$\mathop{\mathbf{E}}_{f\sim Q} \phi(f) \;\;=\;\; \mathop{\mathbf{E}}_{f\sim P} \phi(f) \;=\; \mathop{\mathbf{E}}_{f\sim P} \ln e^{\phi(f)} \;\le\; \ln\left(\mathop{\mathbf{E}}_{f\sim P} e^{\phi(f)}\right).$$

■

Theorem 31 (PAC-Bayesian theorem for aligned posteriors) *For any distribution D on $\mathcal{X}\times\{-1,1\}$, any self-complemented set $\mathcal{H}$ of voters $\mathcal{X} \to [-1,1]$, any prior distribution P on $\mathcal{H}$, any convex function $\mathcal{D} : [0,1] \times [0,1] \to \mathbb{R}$ for which $\mathcal{D}(q,p) = \mathcal{D}(1-q,1-p)$, for any $m' > 0$ and any $\delta \in (0,1]$, we have*

$$\Pr_{S\sim D^m}\left(\begin{array}{l}\textit{For all posteriors } Q \textit{ aligned on } P:\\ \mathcal{D}\big(R_S(G_Q), R_D(G_Q)\big) \;\le\; \dfrac{1}{m'}\left[\ln\left(\dfrac{1}{\delta}\mathop{\mathbf{E}}_{S\sim D^m}\mathop{\mathbf{E}}_{f\sim P} e^{m'\cdot\mathcal{D}\left(\mathbb{E}_S^{\mathcal{L}_\ell}(f),\, \mathbb{E}_D^{\mathcal{L}_\ell}(f)\right)}\right)\right]\end{array}\right) \ge 1-\delta\,.$$

Similarly to Theorem 18, the statement of Theorem 31 above contains a value m' which is likely to be set to m in most cases. However, the distinction between m and m' is mandatory to develop the PAC-Bayesian theory for sample-compressed voters in Section 7. Indeed, in proofs of forthcoming Theorems 39, 41 and 42, we have $m' = m - \lambda$, where λ is the size of the voters compression sequence (this concept is properly defined in Section 7).

Proof The proof follows the exact same steps as the proof of Theorem 18, using the linear loss $\mathcal{L} = \mathcal{L}_\ell$ and replacing the use of the change of measure inequality (Lemma 17) by the change of measure inequality for aligned posteriors (Lemma 30), with $\phi(f) = m' \cdot \mathcal{D}\left(\mathbb{E}_S^{\mathcal{L}_\ell}(f),\, \mathbb{E}_D^{\mathcal{L}_\ell}(f)\right)$. Note that this function has the required property, as

$$\mathcal{D}\left(\mathbb{E}_S^{\mathcal{L}_\ell}(f),\, \mathbb{E}_D^{\mathcal{L}_\ell}(f)\right) = \mathcal{D}\left(1 - \mathbb{E}_S^{\mathcal{L}_\ell}(c(f)),\, 1 - \mathbb{E}_D^{\mathcal{L}_\ell}(c(f))\right) = \mathcal{D}\left(\mathbb{E}_S^{\mathcal{L}_\ell}(c(f)),\, \mathbb{E}_D^{\mathcal{L}_\ell}(c(f))\right) .$$

The other steps of the proof stay exactly the same as the proof of Theorem 18. ■

Appendix B presents more general versions of the last two results.

Let us specialize Theorem 31 to the case where $\mathcal{D}(q,p) = \mathrm{kl}(q\|p)$. Doing so, we recover the classical PAC-Bayesian theorem (Theorem 20), but for aligned posteriors, which therefore has no KL term.

Corollary 32 *For any distribution D on $\mathcal{X}\times\{-1,1\}$, any prior distribution P on a self-complemented set $\mathcal{H}$ of voters $\mathcal{X} \to [-1,1]$, and any $\delta \in (0,1]$, we have*

$$\Pr_{S\sim D^m}\left(\begin{array}{l}\textit{For all posteriors } Q \textit{ aligned on } P:\\ \mathrm{kl}\big(R_S(G_Q)\,\big\|\,R_D(G_Q)\big) \;\le\; \dfrac{1}{m}\left[\ln\dfrac{\xi(m)}{\delta}\right]\end{array}\right) \ge 1-\delta\,,$$

where $\mathrm{kl}(q\|p)$ *and* $\xi(m)$ *and defined by Equations* (21) *and* (22) *respectively.*

Proof This result follows from Theorem 31 by choosing $\mathcal{D}(q,p) = \mathrm{kl}(q,p)$ and $m' = m$. The rest of the proof relies on Lemma 19 (as for the proof of Theorem 20). ■

The following corollary is very similar to the original PAC-Bayesian bound of McAllester (2003a), but without the KL term.

Corollary 33 *For any distribution D on $\mathcal{X}\times\{-1,1\}$, any self-complemented set $\mathcal{H}$ of voters $\mathcal{X} \to [-1,1]$, any prior distribution P on $\mathcal{H}$, and any $\delta \in (0,1]$, we have*

$$\Pr_{S\sim D^m}\left(\begin{array}{l}\textit{For all posteriors } Q \textit{ aligned on } P:\\ R_D(G_Q) \;\le\; R_S(G_Q) + \sqrt{\dfrac{1}{2m}\left[\ln\frac{\xi(m)}{\delta}\right]}\end{array}\right) \ge 1-\delta\,.$$

Proof The result is derived from Corollary 32, by using $2(q-p)^2 \le \mathrm{kl}(q\|p)$ (Pinsker's inequality), and isolating $R_D(G_Q)$ in the obtained inequality. ■

Unlike Theorem 18, Theorem 31 cannot straightforwardly be used for pairs of voters, as we did in the proof of Theorem 25. The reason is that a posterior distribution that is the result of the product of two aligned posteriors is not necessarily aligned itself. So, we have to ensure that we can get rid of the KL term even in that case.

6.3 PAC-Bayesian Theorems without KL for the Expected Disagreement d_Q^D

The following theorem is similar to Theorem 31 for aligned posteriors, but deals with paired-voters. Instead of the linear loss $\mathcal{L}_\ell$, we use the loss $\mathcal{L}_d$ of Equation (27), which is a linear loss defined on a pair of voters. Again, the next two results can be seen as a particular case of the two theorems from Appendix B.

In this subsection, we use the following shorthand notation. Given $f_{ij} = \langle f_i, f_j\rangle$ as defined in Definition 24, the voters f_{i^cj}, f_{ij^c} and $f_{i^cj^c}$ are defined as

$$f_{i^cj}(x) \overset{\text{def}}{=} \langle c(f_i)(x), f_j(x)\rangle,\ f_{ij^c}(x) \overset{\text{def}}{=} \langle f_i(x), c(f_j)(x)\rangle,\ \text{and}\ f_{i^cj^c}(x) \overset{\text{def}}{=} \langle c(f_i)(x), c(f_j)(x)\rangle.$$

Recall that from Equation (26), we have $\mathcal{H}^2 \overset{\text{def}}{=} \{f_{ij} : f_i, f_j \in \mathcal{H}\}$ and $Q^2(f_{ij}) \overset{\text{def}}{=} Q(f_i)\cdot Q(f_j)$. Similarly, we define $P^2(f_{ij}) \overset{\text{def}}{=} P(f_i)\cdot P(f_j)$. Using this notation, let us first generalize the change of measure inequality of Lemma 30 to paired-voters.

Lemma 34 (Change of measure inequality for paired-voters and aligned posteriors) *For any self-complemented set $\mathcal{H}$, for any distribution P on $\mathcal{H}$, any distribution Q aligned on P, and for any measurable function $\phi : \mathcal{H}^2 \to \mathbb{R}$ such that $\phi(f_{ij}) = \phi(f_{i^cj}) = \phi(f_{ij^c}) = \phi(f_{i^cj^c})$ for all $f_{ij} \in \mathcal{H}^2$, we have*

$$\mathop{\mathbf{E}}_{f_{ij}\sim Q^2} \phi(f_{ij}) \ \le\ \ln\left(\mathop{\mathbf{E}}_{f_{ij}\sim P^2} e^{\phi(f_{ij})}\right).$$

Proof First, note that one can change the expectation over Q^2 to an expectation over P^2, using the fact that $\phi(f_{ij}) = \phi(f_{i^cj}) = \phi(f_{ij^c}) = \phi(f_{i^cj^c})$ for any $f_{ij} \in \mathcal{H}^2$, and that Q is aligned on P. More specifically, we have the following.

$$\begin{aligned}
4\cdot &\mathop{\mathbf{E}}_{f_{ij}\sim Q^2} \phi(f_{ij})\\
&= \int_{\mathcal{H}^2} df_{ij} Q^2(f_{ij})\,\phi(f_{ij}) + \int_{\mathcal{H}^2} df_{ij} Q^2(f_{i^cj})\,\phi(f_{i^cj}) + \int_{\mathcal{H}^2} df_{ij} Q^2(f_{ij^c})\,\phi(f_{ij^c}) + \int_{\mathcal{H}^2} df_{ij} Q^2(f_{i^cj^c})\,\phi(f_{i^cj^c})\\
&= \int_{\mathcal{H}^2} df_{ij}\, Q^2(f_{ij})\,\phi(f_{ij}) + \int_{\mathcal{H}^2} df_{ij}\, Q^2(f_{i^cj})\,\phi(f_{ij}) + \int_{\mathcal{H}^2} df_{ij}\, Q^2(f_{ij^c})\,\phi(f_{ij}) + \int_{\mathcal{H}^2} df_{ij}\, Q^2(f_{i^cj^c})\,\phi(f_{ij})\\
&= \int_{\mathcal{H}^2} df_{ij} \Big(Q^2(f_{ij}) + Q^2(f_{i^cj}) + Q^2(f_{ij^c}) + Q^2(f_{i^cj^c})\Big)\phi(f_{ij})\\
&= \int_{\mathcal{H}^2} df_{ij} \Big(P^2(f_{ij}) + P^2(f_{i^cj}) + P^2(f_{ij^c}) + P^2(f_{i^cj^c})\Big)\phi(f_{ij})\\
&\ \ \vdots\\
&= 4\cdot \mathop{\mathbf{E}}_{f_{ij}\sim P^2} \phi(f_{ij})\,.
\end{aligned}$$

The result is then obtained by changing the expectation over Q^2 to an expectation over P^2, and then by applying Jensen's inequality (Lemma 47, in Appendix A).

$$\mathop{\mathbf{E}}_{f_{ij}\sim Q^2} \phi(f_{ij}) \;=\; \mathop{\mathbf{E}}_{f_{ij}\sim P^2} \phi(f_{ij}) \;=\; \mathop{\mathbf{E}}_{f_{ij}\sim P^2} \ln e^{\phi(f_{ij})} \;\leq\; \ln\left(\mathop{\mathbf{E}}_{f_{ij}\sim P^2} e^{\phi(f_{ij})}\right).$$

■

Theorem 35 (PAC-Bayesian theorem for paired-voters and aligned posteriors) *For any distribution D on $\mathcal{X}\times\{-1,1\}$, any self-complemented set $\mathcal{H}$ of voters $\mathcal{X}\to[-1,1]$, any prior distribution P on $\mathcal{H}$, any convex function $\mathcal{D} : [0,1]\times[0,1]\to\mathbb{R}$ for which $\mathcal{D}(q,p)=\mathcal{D}(1-q,1-p)$, for any $m'>0$ and any $\delta\in(0,1]$, we have*

$$\Pr_{S\sim D^m}\left(\begin{array}{l}\textit{For all posteriors } Q \textit{ aligned on } P:\\ \mathcal{D}\big(\,d_Q^S,\, d_Q^D\,\big) \;\leq\; \dfrac{1}{m'}\left[\ln\left(\dfrac{1}{\delta}\mathop{\mathbf{E}}_{S\sim D^m}\mathop{\mathbf{E}}_{f_{ij}\sim P^2} e^{m'\cdot\mathcal{D}(\mathbb{E}_S^{\mathcal{L}_d}(f_{ij}),\mathbb{E}_D^{\mathcal{L}_d}(f_{ij}))}\right)\right]\end{array}\right) \geq 1-\delta\,,$$

where f_{ij} is given in Definition 24, and where $P^2(f_{ij}) \overset{\text{def}}{=} P(f_i)\cdot P(f_j)$.

Proof Theorem 35 is deduced from Theorem 31, by using the change of measure inequality given by Lemma 34 instead of the one from Lemma 30, with $\phi(f_{ij}) = m'\cdot\mathcal{D}(\mathbb{E}_S^{\mathcal{L}_d}(f_{ij}),\,\mathbb{E}_D^{\mathcal{L}_d}(f_{ij}))$. As the loss $\mathcal{L}_d$ is such that

$$\mathbb{E}_{D'}^{\mathcal{L}_d}(f_{i^cj^c}) \;=\; \mathbb{E}_{D'}^{\mathcal{L}_d}(f_{ij})\,, \quad\text{ and }\quad \mathbb{E}_{D'}^{\mathcal{L}_d}(f_{i^cj}) \;=\; \mathbb{E}_{D'}^{\mathcal{L}_d}(f_{ij^c}) \;=\; 1-\mathbb{E}_{D'}^{\mathcal{L}_d}(f_{ij})\,,$$

we then have that $\phi(f_{ij})$ has the required property to apply Lemma 34. ■

Let us now specialize Theorem 35 to $\mathcal{D}(q,p)=\mathrm{kl}(q\|p)$.

Corollary 36 *For any distribution D on $\mathcal{X}\times\{-1,1\}$, any self-complemented set $\mathcal{H}$ of voters $\mathcal{X}\to[-1,1]$, any prior distribution P on $\mathcal{H}$, and any $\delta\in(0,1]$, we have*

$$\Pr_{S\sim D^m}\left(\begin{array}{l}\textit{For all posteriors } Q \textit{ aligned on } P:\\ \mathrm{kl}\big(d_Q^S\,\|\,d_Q^D\big) \;\leq\; \dfrac{1}{m}\left[\ln\frac{\xi(m)}{\delta}\right]\end{array}\right) \geq 1-\delta\,.$$

Proof The result is directly obtained from Theorem 35, by choosing $\mathcal{D}(q,p)=\mathrm{kl}(q,p)$. The rest of the proof relies on Lemma 19. ■

Similarly as for Corollary 33, we can easily derive the following result.

Corollary 37 *For any distribution D on $\mathcal{X}\times\{-1,1\}$, for any self-complemented set $\mathcal{H}$ of voters $\mathcal{X}\to[-1,1]$, any prior distribution P on $\mathcal{H}$, and any $\delta\in(0,1]$, we have*

$$\Pr_{S\sim D^m}\left(\begin{array}{l}\textit{For all posteriors } Q \textit{ aligned on } P:\\ d_Q^D \;\geq\; d_Q^S - \sqrt{\dfrac{1}{2m}\left[\ln\frac{\xi(m)}{\delta}\right]}\end{array}\right) \geq 1-\delta\,.$$

Proof The result is derived from Corollary 36, by using $2(q-p)^2 \leq \mathrm{kl}(q\|p)$ (Pinsker's inequality), and isolating d_Q^D in the obtained inequality. ■

6.4 A Bound for the Risk of the Majority Vote without KL Term

Finally, we make use of these results to bound $\mathcal{C}_Q^D$ – and therefore $R_D(B_Q)$ – for aligned posteriors Q, giving rise to PAC-Bound 3. Aside from the fact that this bound has no KL term, it is similar to PAC-Bound 1, as it separately bounds the Gibbs risk and the expected disagreement. This new PAC-Bayesian bound provides us with a starting point to design the MinCq leaning algorithm introduced in Section 8.

PAC-Bound 3 *For any distribution D on $\mathcal{X}\times\{-1,1\}$, for any self-complemented set $\mathcal{H}$ of voters $\mathcal{X}\to[-1,1]$, for any prior distribution P on $\mathcal{H}$, and any $\delta\in(0,1]$, we have*

$$\Pr_{S\sim D^m}\left(\begin{array}{l}\forall\, Q \textit{ aligned on } P:\\ R_D(B_Q) \;\leq\; 1-\dfrac{\big(1-2\cdot\overline{r}\,\big)^2}{1-2\cdot\underline{d}} \;=\; 1-\dfrac{\big(\,\underline{\mu_1}\,\big)^2}{\overline{\mu_2}}\end{array}\right) \geq 1-\delta\,,$$

where

$$\overline{r} \overset{\text{def}}{=} \min\left(\tfrac{1}{2},\, R_S(G_Q)+\sqrt{\tfrac{1}{2m}\left[\ln\tfrac{\xi(m)}{\delta/2}\right]}\right), \qquad \underline{d} \overset{\text{def}}{=} \max\left(0,\, d_Q^S-\sqrt{\tfrac{1}{2m}\left[\ln\tfrac{\xi(m)}{\delta/2}\right]}\right),$$

$$\underline{\mu_1} \overset{\text{def}}{=} \max\left(0,\, \mu_1(M_Q^S)-\sqrt{\tfrac{2}{m}\left[\ln\tfrac{\xi(m)}{\delta/2}\right]}\right), \qquad \overline{\mu_2} \overset{\text{def}}{=} \min\left(1,\, \mu_2(M_Q^S)+\sqrt{\tfrac{2}{m}\left[\ln\tfrac{\xi(m)}{\delta/2}\right]}\right).$$

Proof The inequality is a consequence of Theorem 11, as well as Corollaries 33 and 37. The equality $1-\frac{(1-2\cdot\overline{r})^2}{1-2\cdot\underline{d}} = 1-\frac{(\underline{\mu_1})^2}{\overline{\mu_2}}$ is a direct application of Equations (7) and (9). ■

PAC-Bound 3' that is presented at the end of Section 7 accepts voters that are kernel functions defined using a part of the training set S. This is unusual in the PAC-Bayesian theory, since the prior P on the set of voters has to be defined before seeing the training set S. To overcome this difficulty, we use the sample compression theory.

7. PAC-Bayesian Theory for Sample-Compressed Voters

PAC-Bayesian theorems of Sections 5 and 6 are not valid when $\mathcal{H}$ consists of a set of functions of the form $\pm k(x_i,\cdot)$ for some kernel $k:\mathcal{X}\times\mathcal{X}\to[-1,1]$, as is the case with the Support Vector Machine classifier (see Equation 1). This is because the definition of each involved voter depends on an example (x_i,y_i) of the training data S. This is problematic from the PAC-Bayesian point of view because the prior on the voters is supposed to be defined before seeing the data S. There are two known methods to overcome this problem.

The first method, introduced by Langford and Shawe-Taylor (2002), considers a surrogate set of voters $\mathcal{H}^k$ of *all* the linear classifiers in the space induced[12] by the kernel k. They

[12] This space is also known as a Reproducible Kernel Hilbert Space (RKHS). For more details, see Cristianini and Shawe-Taylor (2000) and Schölkopf et al. (2001)

then make use of the representer theorem to show that the classification function turns out to be a linear combination of the examples, similar to the Support Vector Machine classifier (Equation 1). To avoid the curse of dimensionality, they propose restricting the choice of the prior and posterior distributions on $\mathcal{H}^k$ to isotropic Gaussian centered on a vector representing a particular linear classifier. Based on this approach, Germain et al. (2009) suggests a learning algorithm for linear classifiers that exactly consists in a PAC-Bayesian bound minimization.

The second method, that is presented in the present section, is based on the sample compression setting of Floyd and Warmuth (1995). It has been adapted to the PAC-Bayesian theory by Laviolette and Marchand (2005, 2007), allowing one to directly deal with the case where voters are constructed using examples in the training set, without involving any RKHS notion nor any representer theorem. Conversely to the first method described above, the sample compression approach allows one not only to deal with kernel functions, but with any kind of similarity measure between examples, hence to deal with any kind of voters.

7.1 The General Sample Compression Setting

In the *sample compression setting*, learning algorithms have access to a data-dependent set of voters, that we refer to as *sc-voters*. Given a training sequence[13] $S = \langle\, (x_1, y_1), \ldots, (x_m, y_m)\, \rangle$, each sc-voter is described by a sequence $S_\mathbf{i}$ of elements of S called the *compression sequence*, and a *message* σ which represents the additional information needed to obtain a voter from $S_\mathbf{i}$. If $\mathbf{i} = \langle i_1, i_2, .., i_k\rangle$, then $S_\mathbf{i} \stackrel{\text{def}}{=} \langle (x_{i_1}, y_{i_1}),\, (x_{i_2}, y_{i_2}), \ldots, (x_{i_k}, y_{i_k})\rangle$. In this paper, repetitions are allowed in $S_\mathbf{i}$, and k, the number of indices present in $\mathbf{i}$ (counting the repetitions), is denoted by $|\mathbf{i}|$.

The fact that each sc-voter is described by a compression sequence and a message implies that there exists a *reconstruction function* $\mathcal{R}(S_\mathbf{i}, \sigma)$ that outputs a classifier when given an arbitrary compression sequence $S_\mathbf{i}$ and a message σ. The message σ is chosen from the set $\Sigma_{S_\mathbf{i}}$ of all messages that can be supplied with the compression sequence $S_\mathbf{i}$. In the PAC-Bayesian setting, $\Sigma_{S_\mathbf{i}}$ must be defined a priori (before observing the training data) for all possible sequences $S_\mathbf{i}$, and can be either a discrete or a continuous set. The sample compression setting strictly generalizes the (classical) non-sample-compressed setting, since the latter corresponds to the case where $|\mathbf{i}| = 0$, the voters being then defined only via the messages.

7.2 A Simplified Sample Compression Setting

For the needs of this paper, we consider a simplified framework where sc-voters have a compression sequence of at most λ examples (possibly with repetitions) and a message string of λ bits that we represent by a sequence of "-1" and "$+1$". Instead of being defined on sc-voters, the weighted distribution Q is defined on $\mathcal{I}_\lambda \times \Sigma_\lambda$, where

$$\mathcal{I}_\lambda \stackrel{\text{def}}{=} \Big\{ \langle i_1, i_2, .., i_k\rangle \,:\, k \in \{0, .., \lambda\} \text{ and } i_j \in \{1, .., m\} \Big\} \quad \text{and} \quad \Sigma_\lambda \stackrel{\text{def}}{=} \Big\{ -1, 1 \Big\}^\lambda. \tag{38}$$

[13] The sample compression theory considers the training examples as a sequence instead of a set, because it refers to the training examples by their indices.

In other words, $Q(\mathbf{i},\boldsymbol{\sigma})$ corresponds to the weight of the sc-voter output by $\mathcal{R}(S_{\mathbf{i}},\boldsymbol{\sigma})$, *i.e.*, the sc-voter of compression sequence $\mathbf{i}=\langle i_1,\ldots,i_{|\mathbf{i}|}\rangle\in\mathcal{I}_\lambda$ and message $\boldsymbol{\sigma}=\langle\sigma_1,\ldots,\sigma_\lambda\rangle\in\Sigma_\lambda$. In particular, a prior (resp., a posterior) on the set of all sc-voters is now simply a prior on the set $\mathcal{I}_\lambda\times\Sigma_\lambda$. Thus, such a prior can really be defined *a priori*, before seeing the data S.[14] The set of sc-voters is therefore only defined when the training sequence S is given, and corresponds to

$$\mathcal{H}^{\mathcal{R}}_{S,\lambda} \overset{\text{def}}{=} \{\mathcal{R}(S_{\mathbf{i}},\boldsymbol{\sigma}) \,:\, \mathbf{i}\in\mathcal{I}_\lambda,\, \boldsymbol{\sigma}\in\Sigma_\lambda\}\,.$$

Finally, given a training sequence S and a reconstruction function $\mathcal{R}$, for a distribution Q on $\mathcal{I}_\lambda\times\Sigma_\lambda$, we define the Bayes classifier as

$$B_{Q,S} \overset{\text{def}}{=} \operatorname{sgn}\left[\mathop{\mathbf{E}}_{(\mathbf{i},\boldsymbol{\sigma})\sim Q} \mathcal{R}(S_{\mathbf{i}},\boldsymbol{\sigma})\right].$$

We then define the Bayes risk $R_{D'}(B_{Q,S})$ and the Gibbs risk $R_{D'}(G_{Q,S})$ of a distribution Q on $\mathcal{I}_\lambda\times\Sigma_\lambda$ relative to D' as

$$\begin{aligned} R_{D'}(B_{Q,S}) &\overset{\text{def}}{=} \mathbb{E}^{\mathcal{L}_{01}}_{D'}\big(B_{Q,S}\big)\,,\\ R_{D'}(G_{Q,S}) &\overset{\text{def}}{=} \mathop{\mathbf{E}}_{(\mathbf{i},\boldsymbol{\sigma})\sim Q} \mathbb{E}^{\mathcal{L}_\ell}_{D'}\big(\mathcal{R}(S_{\mathbf{i}},\boldsymbol{\sigma})\big)\,. \end{aligned}$$

7.3 A First Sample-Compressed PAC-Bayesian Theorem

To derive PAC-Bayesian bounds for majority votes of sc-voters, one must deal with the following issue: even if the training sequence S is drawn i.i.d. from a data-generating distribution D, the empirical risk of the Gibbs $R_S(G_{Q,S})$ is not an unbiased estimate of its true risk $R_D(G_{Q,S})$. For instance, the reconstruction function $\mathcal{R}$ can be such that an sc-voter output by $\mathcal{R}(S_{\mathbf{i}},\boldsymbol{\sigma})$ never errs on an example belonging to its compression sequence $S_{\mathbf{i}}$; this biases the empirical risk because examples of $S_{\mathbf{i}}$ are all in S.

To deal with this bias, the $\frac{1}{m}$ factor in the usual PAC-Bayesian bounds is replaced by a factor of the form $\frac{1}{m-l}$ in their sample compression versions. In Laviolette and Marchand (2005, 2007), l corresponds to the Q-average size of the sample compression sequence. In the present paper, we restrain ourselves to a simpler case, where l is the maximum possible size of a compression sequence (*i.e.*, $l=\lambda$). This simplification allows us to deal with the biased character of the empirical Gibbs risk using a proof approach similar to the one proposed in Germain et al. (2011). The key step of this approach is summarized in the following lemma.

Lemma 38 *Let $\mathcal{R}$ be a reconstruction function that outputs sc-voters of size at most λ (where $\lambda<m$). For any distribution D on $\mathcal{X}\times\{-1,1\}$, and for any prior distribution P on $\mathcal{I}_\lambda\times\Sigma_\lambda$,*

$$\mathop{\mathbf{E}}_{S\sim D^m}\ \mathop{\mathbf{E}}_{(\mathbf{i},\boldsymbol{\sigma})\sim P} e^{(m-\lambda)\cdot 2\cdot\left(\mathbb{E}^{\mathcal{L}_\ell}_S(\mathcal{R}(S_{\mathbf{i}},\boldsymbol{\sigma}))-\mathbb{E}^{\mathcal{L}_\ell}_D(\mathcal{R}(S_{\mathbf{i}},\boldsymbol{\sigma}))\right)^2} \;\le\; e^{4\lambda}\cdot\xi(m-\lambda)\,,$$

where $\xi(\cdot)$ is defined by Equation (22), and therefore we have that $\xi(m-\lambda)\le 2\sqrt{m-\lambda}$.

[14] Laviolette and Marchand (2007) describe a more general setting where, for each $S\in(\mathcal{X}\times\mathcal{Y})^m$, a prior is defined on $\mathcal{I}_\lambda\times\Sigma_{S_{\mathbf{i}}}$. Hence, the messages may depend on the compression sequence $S_{\mathbf{i}}$.

Proof As the the choice of $(\mathbf{i},\boldsymbol{\sigma})$ according to the prior P is independent[15] of S, we have

$$
\begin{aligned}
&\mathop{\mathbf{E}}_{S\sim D^m}\ \mathop{\mathbf{E}}_{(\mathbf{i},\boldsymbol{\sigma})\sim P} e^{(m-\lambda)\cdot 2\cdot\left(\mathbb{E}_S^{\mathcal{L}_\ell}(\mathcal{R}(S_{\mathbf{i}},\boldsymbol{\sigma}))-\mathbb{E}_D^{\mathcal{L}_\ell}(\mathcal{R}(S_{\mathbf{i}},\boldsymbol{\sigma}))\right)^2}\\
&\qquad = \mathop{\mathbf{E}}_{(\mathbf{i},\boldsymbol{\sigma})\sim P}\ \mathop{\mathbf{E}}_{S\sim D^m} e^{(m-\lambda)\cdot 2\cdot\left(\mathbb{E}_S^{\mathcal{L}_\ell}(\mathcal{R}(S_{\mathbf{i}},\boldsymbol{\sigma}))-\mathbb{E}_D^{\mathcal{L}_\ell}(\mathcal{R}(S_{\mathbf{i}},\boldsymbol{\sigma}))\right)^2} \qquad (39)\\
&\qquad = \mathop{\mathbf{E}}_{(\mathbf{i},\boldsymbol{\sigma})\sim P}\ \mathop{\mathbf{E}}_{S_{\mathbf{i}}\sim D^\lambda}\ \mathop{\mathbf{E}}_{S_{\mathbf{i}^c}\sim D^{m-\lambda}} e^{(m-\lambda)\cdot 2\cdot\left(\mathbb{E}_S^{\mathcal{L}_\ell}(\mathcal{R}(S_{\mathbf{i}},\boldsymbol{\sigma}))-\mathbb{E}_D^{\mathcal{L}_\ell}(\mathcal{R}(S_{\mathbf{i}},\boldsymbol{\sigma}))\right)^2}. \qquad (40)
\end{aligned}
$$

Let us now rewrite the empirical loss of an sc-voter as a combination of the loss on its compression sequence $S_{\mathbf{i}}$ and the loss on the other training examples $S_{\mathbf{i}^c}$.

$$
\mathbb{E}_S^{\mathcal{L}_\ell}(\mathcal{R}(S_{\mathbf{i}},\boldsymbol{\sigma})) \;=\; \frac{1}{m}\left[\lambda\cdot\mathbb{E}_{S_{\mathbf{i}}}^{\mathcal{L}_\ell}(\mathcal{R}(S_{\mathbf{i}},\boldsymbol{\sigma}))+(m-\lambda)\cdot\mathbb{E}_{S_{\mathbf{i}^c}}^{\mathcal{L}_\ell}(\mathcal{R}(S_{\mathbf{i}},\boldsymbol{\sigma}))\right].
$$

Since $0\leq\mathbb{E}_{D'}^{\mathcal{L}_\ell}(\mathcal{R}(S_{\mathbf{i}},\boldsymbol{\sigma}))\leq 1$ and $2\cdot(q-p)^2\leq\mathrm{kl}(q\|p)$ (Pinsker's inequality), we have

$$
\begin{aligned}
&(m-\lambda)\cdot 2\cdot\left(\mathbb{E}_S^{\mathcal{L}_\ell}(\mathcal{R}(S_{\mathbf{i}},\boldsymbol{\sigma}))-\mathbb{E}_D^{\mathcal{L}_\ell}(\mathcal{R}(S_{\mathbf{i}},\boldsymbol{\sigma}))\right)^2\\
&\quad= (m-\lambda)\cdot 2\cdot\left(\tfrac{1}{m}\left[\lambda\cdot\mathbb{E}_{S_{\mathbf{i}}}^{\mathcal{L}_\ell}(\mathcal{R}(S_{\mathbf{i}},\boldsymbol{\sigma}))+(m-\lambda)\cdot\mathbb{E}_{S_{\mathbf{i}^c}}^{\mathcal{L}_\ell}(\mathcal{R}(S_{\mathbf{i}},\boldsymbol{\sigma}))\right]-\mathbb{E}_D^{\mathcal{L}_\ell}(\mathcal{R}(S_{\mathbf{i}},\boldsymbol{\sigma}))\right)^2\\
&\quad= (m-\lambda)\cdot 2\cdot\left(\tfrac{\lambda}{m}\left[\mathbb{E}_{S_{\mathbf{i}}}^{\mathcal{L}_\ell}(\mathcal{R}(S_{\mathbf{i}},\boldsymbol{\sigma}))-\mathbb{E}_{S_{\mathbf{i}^c}}^{\mathcal{L}_\ell}(\mathcal{R}(S_{\mathbf{i}},\boldsymbol{\sigma}))\right]+\left[\mathbb{E}_{S_{\mathbf{i}^c}}^{\mathcal{L}_\ell}(\mathcal{R}(S_{\mathbf{i}},\boldsymbol{\sigma}))-\mathbb{E}_D^{\mathcal{L}_\ell}(\mathcal{R}(S_{\mathbf{i}},\boldsymbol{\sigma}))\right]\right)^2\\
&\quad= (m-\lambda)\cdot 2\cdot\Big(\left(\tfrac{\lambda}{m}\right)^2\left[\mathbb{E}_{S_{\mathbf{i}}}^{\mathcal{L}_\ell}(\mathcal{R}(S_{\mathbf{i}},\boldsymbol{\sigma}))-\mathbb{E}_{S_{\mathbf{i}^c}}^{\mathcal{L}_\ell}(\mathcal{R}(S_{\mathbf{i}},\boldsymbol{\sigma}))\right]^2+\left[\mathbb{E}_{S_{\mathbf{i}^c}}^{\mathcal{L}_\ell}(\mathcal{R}(S_{\mathbf{i}},\boldsymbol{\sigma}))-\mathbb{E}_D^{\mathcal{L}_\ell}(\mathcal{R}(S_{\mathbf{i}},\boldsymbol{\sigma}))\right]^2\\
&\qquad\qquad+\tfrac{2\lambda}{m}\left[\mathbb{E}_{S_{\mathbf{i}}}^{\mathcal{L}_\ell}(\mathcal{R}(S_{\mathbf{i}},\boldsymbol{\sigma}))-\mathbb{E}_{S_{\mathbf{i}^c}}^{\mathcal{L}_\ell}(\mathcal{R}(S_{\mathbf{i}},\boldsymbol{\sigma}))\right]\left[\mathbb{E}_{S_{\mathbf{i}^c}}^{\mathcal{L}_\ell}(\mathcal{R}(S_{\mathbf{i}},\boldsymbol{\sigma}))-\mathbb{E}_D^{\mathcal{L}_\ell}(\mathcal{R}(S_{\mathbf{i}},\boldsymbol{\sigma}))\right]\Big)\\
&\quad\leq (m-\lambda)\cdot 2\cdot\left(\left(\tfrac{\lambda}{m}\right)^2+\left[\mathbb{E}_{S_{\mathbf{i}^c}}^{\mathcal{L}_\ell}(\mathcal{R}(S_{\mathbf{i}},\boldsymbol{\sigma}))-\mathbb{E}_D^{\mathcal{L}_\ell}(\mathcal{R}(S_{\mathbf{i}},\boldsymbol{\sigma}))\right]^2+\tfrac{2\lambda}{m}\right)\\
&\quad= 2\,\lambda\cdot\left(2-\tfrac{\lambda}{m}-\left(\tfrac{\lambda}{m}\right)^2\right)+(m-\lambda)\cdot 2\cdot\left[\mathbb{E}_{S_{\mathbf{i}^c}}^{\mathcal{L}_\ell}(\mathcal{R}(S_{\mathbf{i}},\boldsymbol{\sigma}))-\mathbb{E}_D^{\mathcal{L}_\ell}(\mathcal{R}(S_{\mathbf{i}},\boldsymbol{\sigma}))\right]^2\\
&\quad\leq 4\,\lambda+(m-\lambda)\cdot 2\cdot\left[\mathbb{E}_{S_{\mathbf{i}^c}}^{\mathcal{L}_\ell}(\mathcal{R}(S_{\mathbf{i}},\boldsymbol{\sigma}))-\mathbb{E}_D^{\mathcal{L}_\ell}(\mathcal{R}(S_{\mathbf{i}},\boldsymbol{\sigma}))\right]^2\\
&\quad\leq 4\,\lambda+(m-\lambda)\cdot\mathrm{kl}\big(\mathbb{E}_{S_{\mathbf{i}^c}}^{\mathcal{L}_\ell}(\mathcal{R}(S_{\mathbf{i}},\boldsymbol{\sigma}))\,\|\,\mathbb{E}_D^{\mathcal{L}_\ell}(\mathcal{R}(S_{\mathbf{i}},\boldsymbol{\sigma}))\big). \qquad (41)
\end{aligned}
$$

Note that $\mathcal{R}(S_{\mathbf{i}},\boldsymbol{\sigma})$ does not depend on examples contained in $S_{\mathbf{i}^c}$. Thus, from the point of view of $S_{\mathbf{i}^c}$, $\mathcal{R}(S_{\mathbf{i}},\boldsymbol{\sigma})$ is a classical voter (not a sample-compressed one). Therefore, one can apply Lemma 19, replacing $S\sim D^m$ by $S_{\mathbf{i}^c}\sim D^{m-\lambda}$, and f by $\mathcal{R}(S_{\mathbf{i}},\boldsymbol{\sigma})$. Lemma 19, together with Equations (40) and (41), gives

$$
\begin{aligned}
&\mathop{\mathbf{E}}_{(\mathbf{i},\boldsymbol{\sigma})\sim P}\ \mathop{\mathbf{E}}_{S_{\mathbf{i}}\sim D^\lambda}\ \mathop{\mathbf{E}}_{S_{\mathbf{i}^c}\sim D^{m-\lambda}} e^{(m-\lambda)\cdot 2\cdot\left(\mathbb{E}_S^{\mathcal{L}_\ell}(\mathcal{R}(S_{\mathbf{i}},\boldsymbol{\sigma}))-\mathbb{E}_D^{\mathcal{L}_\ell}(\mathcal{R}(S_{\mathbf{i}},\boldsymbol{\sigma}))\right)^2}\\
&\qquad\leq e^{4\,\lambda}\cdot\mathop{\mathbf{E}}_{(\mathbf{i},\boldsymbol{\sigma})\sim P}\ \mathop{\mathbf{E}}_{S_{\mathbf{i}}\sim D^\lambda}\ \mathop{\mathbf{E}}_{S_{\mathbf{i}^c}\sim D^{m-\lambda}} e^{(m-\lambda)\cdot\mathrm{kl}\left(\mathbb{E}_{S_{\mathbf{i}^c}}^{\mathcal{L}_\ell}(\mathcal{R}(S_{\mathbf{i}},\boldsymbol{\sigma}))\,\|\,\mathbb{E}_D^{\mathcal{L}_\ell}(\mathcal{R}(S_{\mathbf{i}},\boldsymbol{\sigma}))\right)}\\
&\qquad\leq e^{4\,\lambda}\cdot\mathop{\mathbf{E}}_{(\mathbf{i},\boldsymbol{\sigma})\sim P}\ \mathop{\mathbf{E}}_{S_{\mathbf{i}}\sim D^\lambda}\xi(m-\lambda) \;=\; e^{4\,\lambda}\cdot\xi(m-\lambda),
\end{aligned}
$$

and we are done. ■

[15] Note that because of this independence, the exchange in the order of the two expectations (Line 39) is trivial. This independence is a direct consequence of our choice to only consider the simplified setting described by Equation (38). In the more general setting of Laviolette and Marchand (2007), this part of the proof is more complicated.

The next PAC-Bayesian theorem presents the generalization of McAllester's PAC-Bayesian bound (Corollary 22) for the sample compression case.

Theorem 39 *Let $\mathcal{R}$ be a reconstruction function that outputs sc-voters of size at most λ (where $\lambda < m$). For any distribution D on $\mathcal{X}\times\{-1,1\}$, for any prior distribution P on $\mathcal{I}_\lambda \times \Sigma_\lambda$, and any $\delta \in (0,1]$, we have*

$$\Pr_{S\sim D^m}\left(\begin{array}{l}\textit{For all posteriors } Q\ : \\ R_D(G_{Q,S}) \ \leq \ R_S(G_{Q,S}) + \sqrt{\dfrac{1}{2(m-\lambda)}\left[\mathrm{KL}(Q\|P) + 4\lambda + \ln\frac{\xi(m-\lambda)}{\delta}\right]}\end{array}\right) \geq\ 1-\delta\,.$$

Proof We apply the exact same steps as in the proof of Theorem 18, with $m' = m - \lambda$, $f = \mathcal{R}(S_{\mathbf{i}}, \boldsymbol{\sigma})$, and $\mathcal{D}(q,p) = 2(q-p)^2$, we obtain

$$\Pr_{S\sim D^m}\left(\begin{array}{l}\text{For all posteriors}\, Q\ : \\ 2\Big(R_S(G_{Q,S}) - R_D(G_{Q,S})\Big)^2 \\ \quad \leq \dfrac{1}{m-\lambda}\left[\mathrm{KL}(Q\|P) + \ln\left(\dfrac{1}{\delta}\mathop{\mathbf{E}}_{S\sim D^m}\ \mathop{\mathbf{E}}_{(\mathbf{i},\boldsymbol{\sigma})\sim P} e^{(m-\lambda)\cdot 2\cdot\left(\mathbb{E}_S^{\mathcal{L}_\ell}(\mathcal{R}(S_{\mathbf{i}},\boldsymbol{\sigma})) - \mathbb{E}_D^{\mathcal{L}_\ell}(\mathcal{R}(S_{\mathbf{i}},\boldsymbol{\sigma}))\right)^2}\right)\right]\end{array}\right) \geq\ 1-\delta\,.$$

The result then follows from Lemma 38 and easy calculations. ■

All the PAC-Bayesian results presented in the preceding sections can be similarly generalized. We leave them to the reader with the exception of the PAC-Bayesian bounds that have no KL, that are used in the next section, as we present the learning algorithm MinCq that minimizes the $\mathcal{C}$-bound.

7.4 Sample-Compressed PAC-Bayesian Bounds without KL

The bounds presented in this section generalize the results presented in Section 6 to the sample compression case. We first need to generalize the notion of self-complement (Definition 29) to sc-voters.

Definition 40 A reconstruction function $\mathcal{R}$ is said to be *self-complemented* if for any training sequence $S \in (\mathcal{X}\times\mathcal{Y})^m$ and any $(\mathbf{i},\boldsymbol{\sigma}) \in \mathcal{I}_\lambda \times \Sigma_\lambda$, we have

$$-\mathcal{R}(S_{\mathbf{i}}, \boldsymbol{\sigma}) \quad = \quad \mathcal{R}(S_{\mathbf{i}}, -\boldsymbol{\sigma})\,,$$

where, if $\boldsymbol{\sigma} = \langle\sigma_1, .., \sigma_\lambda\rangle$, then $-\boldsymbol{\sigma} = \langle -\sigma_1, .., -\sigma_\lambda\rangle$.

7.4.1 A PAC-Bayesian Theorem for the Gibbs Risk of Sc-Voters

Theorem 41 *Let $\mathcal{R}$ be a self-complemented reconstruction function that outputs sc-voters of size at most λ (where $\lambda < m$). For any distribution D on $\mathcal{X}\times\{-1,1\}$, for any prior distribution P on $\mathcal{I}_\lambda \times \Sigma_\lambda$, and any $\delta \in (0,1]$, we have*

$$\Pr_{S\sim D^m}\left(\begin{array}{l}\textit{For all posteriors } Q \textit{ aligned on } P\textit{:} \\ R_D(G_{Q,S}) \ \leq \ R_S(G_{Q,S}) + \sqrt{\dfrac{1}{2(m-\lambda)}\left[4\lambda + \ln\frac{\xi(m-\lambda)}{\delta}\right]}\end{array}\right) \geq\ 1-\delta\,.$$

Proof First note that $2\cdot(q-p)^2 = 2\cdot((1-q)-(1-p))^2$. Then apply the exact same steps as in the proof of Theorem 31 with $m' = m-\lambda$, $f = \mathcal{R}(S_{\mathbf{i}}, \boldsymbol{\sigma})$, and $\mathcal{D}(q,p) = 2(q-p)^2$ to obtain

$$\Pr_{S\sim D^m}\left(\begin{array}{l}\text{For all posteriors}\, Q \text{ aligned on } P:\\ 2\Big(R_S(G_{Q,S})-R_D(G_{Q,S})\Big)^2 \le \frac{1}{m-\lambda}\left[\ln\left(\frac{1}{\delta}\mathop{\mathbf{E}}_{S\sim D^m}\mathop{\mathbf{E}}_{(\mathbf{i},\boldsymbol{\sigma})\sim P} e^{(m-\lambda)\cdot 2\cdot\left(\mathbb{E}_S^{\mathcal{L}_\ell}(\mathcal{R}(S_{\mathbf{i}},\boldsymbol{\sigma}))-\mathbb{E}_D^{\mathcal{L}_\ell}(\mathcal{R}(S_{\mathbf{i}},\boldsymbol{\sigma}))\right)^2}\right)\right]\end{array}\right) \\ \ge 1-\delta\,.$$

The result then follows from Lemma 38 and easy calculations. ■

7.4.2 A PAC-Bayesian Theorem for the Disagreement of Sc-Voters

Given a training sequence S and a reconstruction function $\mathcal{R}$, we define the expected disagreement of a distribution Q on $\mathcal{I}_\lambda \times \Sigma_\lambda$ relative to D' as

$$\begin{aligned} d_{Q,S}^{D'} &\overset{\text{def}}{=} \mathop{\mathbf{E}}_{x\sim D'_{\mathcal{X}}}\ \mathop{\mathbf{E}}_{(\mathbf{i},\boldsymbol{\sigma})\sim Q}\ \mathop{\mathbf{E}}_{(\mathbf{i}',\boldsymbol{\sigma}')\sim Q} \mathcal{L}_\ell\left(\mathcal{R}(S_{\mathbf{i}},\boldsymbol{\sigma})(x), \mathcal{R}(S_{\mathbf{i}'},\boldsymbol{\sigma}')(x)\right) \\ &= \mathop{\mathbf{E}}_{(\mathbf{i},\mathbf{i}',\boldsymbol{\sigma},\boldsymbol{\sigma}')\sim Q^2} \mathbb{E}_{D'}^{\mathcal{L}_d}\Big(\overline{\mathcal{R}}(S_{\mathbf{i},\mathbf{i}'},\boldsymbol{\sigma},\boldsymbol{\sigma}')\Big)\,, \end{aligned}$$

where

$$\begin{aligned} Q^2(\mathbf{i},\mathbf{i}',\boldsymbol{\sigma},\boldsymbol{\sigma}') &\overset{\text{def}}{=} Q(\mathbf{i},\boldsymbol{\sigma})\cdot Q(\mathbf{i}',\boldsymbol{\sigma}')\,, \\ \overline{\mathcal{R}}(S_{\mathbf{i},\mathbf{i}'},\boldsymbol{\sigma},\boldsymbol{\sigma}')(x) &\overset{\text{def}}{=} \left\langle\, \mathcal{R}(S_{\mathbf{i}},\boldsymbol{\sigma})(x), \mathcal{R}(S_{\mathbf{i}'},\boldsymbol{\sigma}')(x)\,\right\rangle\,. \end{aligned}$$

Thus, $\overline{\mathcal{R}}$ is a new reconstruction function that outputs an *sc-paired-voter* which is the sample-compressed version of the paired-voter of Definition 24. From there, we adapt Corollary 37 to sc-voters, and we obtain the following PAC-Bayesian theorem. This result bounds $d_{Q,S}^{D}$ for posterior distributions Q aligned on a prior distribution P.

Theorem 42 *Let $\mathcal{R}$ be a self-complemented reconstruction function that outputs sc-voters of size at most λ (where $\lambda < \lfloor\frac{m}{2}\rfloor$). For any distribution D on $\mathcal{X}\times\{-1,1\}$, for any prior distribution P on $\mathcal{I}_\lambda\times\Sigma_\lambda$, and any $\delta\in(0,1]$, we have*

$$\Pr_{S\sim D^m}\left(\begin{array}{l}\textit{For all posteriors } Q \textit{ aligned on } P:\\ d_{Q,S}^{D} \;\ge\; d_{Q,S}^{S} - \sqrt{\dfrac{1}{2(m-2\,\lambda)}\left[8\lambda+\ln\frac{\xi(m-2\,\lambda)}{\delta}\right]}\end{array}\right) \ge 1-\delta\,.$$

Proof Let $P^2(\mathbf{i},\mathbf{i}',\boldsymbol{\sigma},\boldsymbol{\sigma}') \overset{\text{def}}{=} P(\mathbf{i},\boldsymbol{\sigma})\cdot P(\mathbf{i}',\boldsymbol{\sigma}')$. Now note that $2\cdot(q-p)^2 = 2\cdot((1-q)-(1-p))^2$. Then apply the exact same steps as in the proof of Theorem 35 with $m' = m-2\lambda$, $f_{ij} = \overline{\mathcal{R}}(S_{\mathbf{i},\mathbf{i}'},\boldsymbol{\sigma},\boldsymbol{\sigma}')$ and $\mathcal{D}(q,p) = 2(q-p)^2$ to obtain

$$\Pr_{S\sim D^m}\left(\begin{array}{l}\text{For all posteriors}\, Q \text{ aligned on } P:\\ 2\Big(d_{Q,S}^{S}-d_{Q,S}^{D}\Big)^2 \le \frac{1}{m}\left[\ln\left(\frac{1}{\delta}\mathop{\mathbf{E}}_{S\sim D^m}\mathop{\mathbf{E}}_{(\mathbf{i},\mathbf{i}',\boldsymbol{\sigma},\boldsymbol{\sigma}')\sim P^2} e^{m\cdot 2\cdot\left(\mathbb{E}_S^{\mathcal{L}_d}(\overline{\mathcal{R}}(S_{\mathbf{i},\mathbf{i}'},\boldsymbol{\sigma},\boldsymbol{\sigma}'))-\mathbb{E}_D^{\mathcal{L}_d}(\overline{\mathcal{R}}(S_{\mathbf{i},\mathbf{i}'},\boldsymbol{\sigma},\boldsymbol{\sigma}'))\right)^2}\right)\right]\end{array}\right) \\ \ge\; 1-\delta\,.$$

Calculations similar to the ones of the proof of Lemma 38 (with λ replaced by 2λ) give

$$\mathop{\mathbf{E}}_{S\sim D^m}\ \mathop{\mathbf{E}}_{(\mathbf{i},\mathbf{i}',\boldsymbol{\sigma},\boldsymbol{\sigma}')\sim P^2} e^{(m-2\lambda)\cdot 2\cdot\left(\mathbb{E}_S^{\mathcal{L}_d}(\overline{\mathcal{R}}(S_{\mathbf{i},\mathbf{i}'},\boldsymbol{\sigma},\boldsymbol{\sigma}'))-\mathbb{E}_D^{\mathcal{L}_d}(\overline{\mathcal{R}}(S_{\mathbf{i},\mathbf{i}'},\boldsymbol{\sigma},\boldsymbol{\sigma}'))\right)^2} \quad\le\quad e^{8\,\lambda}\cdot\xi(m-2\lambda)\,.$$

Therefore, we have

$$\Pr_{S\sim D^m}\left(\begin{array}{l}\text{For all posteriors}\, Q \text{ aligned on } P:\\ 2\Big(d^{S}_{Q,S}-d^{D}_{Q,S}\Big)^2\le \dfrac{1}{m-2\lambda}\Big[8\lambda+\ln\frac{\xi(m-2\,\lambda)}{\delta}\Big]\end{array}\right)\ \ge\ 1-\delta\,.$$

and the result is obtained by isolating $d^{D}_{Q,S}$ in the inequality. ■

7.4.3 A Sample Compression Bound for the Risk of the Majority Vote

Let us now exploit Theorems 41 and 42, together with the $\mathcal{C}$-bound of Theorem 11, to obtain a bound on the risk on a majority vote with kernel functions as voters. Given any similarity function (possibly a kernel) $k:\mathcal{X}\times\mathcal{X}\to[-1,1]$ and a training sequence size of m, we consider a majority vote of sc-voters of compression size at most 1 given by the following reconstruction function,

$$\mathcal{R}_k\big(S_{\mathbf{i}},\langle\sigma\rangle\big)(x)\ \stackrel{\text{def}}{=}\ \begin{cases}\sigma & \text{if }\ \mathbf{i}=\langle\,\rangle,\\ \sigma\cdot k(x_i,x) & \text{otherwise (}\ \mathbf{i}=\langle i\rangle\ \text{),}\end{cases}$$

where $\mathbf{i}\in\mathcal{I}_1=\{\langle\,\rangle,\langle 1\rangle,\langle 2\rangle,\ldots,\langle m\rangle\}$ and $\langle\sigma\rangle\in\Sigma_1$ (thus, $\sigma\in\{-1,1\}$). Here, the elements of sets $\mathcal{I}_1$ and Σ_1 are obtained from Equation (38), with $\lambda=1$. Note that $\mathcal{R}_k$ is self-complemented (Definition 40) because $-\mathcal{R}_k\big(S_{\mathbf{i}},\langle\sigma\rangle\big)=\mathcal{R}_k\big(S_{\mathbf{i}},\langle-\sigma\rangle\big)$ for any $(\mathbf{i},\boldsymbol{\sigma})$.

Once the training sequence $S\sim D^m$ is observed, the (self-complemented) reconstruction function $\mathcal{R}_k$ gives rise to the following set of $2m+2$ sc-voters,

$$\mathcal{H}^{\mathcal{R}_k}_{S,1}\ \stackrel{\text{def}}{=}\ \Big\{b(\cdot),k(x_1,\cdot),k(x_2,\cdot),\ldots,k(x_m,\cdot),-b(\cdot),-k(x_1,\cdot),-k(x_2,\cdot),\ldots,-k(x_m,\cdot)\Big\}\,,$$

where $b:\mathcal{X}\to\{1\}$ is a "dummy voter" that always outputs 1 and allows introducing a *bias* value into the majority vote classifier. Note that $\mathcal{H}^{\mathcal{R}_k}_{S,1}$ is a self-complemented set of sc-voters, and the margin of the majority vote given by the distribution Q on $\mathcal{H}^{\mathcal{R}_k}_{S,1}$ is

$$M_{Q,S}(x,y)\ \stackrel{\text{def}}{=}\ y\left(Q\big(\,b(\cdot)\,\big)-Q\big(\ -b(\cdot)\,\big)+\sum_{i=1}^{m}\Big[Q\big(\,k(x_i,\cdot)\,\big)-Q\big(\ -k(x_i,\cdot)\,\big)\Big]\,k(x_i,x)\right)\,.$$

Consequently, the empirical first and second moments of this margin are

$$\mu_1(M^{S}_{Q,S})\ =\ \frac{1}{m}\sum_{i=1}^{m}M_{Q,S}(x_i,y_i),\quad\text{and}\quad \mu_2(M^{S}_{Q,S})\ =\ \frac{1}{m}\sum_{i=1}^{m}\Big[M_{Q,S}(x_i,y_i)\Big]^2.$$

Hence, the empirical Gibbs risk and the empirical expected disagreement can be expressed by

$$R_S(G_{Q,S})\ =\ \frac{1}{2}\left(1-\mu_1(M^{S}_{Q,S})\right),\quad\text{and}\quad d^{S}_{Q,S}\ =\ \frac{1}{2}\left(1-\mu_2(M^{S}_{Q,S})\right)\,. \tag{42}$$

Thus, we obtain the following bound on the risk of a majority vote of kernel voters $R_D(B_{Q,S})$ for aligned posteriors Q.

PAC-Bound 3' *Let $k : \mathcal{X} \times \mathcal{X} \to [-1, 1]$. For any distribution D on $\mathcal{X} \times \{-1, 1\}$, for any prior distribution P on $\mathcal{H}^{\mathcal{R}_k}_{S,1}$, and any $\delta \in (0, 1]$, we have*

$$\Pr_{S\sim D^m}\left(\begin{array}{l} \forall\, Q \text{ aligned on } P : \\ R_D(B_{Q,S}) \;\le\; 1 - \dfrac{\big(1 - 2\cdot \overline{r}\,\big)^2}{1 - 2\cdot \underline{d}} \;=\; 1 - \dfrac{\big(\underline{\mu_1}\big)^2}{\overline{\mu_2}} \end{array} \right) \ge 1 - \delta\,,$$

where

$$\begin{aligned} \overline{r} &\overset{\text{def}}{=} \min\left(\tfrac{1}{2},\, R_S(G_{Q,S}) + \sqrt{\tfrac{1}{2(m-1)}\left[4 + \ln\tfrac{\xi(m-1)}{\delta/2}\right]}\right), \\ \underline{d} &\overset{\text{def}}{=} \max\left(0,\, d^S_{Q,S} - \sqrt{\tfrac{1}{2(m-2)}\left[8 + \ln\tfrac{\xi(m-2)}{\delta/2}\right]}\right), \\ \underline{\mu_1} &\overset{\text{def}}{=} \max\left(0,\, \mu_1(M^S_{Q,S}) - \sqrt{\tfrac{2}{m-1}\left[4 + \ln\tfrac{\xi(m-1)}{\delta/2}\right]}\right), \\ \overline{\mu_2} &\overset{\text{def}}{=} \min\left(1,\, \mu_2(M^S_{Q,S}) + \sqrt{\tfrac{2}{m-2}\left[8 + \ln\tfrac{\xi(m-2)}{\delta/2}\right]}\right). \end{aligned}$$

Proof The proof is almost identical to the one of PAC-Bound 3, except that it relies on sample-compressed PAC-Bayesian bounds. Indeed, the inequality is a consequence of Theorem 11, as well as Theorems 41 and 42. The equality $1 - \frac{(1-2\cdot\overline{r})^2}{1-2\cdot\underline{d}} = 1 - \frac{(\underline{\mu_1})^2}{\overline{\mu_2}}$ is a direct application of Equation (42). ■

PAC-Bounds 3 and 3' are expressed in two forms. The first form relies on bounds on the Gibbs risk and the expected disagreement (denoted $\overline{r}$ and $\underline{d}$). The second form relies on bounds on the first and second moments of the margin (denoted $\underline{\mu_1}$ and $\overline{\mu_2}$). This latter form is used to justify the learning algorithm presented in Section 8.

8. MinCq: Learning by Minimizing the $\mathcal{C}$-bound

In this section, we propose a new algorithm, that we call MinCq, for constructing a weighted majority vote of voters. One version of this algorithm is designed for the supervised inductive framework and minimizes the $\mathcal{C}$-bound. A second version of MinCq that minimizes the $\mathcal{C}$-bound in the transductive (or semi-supervised) setting can be found in Laviolette et al. (2011). Both versions can be expressed as quadratic programs on positive semi-definite matrices.

As is the case for Boosting algorithms (Schapire and Singer, 1999), MinCq is designed to output a Q-weighted majority vote of voters that perform rather poorly individually and, consequently, are often called weak learners. Hence, the decision of each vote is based on a small majority (*i.e.*, with a Gibbs risk just a bit lower than 1/2). Recall that in situations where the Gibbs risk is high (*i.e.*, the first moment of the margin is close to 0), the $\mathcal{C}$-bound can nevertheless remain small if the voters of the majority vote are maximally uncorrelated.

Unfortunately, minimizing the empirical value of the $\mathcal{C}$-bound tends to overfit the data. To overcome this problem, MinCq uses a distribution Q of voters which is constrained to be quasi-uniform (see Equation 37) and for which the first moment of the margin is forced to be not too close to 0. More precisely, the value $\mu_1(M_Q^S)$ is constrained to be bigger than some strictly positive constant μ. This μ then becomes a hyperparameter of the algorithm that has to be fixed by cross-validation, as the parameter C is for SVM. This new learning strategy is justified by PAC-Bound 3, dedicated to quasi-uniform posteriors[16], and PAC-Bound 3', that is specialized to kernel voters. Hence, MinCq can be viewed as the algorithm that simply looks for the majority vote of margin at least μ that minimizes PAC-Bound 3 (or PAC-Bound 3' in the sample compression case).

MinCq is also justified by two important properties of quasi-uniform majority votes. First, as we shall see in Theorem 43, there is no generality loss when restricting ourselves to quasi-uniform distributions. Second, as we shall see in Theorem 44, for any margin threshold $\mu > 0$ and any quasi-uniform distribution Q such that $\mu_1(M_Q^S) \geq \mu$, there is another quasi-uniform distribution Q' whose margin is exactly μ that achieves the same majority vote and therefore has the same $\mathcal{C}$-bound value.

Thus, to minimize the $\mathcal{C}$-bound, the learner must substantially reduce the variance of the margin distribution – *i.e.*, $\mu_2(M_Q^S)$ – while maintaining its first moment – *i.e.*, $\mu_1(M_Q^S)$ – over the threshold μ. Many learning algorithms actually exploit this strategy in different ways. Indeed, the variance of the margin distribution is controlled by Breiman (2001) for producing random forests, by Dredze et al. (2010) in the transfer learning setting, and by Shen and Li (2010) in the Boosting setting. Thus, the idea of minimizing the variance of the margin is well-known and used. We propose a new theoretical justification for all these types of algorithms and propose a novel learning algorithm, called MinCq, that directly minimizes the $\mathcal{C}$-bound.

8.1 From the $\mathcal{C}$-bound to the MinCq Learning Algorithm

We only consider learning algorithms that construct majority votes based on a (finite) self-complemented hypothesis space $\mathcal{H} = \{f_1, \ldots, f_{2n}\}$ of real-valued voters. Recall that these voters can be classifiers such as decision stumps or can be given by a kernel k evaluated on the examples of S such as $f_i(\cdot) = k(x_i, \cdot)$.

We consider the second form of the $\mathcal{C}$-bound, which relies on the first two moments of the margin of the majority vote classifier (see Theorem 11):

$$\mathcal{C}_Q^{D'} \;=\; 1 - \frac{\Big(\mu_1(M_Q^{D'})\Big)^2}{\mu_2(M_Q^{D'})}\,.$$

Our first attempts to minimize the $\mathcal{C}$-bound confronted us with two problems.

Problem 1: an empirical $\mathcal{C}$-bound minimization without any regularization tends to overfit the data.

Problem 2: most of the time, the distributions Q minimizing the $\mathcal{C}$-bound $\mathcal{C}_Q^S$ are such that both $\mu_1(M_Q^S)$ and $\mu_2(M_Q^S)$ are very close to 0. Since $\mathcal{C}_Q^S = 1 - (\mu_1(M_Q^S))^2/\mu_2(M_Q^S)$, this gives a 0/0 numerical instability. Since $(\mu_1(M_Q^D))^2/\mu_2(M_Q^D)$ can only be empirically estimated by $(\mu_1(M_Q^S))^2/\mu_2(M_Q^S)$, Problem 2 amplifies Problem 1.

[16] PAC-Bound 3 is dedicated to posteriors Q that are aligned on a prior distribution P, but in this section we always consider that the prior distribution P is uniform, thus leading to a quasi-uniform posterior Q.

A natural way to resolve Problem 1 is to restrict ourselves to quasi-uniform distributions, *i.e.*, distributions that are aligned on the uniform prior (see Section 6.1 for the definition). In Section 6, we show that with such distributions, one can upper-bound the Bayes risk without needing a KL-regularization term. Hence, according to this PAC-Bayesian theory, these distributions have some “built-in” regularization effect that should prevent overfitting. Section 7 generalizes these results to the sample compression setting, which is necessary in the case where voters such as kernels are defined using the training set.

The next theorem shows that this restriction on Q does not reduce the set of possible majority votes.

Theorem 43 *Let $\mathcal{H}$ be a self-complemented set. For all distributions Q on $\mathcal{H}$, there exists a quasi-uniform distribution Q' on $\mathcal{H}$ that gives the same majority vote as Q, and that has the same empirical and true $\mathcal{C}$-bound values, i.e.,*

$$B_{Q'} = B_Q\,, \quad \mathcal{C}^S_{Q'} = \mathcal{C}^S_Q \quad \textit{and} \quad \mathcal{C}^D_{Q'} = \mathcal{C}^D_Q\,.$$

Proof Let Q be a distribution on $\mathcal{H}=\{f_1,\ldots,f_{2n}\}$, let $M \overset{\text{def}}{=} \max_{i\in\{1,..,n\}}|Q(f_{i+n}) - Q(f_i)|$, and let Q' be defined as

$$Q'(f_i) \overset{\text{def}}{=} \frac{1}{2n} + \frac{Q(f_i)\,-\,Q(f_{i+n})}{2nM}\,,$$

where the indices of f are defined modulo $2n$ (*i.e.*, $f_{(i+n)+n} = f_i$). Then it is easy to show that Q' is a quasi-uniform distribution. Moreover, for any example $x \in \mathcal{X}$, we have

$$\begin{aligned}
\mathop{\mathbf{E}}_{f\sim Q'} f(x) \;&\overset{\text{def}}{=}\; \sum_{i=1}^{2n} Q'(f_i)\, f_i(x) \;=\; \sum_{i=1}^{n}(Q'(f_i) - Q'(f_{i+n}))\, f_i(x)\\
&=\; \sum_{i=1}^{n} \frac{2Q(f_i) - 2Q(f_{i+n})}{2nM}\, f_i(x) \;=\; \frac{1}{nM}\sum_{i=1}^{2n} Q(f_i)\, f_i(x)\\
&=\; \frac{1}{nM}\mathop{\mathbf{E}}_{f\sim Q} f(x)\,.
\end{aligned}$$

Since $nM > 0$, this implies that $B_{Q'}(x) = B_Q(x)$ for all $x \in \mathcal{X}$. It also shows that $M_{Q'}(x,y) = \frac{1}{nM}M_Q(x,y)$, which implies that $\left(\mu_1(M^{D'}_{Q'})\right)^2 = \left(\frac{1}{nM}\mu_1(M^{D'}_Q)\right)^2$ and $\mu_2(M^{D'}_{Q'}) = \left(\frac{1}{nM}\right)^2\mu_2(M^{D'}_Q)$ for both $D' = D$ and $D' = S$.

The theorem then follows from the definition of the $\mathcal{C}$-bound. ■

Theorem 43 points out a nice property of the $\mathcal{C}$-bound: different distributions Q that give rise to a same majority vote have the same (real and empirical) $\mathcal{C}$-bound values. Since the $\mathcal{C}$-bound is a bound on majority votes, this is a suitable property. Moreover, PAC-Bounds 3 and 3’, together with Theorem 43, indicate that restricting ourselves to quasi-uniform distributions is a natural solution to the problem of overfitting (see Problem 1). Unfortunately, Problem 2 remains since a consequence of the next theorem is that, among all the posteriors Q that minimize the $\mathcal{C}$-bound, there is always one whose empirical margin $\mu_1(M^S_Q)$ is as close to 0 as we want.

Theorem 44 *Let $\mathcal{H}$ be a self-complemented set. For all $\mu \in (0,1]$ and for all quasi-uniform distributions Q on $\mathcal{H}$ having an empirical margin $\mu_1(M_Q^S) \geq \mu$, there exists a quasi-uniform distribution Q' on $\mathcal{H}$, having an empirical margin equal to μ, such that Q and Q' induce the same majority vote and have the same empirical and true $\mathcal{C}$-bound values, i.e.,*

$$\mu_1(M_{Q'}^S) = \mu\,, \quad B_{Q'} = B_Q\,, \quad \mathcal{C}_{Q'}^S = \mathcal{C}_Q^S \quad \textit{and} \quad \mathcal{C}_{Q'}^D = \mathcal{C}_Q^D.$$

Proof Let Q be a quasi-uniform distribution on $\mathcal{H}\!=\!\{f_1,\ldots,f_{2n}\}$ such that $\mu_1(M_Q^S) \geq \mu$. We define Q' as

$$Q'(f_i) \stackrel{\text{def}}{=} \frac{\mu}{\mu_1(M_Q^S)} \cdot Q(f_i) + \left(1 - \frac{\mu}{\mu_1(M_Q^S)}\right) \cdot 1/2n\,, \quad i \in \{1,..,2n\}\,.$$

Clearly Q' is a quasi-uniform distribution since it is a convex combination of a quasi-uniform distribution and the uniform one. Then, similarly as in the proof of Theorem 43, one can easily show that $\mathop{\mathbf{E}}_{f\sim Q'} f(x) = \frac{\mu}{\mu_1(M_Q^S)} \mathop{\mathbf{E}}_{f\sim Q} f(x)$, which implies the result. ■

Training set bounds (such as VC-bounds for example) are known to degrade when the capacity of classification increases. As shown by Theorem 44 for the majority vote setting, this capacity increases as μ decreases to 0. Thus, we expect that any training set bound degrades for small μ. This is not the case for the $\mathcal{C}$-bound itself, but the $\mathcal{C}$-bound is not a training set bound. To obtain a training set bound, we have to relate the empirical value $\mathcal{C}_Q^S$ to the true one $\mathcal{C}_Q^D$, which is done via PAC-Bounds 3 and 3'. In these bounds, there is indeed a degradation as μ decreases because the true $\mathcal{C}$-bound is of the form $1-(\mu_1(M_Q^D))^2/\mu_2(M_Q^D)$. Since $\mu = \mu_1(M_Q^S)$, and because a small $\mu_1(M_Q^S)$ tends to produce small $\mu_2(M_Q^S)$, the bounds on $\mathcal{C}_Q^D$ given $\mathcal{C}_Q^S$ that outcomes from PAC-Bounds 3 and 3' are therefore much looser for small μ because of the 0/0 instability. As explained in the introduction of the present section, one way to overcome the instability identified in Problem 2 is to restrict ourselves to quasi-uniform distributions whose empirical margin is greater or equal than some threshold μ. Interestingly, thanks to Theorem 44, this is equivalent to restricting ourselves to distributions having empirical margin *exactly equal to* μ. From Theorems 11 and 44, it then follows that *minimizing the $\mathcal{C}$-bound, under the constraint $\mu_1(M_Q^S)\!\geq\!\mu$, is equivalent to minimizing $\mu_2(M_Q^S)$, under the constraint $\mu_1(M_Q^S)\!=\!\mu$*, from this observation, and the fact that minimizing PAC-Bounds 3 and 3' is equivalent to minimizing the empirical $\mathcal{C}$-bound $\mathcal{C}_Q^S$, we can now define the algorithm MinCq.

In this section, μ always represents a restriction on the margin. Moreover, we say that a value μ is *D'-realizable* if there exists some quasi-uniform distribution Q such that $\mu_1(M_Q^{D'}) = \mu$. The proposed algorithm, called MinCq, is then defined as follows.

Definition 45 (MinCq Algorithm) Given a self-complemented set $\mathcal{H}$ of voters, a training set S, and a S-realizable $\mu > 0$, among all quasi-uniform distributions Q of empirical margin $\mu_1(M_Q^S)$ exactly equal to μ, the algorithm MinCq consists in finding one that minimizes $\mu_2(M_Q^S)$.

This algorithm can be translated as a simple quadratic program (QP) that has only n variables (instead of $2n$), and thus can be easily solved by any QP solver. In the next subsection, we explain how the algorithm of Definition 45 can be turned into a QP.

8.2 MinCq as a Quadratic Program

Given a training set S, and a self-complemented set $\mathcal{H}$ of voters $\{f_1, f_2, \ldots, f_{2n}\}$, let

$$\mathcal{M}_i \stackrel{\text{def}}{=} \mathop{\mathbf{E}}_{(x,y)\sim S} y\, f_i(x) \qquad \text{and} \qquad \mathcal{M}_{i,j} \stackrel{\text{def}}{=} \mathop{\mathbf{E}}_{(x,y)\sim S} f_i(x)\, f_j(x)\,.$$

Let $\mathbf{M}$ be a symmetric $n \times n$ matrix, $\mathbf{a}$ be a column vector of n elements, and $\mathbf{m}$ be a column vector of n elements defined by

$$\mathbf{M} \stackrel{\text{def}}{=} \begin{bmatrix} \mathcal{M}_{1,1} & \mathcal{M}_{1,2} & \ldots & \mathcal{M}_{1,n} \\ \mathcal{M}_{2,1} & \mathcal{M}_{2,2} & \ldots & \mathcal{M}_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ \mathcal{M}_{n,1} & \mathcal{M}_{n,2} & \ldots & \mathcal{M}_{n,n} \end{bmatrix}, \quad \mathbf{a} \stackrel{\text{def}}{=} \begin{bmatrix} \frac{1}{n}\sum_{i=1}^n \mathcal{M}_{i,1} \\ \frac{1}{n}\sum_{i=1}^n \mathcal{M}_{i,2} \\ \vdots \\ \frac{1}{n}\sum_{i=1}^n \mathcal{M}_{i,n} \end{bmatrix}, \quad \text{and} \quad \mathbf{m} \stackrel{\text{def}}{=} \begin{bmatrix} \mathcal{M}_1 \\ \mathcal{M}_2 \\ \vdots \\ \mathcal{M}_n \end{bmatrix}. \tag{43}$$

Finally, let $\mathbf{q}$ be the column vector of n QP-variables, where each element q_i represents the weight $Q(f_i)$.

Using the above definitions and the fact that $\mathcal{H}$ is self-complemented, one can show that

$$\mathcal{M}_{i+n} = -\mathcal{M}_i\,, \quad \mathcal{M}_{i+n,j} = \mathcal{M}_{i,j+n} = -\mathcal{M}_{i,j}\,, \quad \text{and} \quad q_{i+n} = \frac{1}{n} - q_i\,.$$

Moreover, it follows from the definitions of the first two moments of the margin $\mu_1(M_Q^S)$ and $\mu_2(M_Q^S)$ (see Equations 6 and 8) that

$$\mu_1(M_Q^S) \;=\; \sum_{i=1}^{2n} q_i\, \mathcal{M}_i\,, \qquad \text{and} \qquad \mu_2(M_Q^S) \;=\; \sum_{i=1}^{2n}\sum_{j=1}^{2n} q_i q_j\, \mathcal{M}_{i,j}\,.$$

As MinCq consists in finding the quasi-uniform distribution Q that minimizes $\mu_2(M_Q^S)$, with a margin $\mu_1(M_Q^S)$ exactly equal to the hyperparameter μ, let us now rewrite $\mu_2(M_Q^S)$ and $\mu_1(M_Q^S)$ using the vectors and matrices defined in Equation (43). It follows that

$$\begin{aligned} \mu_2(M_Q^S) \;=\; \sum_{i=1}^{2n}\sum_{j=1}^{2n} q_i q_j\, \mathcal{M}_{i,j} \quad &= \quad \sum_{i=1}^{n}\sum_{j=1}^{n} \Big[q_i q_j - q_{i+n} q_j - q_i q_{j+n} + q_{i+n} q_{j+n} \Big] \mathcal{M}_{i,j} \\ &= \quad \sum_{i=1}^{n}\sum_{j=1}^{n} \left[4 q_i q_j - \frac{4}{n}\, q_i + \frac{1}{n^2} \right] \mathcal{M}_{i,j} \\ &= \quad 4 \sum_{i=1}^{n}\sum_{j=1}^{n} q_i q_j\, \mathcal{M}_{i,j} - \frac{4}{n} \sum_{i=1}^{n}\sum_{j=1}^{n} q_i\, \mathcal{M}_{i,j} + \frac{1}{n^2} \sum_{i=1}^{n}\sum_{j=1}^{n} \mathcal{M}_{i,j} \\ &= \quad 4\Big(\mathbf{q}^\top \mathbf{M}\, \mathbf{q} \;-\; \mathbf{a}^\top\, \mathbf{q} \Big) \;+\; \frac{1}{n^2} \sum_{i=1}^{n}\sum_{j=1}^{n} \mathcal{M}_{i,j}\,, \end{aligned} \tag{44}$$

and

$$\begin{aligned}\mu_1(M_Q^S) \;=\; \sum_{i=1}^{2n} q_i\mathcal{M}_i \;&=\; \sum_{i=1}^{n}\big(q_i - q_{i+n}\big)\mathcal{M}_i \;=\; \sum_{i=1}^{n}\left(2q_i - \frac{1}{n}\right)\mathcal{M}_i \;=\; 2\sum_{i=1}^{n} q_i\;\mathcal{M}_i - \frac{1}{n}\sum_{i=1}^{n}\mathcal{M}_i\\ &=\; 2\mathbf{m}^\top\mathbf{q} - \frac{1}{n}\sum_{i=1}^{n}\mathcal{M}_i\,.\end{aligned}$$

As the objective function $\mu_2(M_Q^S)$ and the constraint $\mu_1(M_Q^S) = \mu$ of the QP can be defined using only n variables, there is no need to consider in the QP the weights of the last n voter. These weights can always be recovered from the n first, because $q_{i+n} = \frac{1}{n} - q_i$, for any i. Note however that to be sure that the solution of the QP has the quasi-uniformity property, we have to add the following constraints to the program:

$$q_i \;\in\; [0, \tfrac{1}{n}] \qquad\qquad \text{for any } i\,.$$

Note that the multiplicative constant 4 and the additive constant $\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\mathcal{M}_{i,j}$ from Equation (44) can be omitted, as the optimal solution will stay the same. From all that precedes and given any S-realizable μ, MinCq solves the optimization problem described by Program 1.

Program 1 : **MinCq** - *a quadratic program for classification*

Solve $\quad \text{argmin}_{\mathbf{q}} \quad \mathbf{q}^\top\,\mathbf{M}\,\mathbf{q} \;-\; \mathbf{a}^\top\,\mathbf{q}$

under constraints : $\mathbf{m}^\top\;\mathbf{q} = \frac{\mu}{2} + \frac{1}{2n}\sum_{i=1}^{n}\mathcal{M}_i$

and : $0 \le q_i \le \frac{1}{n} \quad \forall i \in \{1,\ldots,n\}$

To prove that Program 1 is a quadratic program, it suffices to show that $\mathbf{M}$ is a positive semi-definite matrix. This is a direct consequence of the fact that each $\mathcal{M}_{i,j}$ can be viewed as a scalar product, since

$$\mathcal{M}_{i,j} = \left(\sqrt{\tfrac{1}{|S|}}\,f_i(x)\right)_{x\in S_{\mathcal{X}}} \;\cdot\; \left(\sqrt{\tfrac{1}{|S|}}\,f_j(x)\right)_{x\in S_{\mathcal{X}}}, \qquad \text{where } S_{\mathcal{X}} \overset{\text{def}}{=} \{x\colon (x,y)\in S\}.$$

Finally, the Q-weighted majority vote output by MinCq is

$$\begin{aligned}B_Q(x) \;=\; \text{sgn}\left[\mathop{\mathbf{E}}_{f\sim Q} f(x)\right] \;&=\; \text{sgn}\left[\sum_{i=1}^{2n} q_i f_i(x)\right] \;=\; \text{sgn}\left[\sum_{i=1}^{n} q_i f_i(x) + \sum_{i=n+1}^{2n} q_i f_i(x)\right]\\ &=\; \text{sgn}\left[\sum_{i=1}^{n} q_i f_i(x) + \sum_{i=1}^{n}(\tfrac{1}{n} - q_i)\cdot - f_i(x)\right]\\ &=\; \text{sgn}\left[\sum_{i=1}^{n}(2q_i - \tfrac{1}{n}) f_i(x)\right].\end{aligned}$$

8.3 Experiments

We now compare MinCq to state-of-the-art learning algorithms in three different contexts: *handwritten digits recognition*, *classical binary classification tasks*, and *Amazon reviews sentiment analysis*. A *context* (Lacoste et al., 2012) represents a distribution on the different tasks a learning algorithm can encounter, and a sample from a context is a collection of data sets.

For each context, each data set is randomly split into a training set S and a testing set T. When hyperparameters have to be chosen for an algorithm, 5-fold cross-validation is run on the training set S, and the hyperparameter values that minimize the mean cross-validation risk are chosen. Using these values, the algorithm is trained on the whole training set S, and then evaluated on the testing set T.

For the first two contexts, we compare MinCq using decision stumps as voters (referred to as StumpsMinCq), MinCq using RBF kernel functions $k(x, x') = \exp(-\gamma||x - x'||^2)$ as voters (referred to as RbfMinCq), AdaBoost (Schapire and Singer, 1999) using decision stumps (referred to as StumpsAdaBoost), and the soft-margin Support Vector Machine (SVM) (Cortes and Vapnik, 1995) using the RBF kernel, referred to as RbfSVM. For the last context, we compare MinCq using linear kernel functions $k(x, x') = x \cdot x'$ as voters (referred to as LinearMinCq), and the SVM using the same linear kernel, referred to as LinearSVM.

For the three variants of MinCq, the quadratic program is solved using CVXOPT (Dahl and Vandenberghe, 2007), an off-the-shelf convex optimization solver.

StumpsAdaBoost: For StumpsAdaBoost, we use decision stumps as weak learners. For each attribute, 10 decision stumps (and their complement) are generated, for a total of 20 decision stumps per attribute. The number of boosting rounds is chosen among the following 15 values: 10, 25, 50, 75, 100, 125, 150, 175, 200, 225, 250, 275, 500, 750 and 1000.

StumpsMinCq: For StumpsMinCq, we use the same 10 decision stumps per attribute as for StumpsAdaBoost. Note that we do not need to consider the complement stumps in this case, as MinCq automatically considers self-complemented sets of voters. MinCq's hyperparameter μ is chosen among 15 values between 10^{-4} and 10^0 on a logarithmic scale.

RbfSVM: The γ hyperparameter of the RBF kernel and the C hyperparameter of the SVM are chosen among 15 values between 10^{-4} and 10^1 for γ, and among 15 values between 10^0 and 10^8 for C, both on a logarithmic scale.

RbfMinCq: For RbfMinCq, we consider 15 values of μ between 10^{-4} and 10^{-2} on a logarithmic scale, and the same 15 values of γ as in SVM for the RBF kernel voters.

LinearSVM: When using the linear kernel, the C parameter of the SVM is chosen among 15 values between 10^{-4} and 10^2, on a logarithmic scale. All SVM experiments are done using the implementation of Pedregosa et al. (2011).

LinearMinCq: For LinearMinCq, we consider 15 values of μ between 10^{-4} and 10^{-2} on a logarithmic scale.

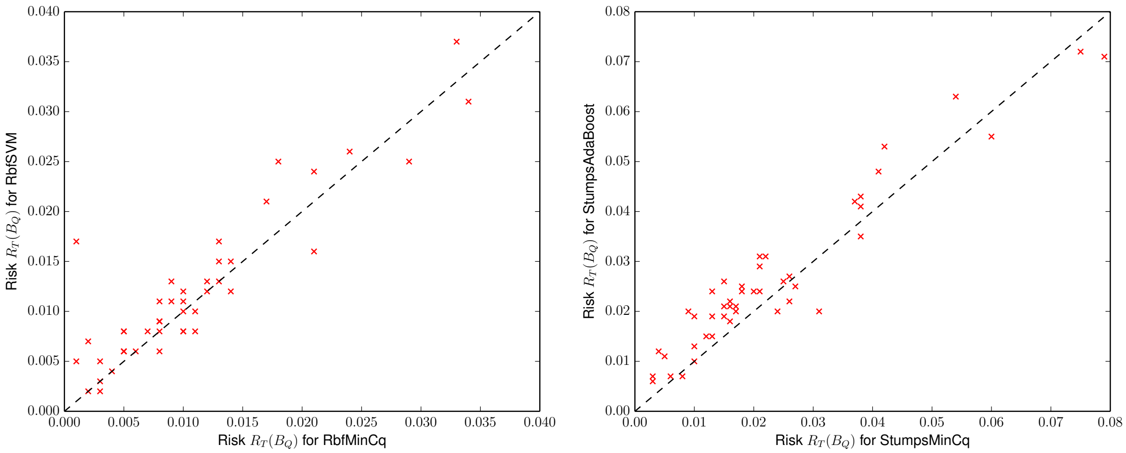


Figure 8: Comparison of the risks on the testing set for each algorithm and each MNIST binary data set. The figure on the left shows a comparison of the risks of RbfMinCq (x-axis) and RbfSVM (y-axis). The figure on the right compares StumpsMinCq (x-axis) and StumpsAdaBoost (y-axis). On each scatter plot, a point represents a pair of risks for a particular MNIST binary data set. A point above the diagonal line indicates better performance for MinCq.

	Statistical Comparison Tests	
	RbfMinCq vs RbfSVM	StumpsMinCq vs StumpsAdaBoost
Poisson binomial test	88%	99%
Sign test (p-value)	0.01	0.00

Table 2: Statistical tests comparing MinCq to either RbfSVM or StumpsAdaBoost. The Poisson binomial test gives the probability that MinCq has a better performance than another algorithm on this context. The sign test gives a p-value representing the probability that the null hypothesis is true (*i.e.*, MinCq and the other algorithm both have the same performance on this context).

When using the RBF kernel for the SVM or MinCq, each data set is normalized using a hyperbolic tangent. For each example x, each attribute $x_1, x_2, \ldots, x_n$ is renormalized with $x'_i = \tanh\left[\frac{x_i - \overline{x_i}}{\sigma_i}\right]$, where $\overline{x_i}$ and σ_i are the mean and standard deviation of the i^{th} attribute respectively, calculated on the training set S. Normalizing the features when using the RBF kernel is a common practice and gives better results for both MinCq and SVM. Empirically, we observe that the performance gain of RbfMinCq with normalized data is even more significant than for RbfSVM.

8.3.1 Handwritten Digits Recognition Context

The first context of interest to compare MinCq with other learning algorithms is the handwritten digits recognition. For this task, we use the *MNIST database of handwritten digits* of Lecun and Cortes. We split the original data set into 45 binary classification tasks, where the union of all binary data sets recovers the original data set, and the intersection of any pair of binary data sets gives the empty set. Therefore, any example from the original data set appears on one and only one binary data set, thus avoiding any correlation between the binary data sets. For each resulting binary data set, we randomly choose 500 examples to be in the training set S, and the testing set T consists of the remaining examples. Figure 8 shows the resulting test risk for each binary data set and each algorithm.

Table 2 shows two statistical tests to compare the algorithms on the handwritten digits recognition context: the Poisson binomial test (Lacoste et al., 2012) and the sign test (Mendenhall, 1983). Both methods suggest that RbfMinCq outperforms RbfSVM on this context, and that StumpsMinCq outperforms StumpsAdaBoost.

8.3.2 Classical Binary Classification Tasks Context

This second context of interest is a more general one: it consists of multiple binary classification data sets coming from the UCI Machine Learning Repository (Blake and Merz, 1998). These data sets are commonly used as a benchmark for learning algorithms, and may help to answer the question "How well may a learning algorithm perform on many unrelated classification tasks". For each data set, half of the examples (up to a maximum of 500) are randomly chosen to be in the training set S, and the remaining examples are in the testing set T. Table 3 shows the resulting test risks on this context, for each algorithm.

Table 3 also shows a statistical comparison of all algorithms on the classical binary classification tasks context, using the Poisson binomial test and the sign test. On this context, both statistical tests show no significant performance difference between RbfMinCq and RbfSVM, and between StumpsMinCq and StumpsAdaBoost, implying that these pairs of algorithms perform similarly well on this general context.

8.3.3 Amazon Reviews Sentiment Analysis

This context contains 4 sentiment analysis data sets, representing product types (*books*, *DVDs*, *electronics* and *kitchen appliances*). The task is to learn from an Amazon.com product user review in natural language, and predict the *polarity* of the review, that is either negative (3 stars or less) or positive (4 or 5 stars). The data sets come from Blitzer et al. (2007), where the natural language reviews have already been converted into a set of *unigrams* and *bigrams* of terms, with a count. For each data set, a training set of 1000 positive reviews and 1000 negative reviews are provided, and the remaining reviews are available in a testing set. The original feature space of these data sets is between $90,000$ and $200,000$ dimensions. However, as most of the unigrams and bigrams are not significant and to reduce the dimensionality, we only consider unigrams and bigrams that appear at least 10 times on the training set (as in Chen et al., 2011), reducing the numbers of dimensions to between 3500 and 6000. Again as in Chen et al. (2011), we apply standard *tf-idf* feature re-weighting (Salton and Buckley, 1988). Table 4 shows the resulting test risks for each algorithm.

Data Set Information			Risk $R_T(B_Q)$ for Each Algorithm			
Name	$\|S\|$	$\|T\|$	RbfMinCq	RbfSVM	StumpsMinCq	StumpsAdaBoost
Australian	345	345	0.142	**0.133**	**0.165**	0.168
Balance	313	312	0.054	**0.042**	0.042	**0.032**
BreastCancer	350	349	**0.037**	0.046	**0.037**	0.060
Car	500	1228	0.074	**0.032**	0.320	**0.291**
Cmc	500	973	**0.303**	0.306	0.140	**0.134**
Credit-A	345	345	**0.122**	0.133	**0.304**	0.308
Cylinder	270	270	**0.204**	0.233	**0.125**	0.148
Ecoli	168	168	0.077	**0.071**	**0.289**	**0.289**
Flags	97	97	**0.289**	0.320	**0.071**	**0.071**
Glass	107	107	**0.206**	**0.206**	**0.268**	0.309
Heart	135	135	0.163	**0.156**	**0.262**	0.271
Hepatitis	78	77	0.169	**0.143**	**0.185**	**0.185**
Horse	184	184	**0.185**	0.196	**0.169**	0.221
Ionosphere	176	175	0.114	**0.069**	0.245	**0.174**
Letter:AB	500	1055	0.007	**0.003**	**0.109**	0.120
Letter:DO	500	1058	0.021	**0.018**	**0.005**	0.010
Letter:OQ	500	1036	**0.023**	0.036	**0.020**	0.048
Liver	173	172	**0.267**	0.285	**0.042**	0.052
Monks	216	216	0.245	**0.208**	0.306	**0.236**
Nursery	500	12459	**0.025**	0.026	**0.025**	0.026
Optdigits	500	3323	0.034	**0.027**	**0.089**	**0.089**
Pageblock	500	4973	**0.045**	0.048	0.059	**0.055**
Pendigits	500	6994	**0.007**	0.008	**0.069**	0.084
Pima	384	384	**0.253**	0.255	0.273	**0.250**
Segment	500	1810	**0.017**	0.018	0.040	**0.022**
Spambase	500	4101	**0.067**	0.077	0.133	**0.070**
Tic-tac-toe	479	479	0.033	**0.025**	**0.330**	0.353
USvote	218	217	**0.051**	**0.051**	**0.051**	**0.051**
Wine	89	89	**0.034**	0.045	0.169	**0.034**
Yeast	500	984	0.286	**0.279**	0.324	**0.306**
Zoo	51	50	**0.040**	0.060	0.060	**0.040**

	Statistical Comparison Tests	
	RbfMinCq vs RbfSVM	StumpsMinCq vs StumpsAdaBoost
Poisson binomial test	54%	48%
Sign test (p-value)	0.36	0.35

Table 3: Risk on the testing set for all algorithms, on the classical binary classification task context. See Table 2 for an explanation of the statistical tests.

Table 4 also shows a statistical comparison of the algorithms on this context, again using the Poisson binomial test and the sign test. LinearMinCq has an edge over LinearSVM, as it wins or draws on each data set. However, both statistical tests show no significant performance difference between LinearMinCq and LinearSVM.

These experiments show that minimizing the $\mathcal{C}$-bound, and thus favoring majority votes for which the voters are maximally uncorrelated, is a sound approach. MinCq is very competitive with both AdaBoost and the SVM on the classical binary tasks context and the Amazon reviews sentiment analysis context. MinCq even shows a highly significant performance gain on the handwritten digits recognition context, implying that on certain types of tasks or data sets, minimizing the $\mathcal{C}$-bound offers a state-of-the-art performance.

Data Set Information			Risk $R_T(B_Q)$ for Each Algorithm	
Name	$\lvert S\rvert$	$\lvert T\rvert$	LinearMinCq	LinearSVM
Books	2000	4465	**0.158**	**0.158**
DVD	2000	3586	**0.162**	0.163
Kitchen	2000	5945	**0.130**	0.131
Electronics	2000	5681	**0.116**	0.118

Statistical Comparison Tests	
	LinearMinCq vs LinearSVM
Poisson binomial test	68%
Sign test (p-value)	0.31

Table 4: Risk on the testing set for all algorithms, on the Amazon reviews sentiment analysis context. See Table 2 for an explanation of the statistical tests.

However, for all above experiments, we observe that the empirical values of the PAC-Bounds are trivial (close to 1). Remember that, inspired by PAC-Bounds 3 and 3', the MinCq algorithm learns the weights of a majority vote by minimizing the second moment of the margin while fixing its first moment μ to some value. In these experiments, the value of μ chosen by cross-validation is always very close to 0 (basically, $\mu = 10^{-4}$). This implies that $\mathcal{C}_Q^S = 1 - \frac{\mu^2}{\mu_2(M_Q^S)}$ is very close to the $1 - \frac{0}{0}$ form, leading to a severe degradation of PAC-Bayesian bounds for $\mathcal{C}_Q^D$. Note that the voters were all *weak* in the former experiments. This explains why very small values of μ were selected by cross-validation.

8.3.4 Experiments with Stronger Voters

In the following experiment, we show that one can obtain much better bound values by using *stronger* voters, that is, voters with a better individual performance. To do so, instead of considering decision stumps, we consider decision trees.[17] We use 100 decision tree classifiers generated with the implementation of Pedregosa et al. (2011) (we set the maximum depth to 10 and the number of features per node to 1). By using these strong voters, it is possible to achieve higher values of μ.[18]

Figure 9 shows the empirical $\mathcal{C}$-bound value and its corresponding PAC-Bayesian bound values for multiple values of μ on the Mushroom UCI data set. From the 8124 examples, 500 have been used to construct the set of voters, 4062 for the training set, and the remaining examples for the testing set. The figure shows the PAC-Bayesian bounds get tighter when μ is increasing. Note however that the empirical $\mathcal{C}$-bound slightly increases from 0.001 to 0.016. The risk on the testing set of the majority vote (not shown in the figure) is 0 for most values of μ, but also increases a bit for the highest values (remaining below 0.001).

[17] A decision stump can be seen as a (weak) decision tree of depth 1.

[18] Note that the set of decision trees was learned on a *fresh* set of examples, disjoint from the training data. We do so to ensure that all computed PAC-Bounds are valid, even if they are not designed to handle *sample-compressed* voters.

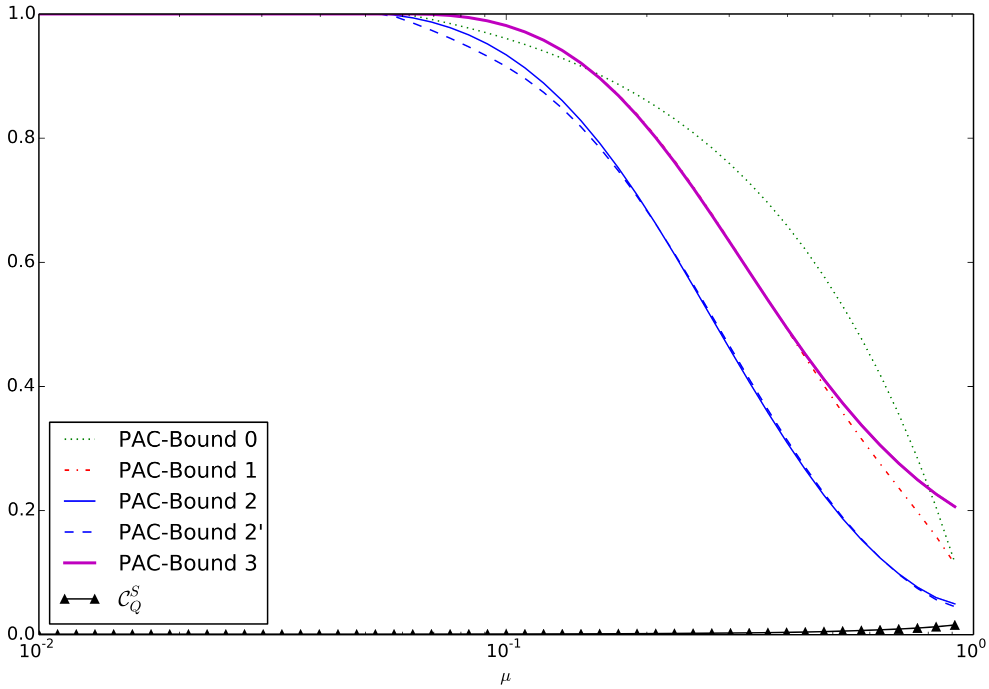


Figure 9: Values of empirical $\mathcal{C}$-bound and corresponding PAC-Bounds 0, 1, 2, 2' and 3 on the majority votes output by MinCq, for multiple values of μ.

Hence, we obtain tight bounds for high values of μ (PAC-Bounds 2 and 2' are under 0.2). Nevertheless, these PAC-Bayesian bounds are not tight enough to precisely guide the selection of μ. This is why we rely on cross-validation to select a good value of μ.

Finally, we also see that PAC-Bound 3 is looser than other bounds over $\mathcal{C}_Q^D$, but this was expected as it was not designed to be as tight as possible. That being said, PAC-Bound 3 has the same behavior than PAC-Bounds 1 and 2. This suggests that we can rely on it to justify the MinCq learning algorithm once the hyperparameter μ is fixed.

9. Conclusion

In this paper, we have revisited the work presented in Lacasse et al. (2006) and Laviolette et al. (2011). We clarified the presentation of previous results and extended them, as well as actualizing the discussion regarding the ever growing development of PAC-Bayesian theory.

We have derived a risk bound (called the $\mathcal{C}$-bound) for the weighted majority vote that depends on the first and the second moment of the associated margin distribution (Theorem 11). The proposed bound is based on the one-sided Chebyshev inequality, which, under the mild condition of Proposition 14, is the tightest inequality for any real-valued random variable given only its first two moments. Also, as shown empirically by Figure 3, this bound has a strong predictive power on the risk of the majority vote.

We have also shown that the original PAC-Bayesian theorem, together with new ones, can be used to obtain high-confidence estimates of this new risk bound that holds uniformly for *all* posterior distributions. We have generalized these PAC-Bayesian results to the (more general) sample compression setting, allowing one to make use of voters that are constructed with elements of the training data, such as kernel functions $y_i k(x_i, \cdot)$. Moreover, we have presented PAC-Bayesian bounds that have the uncommon property of having no Kullback-Leibler divergence term (PAC-Bounds 3 and 3'). These bounds, together with the $\mathcal{C}$-bound, gave the theoretical foundation to the learning algorithm introduced at the end of the paper, that we have called MinCq. The latter turns out to be expressible in the nice form of a quadratic program. MinCq is not only based on solid theoretical guarantees, it also performs very well on natural data, namely when compared with the state-of-the-art SVM.

This work tackled the simplest problem in machine learning (the supervised binary classification in presence of i.i.d. data), and we now consider that the PAC-Bayesian theory is mature enough to embrace a variety of more sophisticated frameworks. Indeed, in the recent years several authors applied this theory to many more complex paradigms: Transductive Learning (Derbeko et al., 2004; Catoni, 2007; Bégin et al., 2014), Domain Adaptation (Germain et al., 2013), Density Estimation (Seldin and Tishby, 2009; Higgs and Shawe-Taylor, 2010), Structured output Prediction (McAllester, 2007; Giguère et al., 2013; London et al., 2014), Co-clustering (Seldin and Tishby, 2009, 2010), Martingales (Seldin et al., 2012), U-Statistics of higher order (Lever et al., 2013) or other non-i.i.d. settings (Ralaivola et al., 2010), Multi-armed Bandit (Seldin et al., 2011) and Reinforcement Learning (Fard and Pineau, 2010; Fard et al., 2011).

Acknowledgements

This work has been supported by National Science and Engineering Research Council (NSERC) Discovery grants 262067 and 0122405. Computations were performed on the Colosse supercomputer grid at Université Laval, under the auspices of Calcul Québec and Compute Canada. The operations of Colosse are funded by the NSERC, the Canada Foundation for Innovation (CFI), NanoQuébec, and the Fonds de recherche du Québec – Nature et technologies (FRQNT).

We would also like to thank Nicolas Usunier for his insightful comments and participation in the preliminary work on majority votes, and Malik Younsi for solving one of our conjectures leading to the nice $\xi(m)+m$ term in the statement of Theorem 28.

Finally, we sincerely thank the three anonymous reviewers for the exceptional quality of their work.

Appendix A. Auxiliary mathematical results

Lemma 46 (Markov's inequality) *For any random variable X such that $\mathbf{E}(X) = \mu$, and for any $a > 0$, we have*

$$\Pr\left(|X| \geq a\right) \;\leq\; \frac{\mu}{a}\,.$$

Lemma 47 (Jensen's inequality) *For any random variable X and any convex function f, we have*

$$f(\mathbf{E}\,[X]) \;\leq\; \mathbf{E}\,[f(X)]\,.$$

Lemma 48 (One-sided Chebyshev inequality) *For any random variable X such that $\mathbf{E}(X) = \mu$ and $\mathbf{Var}(X) = \sigma^2$, and for any $a > 0$, we have*

$$\Pr\Big(X - \mu \geq a\Big) \;\leq\; \frac{\sigma^2}{\sigma^2 + a^2}\,.$$

Proof First observe that $\Pr\Big(X-\mu \geq a\Big) \leq \Pr\Big(\,\Big[X - \mu + \frac{\sigma^2}{a}\Big]^2 \geq \Big[a + \frac{\sigma^2}{a}\Big]^2\Big)$. Let us now apply Markov's inequality (Lemma 46) to bound this probability. We obtain

$$\begin{aligned}
\Pr\left(\left[X - \mu + \frac{\sigma^2}{a}\right]^2 \geq \left[a + \frac{\sigma^2}{a}\right]^2\right) \;&\leq\; \frac{\mathbf{E}\left[X - \mu + \frac{\sigma^2}{a}\right]^2}{\left[a + \frac{\sigma^2}{a}\right]^2} && \text{(Markov's inequality)}\\
&=\; \frac{\mathbf{E}\,(X-\mu)^2 + 2\left(\frac{\sigma^2}{a}\right)\mathbf{E}\,(X-\mu) + \left(\frac{\sigma^2}{a}\right)^2}{\left[a + \frac{\sigma^2}{a}\right]^2}\\
&=\; \frac{\sigma^2 + \left(\frac{\sigma^2}{a}\right)^2}{\left[a + \frac{\sigma^2}{a}\right]^2} \;=\; \frac{\sigma^2\left(1 + \frac{\sigma^2}{a^2}\right)}{(\sigma^2 + a^2)\left(1 + \frac{\sigma^2}{a^2}\right)} \;=\; \frac{\sigma^2}{\sigma^2 + a^2}\,,
\end{aligned}$$

because $\mathbf{E}\,(X-\mu)^2 = \mathbf{Var}(X) = \sigma^2$ and $\mathbf{E}\,(X-\mu) = \mathbf{E}(X) - \mathbf{E}(X) = 0$. ■

Note that the proof Theorem 49 (below) by Cover and Thomas (1991) considers that probability distributions Q and P are discrete, but their argument is straightforwardly generalizable to continuous distributions.

Theorem 49 (Cover and Thomas, 1991, Theorem 2.7.2) *The Kullback-Leibler divergence* $\mathrm{KL}(Q\|P)$ *is convex in the pair* (Q,P), i.e., *if* (Q_1,P_1) *and* (Q_2,P_2) *are two pairs of probability distributions, then*

$$\mathrm{KL}\big(\,\lambda Q_1 + (1{-}\lambda)Q_2 \,\big\|\, \lambda P_1 + (1{-}\lambda)P_2\,\big) \;\leq\; \lambda\,\mathrm{KL}\big(Q_1\big\|P_1\big) + (1{-}\lambda)\,\mathrm{KL}\big(Q_2\big\|P_2\big)\,,$$

for all $\lambda \in [0,1]\,.$

Corollary 50 *Both following functions are convex:*

1. *The function* $\mathrm{kl}(q\|p)$ *of Equation* (21), i.e., *the Kullback-Leibler divergence between two Bernoulli distributions;*

2. *The function* $\mathrm{kl}(q_1,q_2\|p_1,p_2)$ *of Equation* (31), i.e., *the Kullback-Leibler divergence between two distributions of trivalent random variables.*

Proof Straightforward consequence of Theorem 49. ■

Lemma 51 (Maurer, 2004) *Let X be any random variable with values in $[0,1]$ and expectation $\mu=\mathbf{E}(X)$. Denote $\mathbf{X}$ the vector containing the results of n independent realizations of X. Then, consider a Bernoulli random variable X' ($\{0,1\}$-valued) of probability of success μ,* i.e., $\Pr(X'=1)=\mu$. *Denote $\mathbf{X}'\in\{0,1\}^n$ the vector containing the results of n independent realizations of X'.*

If function $f:[0,1]^n\to\mathbb{R}$ is convex, then

$$\mathbf{E}\big[f(\mathbf{X})\big] \;\le\; \mathbf{E}\big[f(\mathbf{X}')\big]\,.$$

The proof of Lemma 52 (below) follows the key steps of the proof of Lemma 51 by Maurer (2004), but we include a few more mathematical details for completeness. Interestingly, the proof highlights that one can generalize Maurer's lemma even more, to embrace random variables of any (countable) number of possible outputs. Note that another generalization of Maurer's lemma is given in Seldin et al. (2012) to embrace the case where the random variables $X_1,\ldots,X_n$ are a martingale sequence instead of being independent.

Lemma 52 (Generalization of Lemma 51) *Let the tuple (X,Y) be a random variable with values in $[0,1]^2$, such that $X+Y\le 1$, and with expectation $(\mu_X,\mu_Y)=(\mathbf{E}(X),\mathbf{E}(Y))$. Given n independent realizations of (X,Y), denote $\mathbf{X}=(X_1,\ldots,X_n)$ the vector of corresponding X-values and $\mathbf{Y}=(Y_1,\ldots,Y_n)$ the vector of corresponding Y-values. Then, consider a random variable (X',Y') with three possible outcomes, $(1,0)$, $(0,1)$ and $(0,0)$, of expectations μ_X, μ_Y and $1-\mu_X-\mu_Y$, respectively. Denote $\mathbf{X}',\mathbf{Y}'\in\{0,1\}^n$ the vectors of n independent realizations of (X',Y').*

If a function $f:[0,1]^n\times[0,1]^n\to\mathbb{R}$ is convex, then

$$\mathbf{E}\big[f(\mathbf{X},\mathbf{Y})\big] \;\le\; \mathbf{E}\big[f(\mathbf{X}',\mathbf{Y}')\big]\,.$$

Proof Given two vectors $\mathbf{x}=(x_1,\ldots,x_n),\mathbf{y}=(y_1,\ldots,y_n)\in[0,1]^n$, let us define

$$(\mathbf{x},\mathbf{y}) \;\stackrel{\text{def}}{=}\; \big((x_1,y_1),(x_2,y_2),\ldots,(x_n,y_n)\big) \;\in\; ([0,1]\times[0,1])^n\,.$$

Consider $H=\{(1,0),(0,1),(0,0)\}$. Lemma 53 (below) shows that any point $(\mathbf{x},\mathbf{y})$ can be written as a convex combination of the extreme points $\boldsymbol{\eta}=(\eta_1,\eta_2,\ldots,\eta_n)\in H^n$:

$$(\mathbf{x},\mathbf{y}) \;=\; \sum_{\boldsymbol{\eta}\in H^n}\left[\left(\prod_{i:\eta_i=(1,0)} x_i\right)\left(\prod_{i:\eta_i=(0,1)} y_i\right)\left(\prod_{i:\eta_i=(0,0)} 1-x_i-y_i\right)\right]\cdot\boldsymbol{\eta}\,. \tag{45}$$

Convexity of function f implies

$$f(\mathbf{x},\mathbf{y}) \;\le\; \sum_{\boldsymbol{\eta}\in H^n}\left[\left(\prod_{i:\eta_i=(1,0)} x_i\right)\left(\prod_{i:\eta_i=(0,1)} y_i\right)\left(\prod_{i:\eta_i=(0,0)} 1-x_i-y_i\right)\right]\cdot f(\boldsymbol{\eta})\,, \tag{46}$$

with equality if $(\mathbf{x},\mathbf{y})\in H^n=\{(1,0),(0,1),(0,0)\}^n$, because the elements of the sum are $0\cdot f(\boldsymbol{\eta})$ for all $\boldsymbol{\eta}\in H^n\setminus\{(\mathbf{x},\mathbf{y})\}$ and $1\cdot f(\boldsymbol{\eta})$ only for $\boldsymbol{\eta}=(\mathbf{x},\mathbf{y})$.

Given that realizations of random variable (X,Y) are independent and that for a given $\eta_i \in H$, only one of the three products is computed[19], we get

$$\begin{aligned}
\mathbf{E}\left[f(\mathbf{X},\mathbf{Y})\right] &\leq \mathbf{E}\left[\sum_{\boldsymbol{\eta}\in H^n}\left[\left(\prod_{i:\eta_i=(1,0)} X_i\right)\left(\prod_{i:\eta_i=(0,1)} Y_i\right)\left(\prod_{i:\eta_i=(0,0)} 1-X_i-Y_i\right)\right]\cdot f(\boldsymbol{\eta})\right]\\
&= \sum_{\boldsymbol{\eta}\in H^n}\mathbf{E}\left[\left(\prod_{i:\eta_i=(1,0)} X_i\right)\left(\prod_{i:\eta_i=(0,1)} Y_i\right)\left(\prod_{i:\eta_i=(0,0)} 1-X_i-Y_i\right)\right]\cdot f(\boldsymbol{\eta})\\
&= \sum_{\boldsymbol{\eta}\in H^n}\left[\left(\prod_{i:\eta_i=(1,0)} \mathbf{E}(X_i)\right)\left(\prod_{i:\eta_i=(0,1)} \mathbf{E}(Y_i)\right)\left(\prod_{i:\eta_i=(0,0)} 1-\mathbf{E}(X_i)-\mathbf{E}(Y_i)\right)\right]\cdot f(\boldsymbol{\eta})\\
&= \sum_{\boldsymbol{\eta}\in H^n}\left[\left(\prod_{i:\eta_i=(1,0)} \mu_X\right)\left(\prod_{i:\eta_i=(0,1)} \mu_Y\right)\left(\prod_{i:\eta_i=(0,0)} 1-\mu_X-\mu_Y\right)\right]\cdot f(\boldsymbol{\eta})\,.
\end{aligned}$$

This becomes an equality when $(\mathbf{X},\mathbf{Y})$ takes values in H^n (as we explain after equation 46). We therefore conclude that $\mathbf{E}\big[f(\mathbf{X},\mathbf{Y})\big] \leq \mathbf{E}\big[f(\mathbf{X}',\mathbf{Y}')\big]$. ■

Lemma 53 (Proof of Equation 45) *Consider* $H=\{(1,0),(0,1),(0,0)\}$ *and an integer* $n>0$. *Any point* $(\mathbf{x},\mathbf{y})\in\big([0,1]\times[0,1]\big)^n$ *can be written as a convex combination of the extreme points* $\boldsymbol{\eta}=(\eta_1,\eta_2,\ldots,\eta_n)\in H^n$*:*

$$(\mathbf{x},\mathbf{y}) \;=\; \sum_{\boldsymbol{\eta}\in H^n}\rho_{\boldsymbol{\eta}}(\mathbf{x},\mathbf{y})\cdot\boldsymbol{\eta}\,,$$

where

$$\rho_{\boldsymbol{\eta}}(\mathbf{x},\mathbf{y}) \;\stackrel{\text{def}}{=}\; \left(\prod_{i:\eta_i=(1,0)} x_i\right)\left(\prod_{i:\eta_i=(0,1)} y_i\right)\left(\prod_{i:\eta_i=(0,0)} 1-x_i-y_i\right).$$

Proof We prove the result by induction over vector size n.
Proof for $n=1$:

$$\begin{aligned}
\sum_{\boldsymbol{\eta}\in H}\rho_{\boldsymbol{\eta}}((x_1,y_1))\cdot\boldsymbol{\eta} &= x_1\cdot((1,0))+y_1\cdot((0,1))+(1-x_1-y_1)\cdot((0,0))\\
&= ((x_1,y_1))\,.
\end{aligned}$$

Proof for $n>1$: We suppose that the result is true for any vector $(\mathbf{x},\mathbf{y})$ of a particular size n (this is our induction hypothesis) and we prove that it implies

$$\sum_{(\boldsymbol{\eta},\eta_{n+1})\in H^{n+1}}\Big[\rho_{(\boldsymbol{\eta},\eta_{n+1})}\big((\mathbf{x},\mathbf{y}),(x_{n+1},y_{n+1})\big)\Big]\cdot(\boldsymbol{\eta},\eta_{n+1}) \;=\; \big((\mathbf{x},\mathbf{y}),\,(x_{n+1},y_{n+1})\big)\,,$$

where $(\mathbf{a},\,b)$ denotes a vector $\mathbf{a}$, augmented by one element b.

[19] The equality between the second and third lines follows from the fact that each expectation inside the sum of Line 2 can be rewritten as the following product of independent random variables:

$$\mathbf{E}\Big[\prod_{\eta_i} g_{\eta_i}(X_i,Y_i)\Big] \quad\text{with}\quad g_{\eta_i}(X_i,Y_i) \stackrel{\text{def}}{=} \begin{cases} X_i & \text{if } \eta_i=(1,0)\\ Y_i & \text{if } \eta_i=(0,1)\\ 1-X_i-Y_i & \text{otherwise.}\end{cases}$$

We have

$$
\begin{aligned}
&\sum_{(\boldsymbol{\eta},\eta_{n+1})\,\in H^{n+1}} \Big[\rho_{(\boldsymbol{\eta},\eta_{n+1})}\big((\mathbf{x},\mathbf{y}),(x_{n+1},y_{n+1})\big)\Big]\cdot(\boldsymbol{\eta},\eta_{n+1})\\
&\quad= \sum_{\boldsymbol{\eta}\in H^n}\rho_{\boldsymbol{\eta}}(\mathbf{x},\mathbf{y})\cdot x_{n+1}\cdot\big(\boldsymbol{\eta},(1,0)\big)+\sum_{\boldsymbol{\eta}\in H^n}\rho_{\boldsymbol{\eta}}(\mathbf{x},\mathbf{y})\cdot y_{n+1}\cdot\big(\boldsymbol{\eta},(0,1)\big)\\
&\qquad\qquad+\sum_{\boldsymbol{\eta}\in H^n}\rho_{\boldsymbol{\eta}}(\mathbf{x},\mathbf{y})\cdot(1-x_{n+1}-y_{n+1})\cdot\big(\boldsymbol{\eta},(0,0)\big)\\
&\quad= \left(\sum_{\boldsymbol{\eta}\in H^n}\rho_{\boldsymbol{\eta}}(\mathbf{x},\mathbf{y})\cdot(x_{n+1}+y_{n+1}+1-x_{n+1}-y_{n+1})\cdot\boldsymbol{\eta},\ \sum_{\boldsymbol{\eta}\in H^n}\rho_{\boldsymbol{\eta}}(\mathbf{x},\mathbf{y})\cdot\big(x_{n+1},y_{n+1}\big)\right)\\
&\quad= \left(\sum_{\boldsymbol{\eta}\in H^n}\rho_{\boldsymbol{\eta}}(\mathbf{x},\mathbf{y})\cdot\boldsymbol{\eta},\ \sum_{\boldsymbol{\eta}\in H^n}\rho_{\boldsymbol{\eta}}(\mathbf{x},\mathbf{y})\cdot\big(x_{n+1},y_{n+1}\big)\right)\\
&\quad= \Big((\mathbf{x},\mathbf{y}),\ \big(x_{n+1},y_{n+1}\big)\Big)\,.
\end{aligned}
$$

For the last equality, the $(\mathbf{x},\mathbf{y})$ term of the vector above is obtained from the induction hypothesis and the last couple is a direct consequence of the following equality:

$$
\sum_{\boldsymbol{\eta}\in H^n}\rho_{\boldsymbol{\eta}}(\mathbf{x},\mathbf{y}) \;=\; \prod_{i=1}^{n}\Big(x_i+y_i+1-x_i-y_i\Big) \;=\; 1\,.
$$

■

Proposition 54 (Concavity of Equation 36) *The function $F_c(d,e)$ is concave.*

Proof We show that the Hessian matrix of $F_c(d,e)$ is a negative semi-definite matrix. In other words, we need to prove that

$$
\frac{\partial^2 F_c(d,e)}{\partial d^2}\le 0\,;\qquad \frac{\partial^2 F_c(d,e)}{\partial e^2}\le 0\,;\qquad \frac{\partial^2 F_c(d,e)}{\partial d^2}\frac{\partial^2 F_c(d,e)}{\partial e^2}-\left(\frac{\partial^2 F_c(d,e)}{\partial d\partial e}\right)^2\ge 0\,.
$$

Indeed, we have

$$
\begin{aligned}
\frac{\partial^2 F_c(d,e)}{\partial d^2} &= \frac{2(1-4e)^2}{(2d-1)^3}\le 0 \quad \forall e\in[0,1], d\in\left[0,\frac{1}{2}\right],\\
\frac{\partial^2 F_c(d,e)}{\partial e^2} &= \frac{8}{2d-1}\le 0 \qquad\quad \forall e\in[0,1], d\in\left[0,\frac{1}{2}\right],\\
\frac{\partial^2 F_c(d,e)}{\partial d^2}\frac{\partial^2 F_c(d,e)}{\partial e^2}-\left(\frac{\partial^2 F_c(d,e)}{\partial d\partial e}\right)^2 &= \frac{2(1-4e)^2}{(2d-1)^3}\cdot\frac{8}{2d-1}-\left(\frac{4-16e}{(1-2d)^2}\right)^2=0\,.
\end{aligned}
$$

■

Appendix B. A General PAC-Bayesian Theorem for Tuples of Voters and Aligned Posteriors

This section presents a change of measure inequality that generalizes both Lemmas 30 and 34, and a PAC-Bayesian theorem that generalizes both Theorems 31 and 35. As these generalizations require more complex notation and ideas, it is provided as an appendix and the simpler versions of the main paper have separate proofs.

Let $\mathcal{H}$ be a countable self-complemented set real-valued functions. In the general setting, we recall that $\mathcal{H}$ is self-complemented if there exists a bijection $c : \mathcal{H} \to \mathcal{H}$ such that $c(f) = -f$ for any $f \in \mathcal{H}$. Moreover, for a distribution Q aligned on a prior distribution P and for any $f \in \mathcal{H}$, we have

$$Q(f) + Q(c(f)) \quad = \quad P(f) + P(c(f))\,.$$

First, we need to define the following notation. Let $\mathbf{k}$ be a sequence of length k, containing numbers representing indices of voters. Let $f_{\mathbf{k}} : \mathcal{X} \to \overline{\mathcal{Y}}^k$ be a function that outputs a tuple of votes, such that $f_{\mathbf{k}}(x) \overset{\text{def}}{=} \langle f_{\mathbf{k}_1}(x), \ldots, f_{\mathbf{k}_k}(x)\rangle\,.$

Let us recall that P^k and Q^k are Cartesian products of probability distributions P and Q. Thus, the probability of drawing $f_{\mathbf{k}} \sim Q^k$ is given by

$$Q^k(f_{\mathbf{k}}) \overset{\text{def}}{=} Q(f_{\mathbf{k}_1}) \cdot Q(f_{\mathbf{k}_2}) \cdot \ldots \cdot Q(f_{\mathbf{k}_k}) = \prod_{i=1}^{k} Q(f_{\mathbf{k}_i})\,.$$

Finally, for each $f_{\mathbf{k}}$ and each $j \in \{0, \ldots, 2^k - 1\}$, let

$$f_{\mathbf{k}}^{[j]}(x) \overset{\text{def}}{=} \langle f_{\mathbf{k}_1}^{(s_1^j)}(x), \ldots, f_{\mathbf{k}_k}^{(s_k^j)}(x)\rangle\ ,$$

where $s_1^j s_2^j \ldots s_k^j$ is the binary representation of the number j, and where $f^{(0)} = f$ and $f^{(1)} = c(f)$. Note that $f_{\mathbf{k}}^{[0]} = f_{\mathbf{k}}$.

To prove the next PAC-Bayesian theorem, we make use of the following change of measure inequality.

Theorem 55 (Change of measure inequality for tuples of voters and aligned posteriors) *For any self-complemented set $\mathcal{H}$, for any distribution P on $\mathcal{H}$, for any distribution Q aligned on P, and for any measurable function $\phi : \mathcal{H}^k \to \mathbb{R}$ for which $\phi(f_{\mathbf{k}}^{[j]}) = \phi(f_{\mathbf{k}}^{[j']})$ for any $j, j' \in \{0, \ldots, 2^k - 1\}$, we have*

$$\mathop{\mathbf{E}}_{f_{\mathbf{k}} \sim Q^k} \phi(f_{\mathbf{k}}) \ \leq\ \ln\left(\mathop{\mathbf{E}}_{f_{\mathbf{k}} \sim P^k} e^{\phi(f_{\mathbf{k}})}\right).$$

Proof First, note that one can change the expectation over Q^k to an expectation over P^k, using the fact that $\phi(f_{\mathbf{k}}^{[j]}) = \phi(f_{\mathbf{k}}^{[j']})$ for any $j, j' \in \{0, \ldots, 2^k - 1\}$ and that Q is aligned on P.

$$
\begin{aligned}
& 2^k \cdot \mathop{\mathbf{E}}_{f_{\mathbf{k}} \sim Q^k} \phi(f_{\mathbf{k}}) \\
&= \int_{\mathcal{H}^k} df_{\mathbf{k}}\; Q^k(f_{\mathbf{k}}^{[0]})\, \phi(f_{\mathbf{k}}^{[0]}) + \int_{\mathcal{H}^k} df_{\mathbf{k}}\; Q^k(f_{\mathbf{k}}^{[1]})\, \phi(f_{\mathbf{k}}^{[1]}) + \ldots + \int_{\mathcal{H}^k} df_{\mathbf{k}}\; Q^k(f_{\mathbf{k}}^{[2^k-1]})\, \phi(f_{\mathbf{k}}^{[2^k-1]}) \\
&= \int_{\mathcal{H}^k} df_{\mathbf{k}}\; Q^k(f_{\mathbf{k}}^{[0]})\, \phi(f_{\mathbf{k}}) \;+\; \int_{\mathcal{H}^k} df_{\mathbf{k}}\; Q^k(f_{\mathbf{k}}^{[1]})\, \phi(f_{\mathbf{k}}) \;+\; \ldots \;+\; \int_{\mathcal{H}^k} df_{\mathbf{k}}\; Q^k(f_{\mathbf{k}}^{[2^k-1]})\, \phi(f_{\mathbf{k}}) \\
&= \int_{\mathcal{H}^k} df_{\mathbf{k}} \sum_{j=0}^{2^k-1} \Big(Q^k(f_{\mathbf{k}}^{[j]}) \Big)\; \phi(f_{\mathbf{k}}) \\
&= \int_{\mathcal{H}^k} df_{\mathbf{k}} \sum_{j=0}^{2^k-1} \left(\prod_{i=1}^{k} \left[Q(f_{\mathbf{k}_i}^{(s_i^j)}) \right] \right) \phi(f_{\mathbf{k}}) && (47) \\
&= \int_{\mathcal{H}^k} df_{\mathbf{k}} \prod_{i=1}^{k} \left[Q(f_{\mathbf{k}_i}^{(0)}) + Q(f_{\mathbf{k}_i}^{(1)}) \right] \phi(f_{\mathbf{k}}) && (48) \\
&= \int_{\mathcal{H}^k} df_{\mathbf{k}} \prod_{i=1}^{k} [Q(f_{\mathbf{k}_i}) + Q(c(f_{\mathbf{k}_i}))]\; \phi(f_{\mathbf{k}}) \\
&= \int_{\mathcal{H}^k} df_{\mathbf{k}} \prod_{i=1}^{k} [P(f_{\mathbf{k}_i}) + P(c(f_{\mathbf{k}_i}))]\; \phi(f_{\mathbf{k}}) \\
& \vdots \\
&= \; 2^k \cdot \mathop{\mathbf{E}}_{f_{\mathbf{k}} \sim P^k} \phi(f_{\mathbf{k}})\,,
\end{aligned}
$$

where we obtain Line (48) from Line (47) by developing the terms of the product of Line (48).

The result is obtained by changing the expectation over Q^k to an expectation over P^k, and then by applying Jensen's inequality (Lemma 47, in Appendix A).

$$
\mathop{\mathbf{E}}_{f_{\mathbf{k}} \sim Q^k} \phi(f_{\mathbf{k}}) \;=\; \mathop{\mathbf{E}}_{f_{\mathbf{k}} \sim P^k} \phi(f_{\mathbf{k}}) \;=\; \mathop{\mathbf{E}}_{f_{\mathbf{k}} \sim P^k} \ln e^{\phi(f_{\mathbf{k}})} \;\leq\; \ln \left(\mathop{\mathbf{E}}_{f_{\mathbf{k}} \sim P^k} e^{\phi(f_{\mathbf{k}})} \right) .
$$

■

Theorem 56 (General PAC-Bayesian theorem for tuples of voters and aligned posteriors) *For any distribution D on $\mathcal{X}\times\mathcal{Y}$, any self-complemented set $\mathcal{H}$ of voters $\mathcal{X}\to\overline{\mathcal{Y}}$, any prior distribution P on $\mathcal{H}$, any integer $k\geq 1$, any convex function $\mathcal{D}:[0,1]\times[0,1]\to\mathbb{R}$ and loss function $\mathcal{L}:\overline{\mathcal{Y}}^k\times\mathcal{Y}^k\to[0,1]$ for which $\mathcal{D}\left(\mathbb{E}^{\mathcal{L}}_S(f^{[j]}_{\mathbf{k}}),\,\mathbb{E}^{\mathcal{L}}_D(f^{[j]}_{\mathbf{k}})\right)=\mathcal{D}\left(\mathbb{E}^{\mathcal{L}}_S(f^{[j']}_{\mathbf{k}}),\,\mathbb{E}^{\mathcal{L}}_D(f^{[j']}_{\mathbf{k}})\right)$, for any $j,j'\in\{0,\ldots,2^k-1\}$, for any $m'>0$ and any $\delta\in(0,1]$, we have*

$$\Pr_{S\sim D^m}\left(\begin{array}{l}\textit{For all posteriors } Q \textit{ aligned on } P:\\ \mathcal{D}\left(\mathop{\mathbf{E}}_{f_{\mathbf{k}}\sim Q^k}\mathbb{E}^{\mathcal{L}}_S(f_{\mathbf{k}}),\mathop{\mathbf{E}}_{f_{\mathbf{k}}\sim Q^k}\mathbb{E}^{\mathcal{L}}_D(f_{\mathbf{k}})\right)\leq\dfrac{1}{m'}\left[\ln\left(\dfrac{1}{\delta}\mathop{\mathbf{E}}_{S\sim D^m}\mathop{\mathbf{E}}_{f_{\mathbf{k}}\sim P^k}e^{m'\cdot\mathcal{D}\left(\mathbb{E}^{\mathcal{L}}_S(f_{\mathbf{k}}),\mathbb{E}^{\mathcal{L}}_D(f_{\mathbf{k}})\right)}\right)\right]\end{array}\right)\geq 1-\delta\,.$$

Proof This proof follows most of the steps of Theorem 18.
We have that $\mathop{\mathbf{E}}_{f_{\mathbf{k}}\sim P^k}e^{m'\cdot\mathcal{D}\left(\mathbb{E}^{\mathcal{L}}_S(f_{\mathbf{k}}),\mathbb{E}^{\mathcal{L}}_D(f_{\mathbf{k}})\right)}$ is a non-negative random variable. By Markov's inequality, we have

$$\Pr_{S\sim D^m}\left(\mathop{\mathbf{E}}_{f_{\mathbf{k}}\sim P^k}e^{m'\cdot\mathcal{D}\left(\mathbb{E}^{\mathcal{L}}_S(f_{\mathbf{k}}),\mathbb{E}^{\mathcal{L}}_D(f_{\mathbf{k}})\right)}\leq\frac{1}{\delta}\mathop{\mathbf{E}}_{S\sim D^m}\mathop{\mathbf{E}}_{f_{\mathbf{k}}\sim P^k}e^{m'\cdot\mathcal{D}\left(\mathbb{E}^{\mathcal{L}}_S(f_{\mathbf{k}}),\mathbb{E}^{\mathcal{L}}_D(f_{\mathbf{k}})\right)}\right)\geq 1-\delta\,.$$

Hence, by taking the logarithm on each side of the innermost inequality, we obtain

$$\Pr_{S\sim D^m}\left(\ln\left[\mathop{\mathbf{E}}_{f_{\mathbf{k}}\sim P^k}e^{m'\cdot\mathcal{D}\left(\mathbb{E}^{\mathcal{L}}_S(f_{\mathbf{k}}),\mathbb{E}^{\mathcal{L}}_D(f_{\mathbf{k}})\right)}\right]\leq\ln\left[\frac{1}{\delta}\mathop{\mathbf{E}}_{S\sim D^m}\mathop{\mathbf{E}}_{f_{\mathbf{k}}\sim P^k}e^{m'\cdot\mathcal{D}\left(\mathbb{E}^{\mathcal{L}}_S(f_{\mathbf{k}}),\mathbb{E}^{\mathcal{L}}_D(f_{\mathbf{k}})\right)}\right]\right)\geq 1-\delta\,.$$

Now, instead of using the change of measure inequality of Lemma 17, we use the change of measure inequality of Theorem 55 on the left side of innermost inequality, with $\phi(f_{\mathbf{k}})=m'\cdot\mathcal{D}\left(\mathbb{E}^{\mathcal{L}}_S(f_{\mathbf{k}}),\,\mathbb{E}^{\mathcal{L}}_D(f_{\mathbf{k}})\right)$. We then use Jensen's inequality (Lemma 47, in Appendix A), exploiting the convexity of $\mathcal{D}$.

$$\begin{aligned}\forall\, Q \text{ aligned on } P:\ \ln\left[\mathop{\mathbf{E}}_{f_{\mathbf{k}}\sim P^k}e^{m'\cdot\mathcal{D}(\mathbb{E}^{\mathcal{L}}_S(f_{\mathbf{k}}),\mathbb{E}^{\mathcal{L}}_D(f_{\mathbf{k}}))}\right]\ &\geq\ m'\cdot\mathop{\mathbf{E}}_{f_{\mathbf{k}}\sim Q^k}\mathcal{D}(\mathbb{E}^{\mathcal{L}}_S(f_{\mathbf{k}}),\mathbb{E}^{\mathcal{L}}_D(f_{\mathbf{k}}))\\ &\geq\ m'\cdot\mathcal{D}(\mathop{\mathbf{E}}_{f_{\mathbf{k}}\sim Q^k}\mathbb{E}^{\mathcal{L}}_S(f_{\mathbf{k}}),\mathop{\mathbf{E}}_{f_{\mathbf{k}}\sim Q^k}\mathbb{E}^{\mathcal{L}}_D(f_{\mathbf{k}}))\,.\end{aligned}$$

We therefore have

$$\Pr_{S\sim D^m}\left(\begin{array}{l}\text{For all posteriors } Q \text{ aligned on } P:\\ m'\cdot\mathcal{D}(\mathop{\mathbf{E}}_{f_{\mathbf{k}}\sim Q^k}\mathbb{E}^{\mathcal{L}}_S(f_{\mathbf{k}}),\mathop{\mathbf{E}}_{f_{\mathbf{k}}\sim Q^k}\mathbb{E}^{\mathcal{L}}_D(f_{\mathbf{k}}))\leq\ln\left[\frac{1}{\delta}\mathop{\mathbf{E}}_{S\sim D^m}\mathop{\mathbf{E}}_{f_{\mathbf{k}}\sim P^k}e^{m'\cdot\mathcal{D}(\mathbb{E}^{\mathcal{L}}_S(f_{\mathbf{k}}),\mathbb{E}^{\mathcal{L}}_D(f_{\mathbf{k}}))}\right]\end{array}\right)\geq 1-\delta\,.$$

The result then follows from easy calculations. ■

References

Arindam Banerjee. On Bayesian bounds. In *ICML*, pages 81–88, 2006.

Luc Bégin, Pascal Germain, François Laviolette, and Jean-Francis Roy. PAC-Bayesian theory for transductive learning. In *AISTATS*, pages 105–113, 2014.

C.L. Blake and C.J. Merz. *UCI Repository of Machine Learning Databases*. Department of Information and Computer Science, Irvine, CA: University of California, http://www.ics.uci.edu/∼mlearn/MLRepository.html, 1998.

John Blitzer, Mark Dredze, and Fernando Pereira. Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In *Annual Meeting-Association for Computational Linguistics*, volume 45, page 440, 2007.

Leo Breiman. Bagging predictors. *Machine Learning*, 24(2):123–140, 1996.

Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

Olivier Catoni. *PAC-Bayesian Supervised Classification: The Thermodynamics of Statistical Learning*. Monograph series of the Institute of Mathematical Statistics, http://arxiv.org/abs/0712.0248, 2007.

Minmin Chen, Kilian Q. Weinberger, and John Blitzer. Co-training for domain adaptation. In *NIPS*, pages 2456–2464, 2011.

Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine Learning*, 20(3): 273–297, 1995.

Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*, chapter 12. Wiley, 1991.

Nello Cristianini and John Shawe-Taylor. *An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods*. Cambridge University Press, Cambridge, U.K., 2000.

Joachim Dahl and Lieven Vandenberghe. CVXOPT, 2007. `http://mloss.org/software/view/34/`.

Amit Daniely, Sivan Sabato, Shai Ben-David, and Shai Shalev-Shwartz. Multiclass learnability and the ERM principle. *CoRR*, abs/1308.2893, 2013.

Philip Derbeko, Ran El-Yaniv, and Ron Meir. Explicit learning curves for transduction and application to clustering and compression algorithms. *J. Artif. Intell. Res. (JAIR)*, 22: 117–142, 2004.

Mark Dredze, Alex Kulesza, and Koby Crammer. Multi-domain learning by confidence-weighted parameter combination. *Machine Learning*, 79(1-2):123–149, 2010.

Mahdi Milani Fard and Joelle Pineau. PAC-Bayesian model selection for reinforcement learning. In *NIPS*, pages 1624–1632, 2010.

Mahdi Milani Fard, Joelle Pineau, and Csaba Szepesvári. PAC-Bayesian policy evaluation for reinforcement learning. In *UAI*, pages 195–202, 2011.

Sally Floyd and Manfred Warmuth. Sample compression, learnability, and the Vapnik-Chervonenkis dimension. *Machine Learning*, 21(3):269–304, 1995.

A. Gelman, J.B. Carlin, H.S. Stern, and D.B. Rubin. *Bayesian Data Analysis*. Chapman & Hall/CRC, 2004. ISBN 9781584883883.

Pascal Germain, Alexandre Lacasse, François Laviolette, and Mario Marchand. PAC-Bayesian learning of linear classifiers. In *ICML*, page 45, 2009.

Pascal Germain, Alexandre Lacoste, François Laviolette, Mario Marchand, and Sara Shanian. A PAC-Bayes sample-compression approach to kernel methods. In *ICML*, pages 297–304, 2011.

Pascal Germain, Amaury Habrard, François Laviolette, and Emilie Morvant. A PAC-Bayesian approach for domain adaptation with specialization to linear classifiers. In *ICML (3)*, pages 738–746, 2013.

Sébastien Giguère, François Laviolette, Mario Marchand, and Khadidja Sylla. Risk bounds and learning algorithms for the regression approach to structured output prediction. In *ICML (1)*, pages 107–114, 2013.

Matthew Higgs and John Shawe-Taylor. A PAC-Bayes bound for tailored density estimation. In *ALT*, pages 148–162, 2010.

Alexandre Lacasse, François Laviolette, Mario Marchand, Pascal Germain, and Nicolas Usunier. PAC-Bayes bounds for the risk of the majority vote and the variance of the Gibbs classifier. In *NIPS*, pages 769–776, 2006.

Alexandre Lacoste, François Laviolette, and Mario Marchand. Bayesian comparison of machine learning algorithms on single and multiple datasets. In *AISTATS*, pages 665–675, 2012.

John Langford. Tutorial on practical prediction theory for classification. *Journal of Machine Learning Research*, 6:273–306, 2005.

John Langford and Matthias Seeger. Bounds for averaging classifiers. Technical report, Carnegie Mellon, Departement of Computer Science, 2001.

John Langford and John Shawe-Taylor. PAC-Bayes & margins. In *NIPS*, pages 423–430, 2002.

François Laviolette and Mario Marchand. PAC-Bayes risk bounds for sample-compressed Gibbs classifiers. In *ICML*, pages 481–488, 2005.

François Laviolette and Mario Marchand. PAC-Bayes risk bounds for stochastic averages and majority votes of sample-compressed classifiers. *Journal of Machine Learning Research*, 8:1461–1487, 2007.

François Laviolette, Mario Marchand, and Jean-Francis Roy. From PAC-Bayes bounds to quadratic programs for majority votes. In *ICML*, pages 649–656, 2011.

Yann Lecun and Corinna Cortes. The MNIST database of handwritten digits. URL `http://yann.lecun.com/exdb/mnist/`.

Guy Lever, François Laviolette, and John Shawe-Taylor. Distribution-dependent PAC-Bayes priors. In *ALT*, pages 119–133, 2010.

Guy Lever, François Laviolette, and John Shawe-Taylor. Tighter PAC-Bayes bounds through distribution-dependent priors. *Theoretical Computer Science*, 473:4–28, 2013.

Ben London, Bert Huang, Benjamin Taskar, and Lise Getoor. PAC-Bayesian collective stability. In *AISTATS*, pages 585–594, 2014.

Andreas Maurer. A note on the PAC-Bayesian theorem. *CoRR*, cs.LG/0411099, 2004.

David McAllester. Some PAC-Bayesian theorems. *Machine Learning*, 37(3):355–363, 1999.

David McAllester. PAC-Bayesian stochastic model selection. *Machine Learning*, 51(1):5–21, 2003a.

David McAllester. Simplified PAC-Bayesian margin bounds. In *COLT*, pages 203–215, 2003b.

David McAllester. Generalization bounds and consistency for structured labeling. In Gökhan Bakır, Thomas Hofmann, Bernhard Schölkopf, Alexander J. Smola, Ben Taskar, and S. V. N. Vishwanathan, editors, *Predicting Structured Data*, chapter 11, pages 247–261. MIT Press, Cambridge, MA, 2007.

David McAllester. A PAC-Bayesian tutorial with a dropout bound. *CoRR*, abs/1307.2118, 2013.

W. Mendenhall. Nonparametric statistics. *Introduction to Probability and Statistics*, 604, 1983.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine Learning in Python . *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Liva Ralaivola, Marie Szafranski, and Guillaume Stempfel. Chromatic PAC-Bayes bounds for non-iid data: Applications to ranking and stationary β-mixing processes. *Journal of Machine Learning Research*, 11:1927–1956, 2010.

Gerard Salton and Christopher Buckley. Term-weighting approaches in automatic text retrieval. *Information Processing & Management*, 24(5):513–523, 1988.

Robert E. Schapire and Yoram Singer. Improved boosting using confidence-rated predictions. *Machine Learning*, 37(3):297–336, 1999.

Robert E. Schapire, Yoav Freund, Peter Bartlett, and Wee Sun Lee. Boosting the margin: A new explanation for the effectiveness of voting methods. *The Annals of Statistics*, 26: 1651–1686, 1998.

Bernhard Schölkopf, Ralf Herbrich, and Alex J. Smola. A generalized representer theorem. In *COLT/EuroCOLT*, pages 416–426, 2001.

Matthias Seeger. PAC-Bayesian generalization bounds for Gaussian processes. *Journal of Machine Learning Research*, 3:233–269, 2002.

Matthias Seeger. *Bayesian Gaussian Process Models: PAC-Bayesian Generalisation Error Bounds and Sparse Approximations.* PhD thesis, University of Edinburgh, 2003.

Yevgeny Seldin and Naftali Tishby. PAC-Bayesian generalization bound for density estimation with application to co-clustering. In *AISTATS*, pages 472–479, 2009.

Yevgeny Seldin and Naftali Tishby. PAC-Bayesian analysis of co-clustering and beyond. *Journal of Machine Learning Research*, 11:3595–3646, 2010.

Yevgeny Seldin, Peter Auer, François Laviolette, John Shawe-Taylor, and Ronald Ortner. PAC-Bayesian analysis of contextual bandits. In *NIPS*, pages 1683–1691, 2011.

Yevgeny Seldin, François Laviolette, Nicolò Cesa-Bianchi, John Shawe-Taylor, and Peter Auer. PAC-Bayesian inequalities for martingales. *IEEE Transactions on Information Theory*, 58(12):7086–7093, 2012.

Chunhua Shen and Hanxi Li. Boosting through optimization of margin distributions. *IEEE Transactions on Neural Networks*, 21(4):659–666, 2010.

Ilya O. Tolstikhin and Yevgeny Seldin. PAC-Bayes-empirical-bernstein inequality. In *NIPS*, pages 109–117, 2013.

Malik Younsi. Proof of a combinatorial conjecture coming from the PAC-Bayesian machine learning theory. *ArXiv E-Prints*, 2012. URL `http://arxiv.org/abs/1209.0824v1`.