Risk upper bounds for general ensemble methods with an application to multiclass classification

François Laviolette, Emilie Morvant, Liva Ralaivola, Jean-Francis Roy

Risk Upper Bounds for General Ensemble Methods with an application to Multiclass Classification

François Laviolette[1] Emilie Morvant[2] Liva Ralaivola[3] Jean-Francis Roy[1]
[1] Département d'informatique et de génie logiciel, Université Laval, Québec, Canada
[2] Laboratoire Hubert Curien, Université Jean Monnet, UMR CNRS 5516, Saint-Etienne, France
[3] Aix-Marseille Univ., LIF-QARMA, CNRS, UMR 7279, F-13013, Marseille, France

June 4, 2018

This work is published as:
Franois Laviolette, Emilie Morvant, Liva Ralaivola, Jean-Francis Roy. **Risk upper bounds for general ensemble methods with an application to multiclass classification.** Neurocomputing, Elsevier, 2017, 219, pp.15–25.
Link to the published version:
https://www.sciencedirect.com/science/article/pii/S0925231216310177

Abstract

This paper generalizes a pivotal result from the PAC-Bayesian literature —the $\mathcal{C}$-bound— primarily designed for binary classification to the general case of ensemble methods of voters with arbitrary outputs. We provide a generic version of the $\mathcal{C}$-bound, an upper bound over the risk of models expressed as a weighted majority vote that is based on the first and second statistical moments of the vote's *margin*. On the one hand, this bound may advantageously be applied on more complex outputs than mere binary outputs, such as multiclass labels and multilabel, and on the other hand, it allows us to consider margin relaxations. We provide a specialization of the bound to multiclass classification together with empirical evidence that the presented theoretical result is tightly bound to the risk of the majority vote classifier. We also give insights as to how the proposed bound may be of use to characterize the risk of multilabel predictors.

1 Introduction

Complex Output Prediction It is well known that learning predictive models capable of dealing with outputs richer than binary outputs (*e.g.,* multiclass or multilabel) and for which theoretical guarantees exist is still a realm of intensive investigations. From a practical standpoint, a lot of relaxations for learning with complex outputs have been devised. A common approach consists in decomposing the output space into "simpler" spaces so that the learning problem at hand is reduced to a few easier (*i.e.,* binary) learning tasks. For instance, this is the idea spurred by the Error-Correcting Output Codes (Dietterich and Bakiri, 1995) that makes it possible to reduce multiclass or multilabel problems into binary classification tasks (*e.g.*, Allwein et al. (2001); Mroueh et al. (2012); Read et al. (2011); Tsoumakas and Vlahavas (2007); Zhang and Schneider (2012)). In our work, we study the problem of complex output prediction by focusing on prediction functions that take the form of a weighted majority vote over a set of complex output classifiers that we call *voters.*

Majority Vote Predictors Studying majority vote predictors actually allows us to provide results that are applicable to a wide range of classification methods. For instance, *ensemble methods*—of which Bagging (Breiman, 1996), Boosting (Schapire and Singer, 1999) and Random Forests (Breiman, 2001) are representative—can all be seen as majority vote learning procedures (Dietterich, 2000; Re and Valentini, 2012); from this standpoint, our work is also related to that of Cortes et al. (2014), who have proposed various ensemble methods for the structured output prediction problem. Majority votes are also central to the Bayesian approach (Gelman et al., 2004) through the notion of Bayesian model averaging (Domingos, 2000; Haussler et al., 1994). Also, most if not all kernel-based predictors, such as the Support Vector Machines (Boser et al., 1992; Cortes and Vapnik, 1995) may be viewed as weighted majority votes as well:

for a kernel classifier built from training set $\{(x_i, y_i)\}$ and kernel function k, the predicted class for some input x is usually computed as the sign of $\sum_i \alpha_i\, y_i\, k(x_i, x)$, each voter is simply given by $x \mapsto y_i k(x_i, x)$.

PAC-Bayesian Analysis of the Risk From a theoretical perspective, as far as binary classification is concerned, the notion of *margin* is often the crux to establish the generalization ability of a majority vote predictor; the margin of a majority vote realized on an example is then defined as the difference between the total weight of the voters that predicted the correct class minus the total weight given to the incorrect one. In the PAC-Bayesian framework, which is our working setup, one way to provide generalization bounds for a majority vote classifier is to relate it to a stochastic classifier, the *Gibbs* classifier, whose risk is the weighted risk of the individual voters involved in the majority vote. Up to a linear transformation, the Gibbs risk is equivalent to the first statistical moment of the margin (Laviolette et al., 2011; Germain et al., 2015). Folk PAC-Bayesian results can be very accurate when the Gibbs risk is low, as in the situation where the voters having large weights are performing well (Germain et al., 2009; Langford and Shawe-Taylor, 2002; McAllester, 2009). However, for general ensemble methods, it is not unusual to be in the situation where, on the one hand, the voters achieve performances only slightly above the chance level—which makes it impossible to find weights that induce a small Gibbs risk—and, on the other hand, the risk of the majority vote itself is very low. Hence, to better capture the accuracy of a majority vote in a PAC-Bayesian fashion, it is required to consider more than the Gibbs risk, *i.e.*, more than only the first statistical moment of the margin. This idea, which has been investigated in the context of ensemble methods by Blanchard (2004) and Breiman (2001), has been revisited as the $\mathcal{C}$-bound by Lacasse et al. (2007) in the PAC-Bayesian framework. This bound sheds light on an essential feature of weighted majority votes: how good the voters individually are is just as important as how correlated their predictions are; this has inspired a new ensemble method for binary classification with PAC-Bayesian generalization guarantees named MinCq (Laviolette et al., 2011), whose performances are on par with the most advanced binary classification methods. In the multiclass setting, there exists one PAC-Bayesian bound, which is based on the confusion matrix of the Gibbs classifier (Morvant et al., 2012). Kuznetsov et al. (2014) have proposed an improved Rademacher bound for multiclass prediction that is based on the notion of the multiclass margin of Breiman (2001) (Definition 1 in the present paper). However, as for the binary case, these bounds suffer from the same lack of tightness when the voters of the majority vote perform poorly.

Contributions Here, we generalize the $\mathcal{C}$-bound to more complex situations than mere binary classification. We first propose a formulation of the $\mathcal{C}$-bound for ensemble methods in complex output settings. To do so, we start from complex output predictors, with the objective to build a majority vote predictor out of these (as, *e.g.*, in Cortes et al. (2014); Li et al. (2013)). This new formulation makes it possible to generalize all the classification-based results of Lacasse et al. (2007), Laviolette et al. (2011) and Germain et al. (2015). Since for complex output prediction the usual margin relies on the deviation between the total weight given to the correct output minus the maximal total weight given to the "runner-up" incorrect one, we base our theory on a notion of margin that is a relevant extension of the usual (binary) margin. As for binary classification (Lacasse et al., 2007; Laviolette et al., 2011; Germain et al., 2015), we derive a PAC-Bayesian generalization bound and show how we can estimate such $\mathcal{C}$-bounds from a sample. Starting from this general theoretical result, we propose specializations suitable for multiclass classification with ensemble methods based on the true margin and on a relaxation that we call ω-margin. We report and analyze the behavior of these $\mathcal{C}$-bounds through an empirical study.

Organization of the paper The rest of the paper is organized as follows. Section 2 recalls the binary $\mathcal{C}$-bound, which is generalized to a more general setting in Section 3. We then specialize this bound to multiclass prediction and provide empirical results in Section 4.We conclude in Section 5 and we provide an insight as to how our results might readily be of use to address the multi-label classification.

2 Ensemble Methods in Binary Classification

For binary classification with majority vote-based ensemble methods, we often consider an arbitrary input space $\mathcal{X}$, an output space $\mathcal{Y} = \{-1, +1\}$ made of two classes, and a set $\mathcal{H} \subseteq \{h : \mathcal{X} \to [-1, +1]\}$ of *voters*. We consider the possibility that a voter outputs any value in $[-1, +1]$, interpretable as a level of confidence of the voter into the predicted label, which is $+1$ if the output is positive and -1 otherwise. A voter that always outputs values in $\{-1, +1\}$ is called a *(binary) classifier*. The *binary ρ-weighted majority vote*

$B_\rho(\cdot)$ is the classifier returning either of the two options that has obtained the larger weight in the vote, *i.e.*,

$$\forall x \in \mathcal{X}, \quad B_\rho(x) \doteq \underset{y\in\{-1,+1\}}{\operatorname{argmax}} \underset{h\sim\rho}{\mathbf{E}} \left(\Big|h(x)\Big| \, \mathrm{I}\Big[\operatorname{sign}\big(h(x)\big) = y\Big] \right)$$
$$= \operatorname{sign}\left[\underset{h\sim\rho}{\mathbf{E}} h(x) \right],$$

where $\mathrm{I}[a] = 1$ if predicate a is true and 0 otherwise.

Given a training set $S = \{(x_i, y_i)\}_{i=1}^m$ of observed data in which each example $(x_i, y_i) \in S$ is independently and identically distributed (*i.i.d.*) according to a fixed yet unknown probability distribution D over $\mathcal{X} \times \{-1, +1\}$, the learner aims at finding a weigthing distribution ρ on $\mathcal{H}$ inducing a low-error majority vote. In other words, minimizing the *true risk*

$$R_D(B_\rho) \doteq \underset{(x,y)\sim D}{\mathbf{E}} \mathrm{I}\left[B_\rho(x) \neq y\right]$$

of the ρ-weighted majority vote under the 0-1-loss is aimed at. One route towards this goal is to implement the Empirical Risk Minimization (ERM) principle which consists in minimizing the *empirical risk*

$$R_S(B_\rho) \doteq \frac{1}{m} \sum_{i=1}^{m} \mathrm{I}\left[B_\rho(x_i) \neq y_i\right]$$

of the majority vote computed on the training set S. However, a well-known issue to learn such weights is that the direct minimization of $R_S(B_\rho)$ is an $\mathcal{NP}$-hard problem. In addition, there are necessary conditions for the ERM approach to have consistency guarantees; with these conditions failing to be met learning may be prone to overfitting. To overcome these two issues, we may use relaxations of the risk, look for estimators or bounds of the true risk that are simultaneously valid for all possible distributions ρ on $\mathcal{H}$, and try to minimize them. In the PAC-Bayesian theory[1], such an estimator is given by the *Gibbs risk*

$$R_D^{\mathrm{Gibbs}}(G_\rho) \doteq \underset{h\sim\rho}{\mathbf{E}} R_D(h)$$

of a ρ-weighted majority vote which is simply the ρ-average risk of the voters. Indeed, it is well known (see, e.g., (McAllester, 1999)) that the risk of the ρ-weighted majority vote $B_\rho(\cdot)$ is bounded by twice its Gibbs risk:

$$R_D(B_\rho) \;\leq\; 2\, R_D(G_\rho). \tag{1}$$

With this relation, the PAC-Bayesian theory indirectly gives generalization bounds for ρ-weighted majority votes. Unfortunately, even if they tightly bound the true risk $R_D^{\mathrm{Gibbs}}(G_\rho)$ in terms of its empirical counterpart

$$R_S^{\mathrm{Gibbs}}(G_\rho) \doteq \underset{h\sim\rho}{\mathbf{E}} R_S(h),$$

this tightness might not carry over to the bound on the majority vote. Indeed, even if there exist situations for which Inequality (1) is an equality, ensemble methods (especially when the voters are 'weak') build on the idea that the risk of the majority vote might be way below the average of its voters' risk. Indeed, it is notorious that voting may dramatically improve performances when the community of voters tends to compensate the individual errors. The "classical" PAC-Bayesian framework (McAllester, 1999) does not make it possible to evaluate whether or not this compensation occurs. To overcome this problem, Lacasse et al. (2007) proposed not only to take into account the mean of the errors of the associated Gibbs predictor G_ρ, but also its variance, and they proposed a new bound, called the *$\mathcal{C}$-bound*, that replaces the loose factor of 2 in Inequality (1). They also extended the PAC-Bayesian theory in such a way that both the mean and the variance of the Gibbs classifier can be estimated from the training data simultaneously for all ρ's. Laviolette et al. (2011) have reformulated this approach in terms of the first and second statistical moment of the *margin* realized by the ρ-weighted majority vote, where the margin $M_\rho(x, y)$ of a ρ-weighted majority vote on an example $(x, y) \in \mathcal{X} \times \mathcal{Y}$ is

$$M_\rho(x, y) \doteq y \underset{h\sim\rho}{\mathbf{E}} h(x).$$

The proposed result pertains to known results in non-PAC-Bayesian frameworks (*e.g.,* (Blanchard, 2004; Breiman, 2001)).

In terms of margin $M_\rho(x, y)$, the $\mathcal{C}$-bound is defined as follows.

[1] The PAC-Bayesian theory was first introduced by McAllester (1999).

Theorem 1 ($\mathcal{C}$-bound of Laviolette et al. (2011); Germain et al. (2015))**.** *For every distribution ρ on a set of voters $\mathcal{H}$, and for every distribution D on $\mathcal{X} \times \mathcal{Y}$, if $\mathbf{E}_{(x,y)\sim D}\, y\, \mathbf{E}_{h\sim\rho}\, h(x) > 0$, then we have:*

$$R_D(B_\rho) \;\leq\; 1 - \frac{\left(\mathop{\mathbf{E}}_{(x,y)\sim D} y \mathop{\mathbf{E}}_{h\sim\rho} h(x)\right)^2}{\mathop{\mathbf{E}}_{(x,y)\sim D}\left(y \mathop{\mathbf{E}}_{h\sim\rho} h(x)\right)^2} = 1 - \frac{\left(\mathop{\mathbf{E}}_{(x,y)\sim D} M_\rho(x,y)\right)^2}{\mathop{\mathbf{E}}_{(x,y)\sim D} M_\rho^2(x,y)}.$$

Proof. First, note that

$$R_D(B_\rho) = \mathop{\mathbf{Pr}}_{(x,y)\sim D}\left(M_\rho(x,y) \leq 0\right),$$

then to upper bound $R_D(B_\rho)$, it suffices to upper bound $\mathbf{Pr}_{(x,y)\sim D}\left(M_\rho(x,y) \leq 0\right)$. Making use of the Cantelli-Chebyshev inequality stating that for any random variable Z,

$$\forall a > 0, \quad \mathbf{Pr}\left(Z \leq \mathbf{E}\left[Z\right] - a\right) \;\leq\; \frac{\mathbf{Var}\, Z}{\mathbf{Var}\, Z + a^2},$$

we get the desired result if $Z = M_\rho(x,y)$, with $a = \mathbf{E}_{(x,y)\sim D}\, M_\rho(x,y)$ combined with the definition of the variance. The constraint $\mathbf{E}_{(x,y)\ \sim D}\, y\, \mathbf{E}_{h\sim\rho}\, h(x) > 0$ comes from this inequality being valid when $a > 0$. □

The $\mathcal{C}$-bound involves both the ρ-weighted majority vote confidence via $\mathbf{E}_{(x,y)}(y\, \mathbf{E}_{h\sim\rho}\, h(x)) = \mathbf{E}_{(x,y)}\, M_\rho(x,y)$ and the average correlation between the voters via $\mathbf{E}_{(x,y)}(y\, \mathbf{E}_{h\sim\rho}\, h(x))(y\, \mathbf{E}_{h'\sim\rho}\, h'(x)) = \mathbf{E}_{(x,y)}\, M_\rho^2(x,y)$. Minimizing its empirical counterpart appears as a natural solution for learning a distribution ρ leading to a well-performing binary ρ-weighted majority vote. Moreover, this strategy is justified by a PAC-Bayesian generalization bound over the $\mathcal{C}$-bound (similar to Theorem 3 of the present paper but restricted to the case where $\mathcal{Y} = \{-1, +1\}$), and has given the MinCq algorithm (Laviolette et al., 2011; Germain et al., 2015).

As announced earlier, we here intend to generalize the $\mathcal{C}$-bound theory to more complex outputs than binary outputs. Our contributions first consist in generalizing—in Section 3—this important result to a broader ensemble method setting, along with PAC-Bayesian generalization bounds.

3 A General Setting for Majority Votes over a Set of Complex Output Voters

In this section, we propose a general setting in which one can consider predicting with ρ-weighted majority votes. We present a general definition of the margin and propose a $\mathcal{C}$-bound designed for majority vote-based ensemble methods when one wants to combine complex output predictors (or experts). We recall that these predictors are assumed to be generated *a priori* and thus are treated as black boxes. We also discuss how to estimate this bound from a set S of m examples drawn *i.i.d.* from D. To do so, we derive a PAC-Bayesian theorem that bounds the *true* risk $R_D(B_\rho)$of the ρ-weighted majority vote B_ρ by using the empirical estimation of our new $\mathcal{C}$-bound on the training sample S.

3.1 A General $\mathcal{C}$-bound for Complex Output Prediction

Given some input space $\mathcal{X}$ and a *finite* output space $\mathcal{Y}$, we suppose that there exists a feature map $\mathbf{Y} : \mathcal{Y} \to \mathbf{H}_\mathcal{Y}$, where $\mathbf{H}_\mathcal{Y}$ is a vector space such, e.g., a Hilbert space. For the sake of clarity, we suppose that all the vectors of $\mathrm{Im}\ \mathcal{Y} \doteq \mathbf{Y}(\mathcal{Y})$, the image of $\mathcal{Y}$ under $\mathbf{Y}(\cdot)$, are unit-norm vectors; most of the following results remain true without this assumption but have to be stated in a more complicated form. Let $\mathrm{conv}(\mathrm{Im}\ \mathcal{Y})$ ($\subseteq \mathbf{H}_\mathcal{Y}$) denote the convex hull of $\mathrm{Im}\ \mathcal{Y}$. We consider a (non-necessarily finite) set of *voters* $\boldsymbol{\mathcal{H}} \subseteq \{\mathbf{h} : \mathcal{X} \to \mathrm{conv}(\mathrm{Im}\ \mathcal{Y})\}$; we use the **bold** notation to distinguish hypotheses and functions that output vector values from the real-valued hypotheses and functions considered in the binary case.

Remark 1. *Let us point out that assuming the existence of a feature map $\mathbf{Y} : \mathcal{Y} \to \mathbf{H}_\mathcal{Y}$ is frequent in kernel methods (see, e.g., Cortes et al. (2007); Brouard et al. (2011); Giguere et al. (2014)). Indeed, there is always such a feature map if one considers an output kernel $k_\mathcal{Y} : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$. Moreover, considering that the vectors in $\mathbf{Y}(\mathcal{Y})$ are unit-norm is equivalent to assuming that such output kernels $k_\mathcal{Y}(\cdot,\cdot)$ are normalized, which can always be done. It is interesting to remark that the "kernel trick" applies here, hence one can consider the dual form, referring only to the kernel $k_\mathcal{Y}(\cdot,\cdot)$ and never explicitly use the*

feature map $\mathbf{Y}(\cdot)$ *that can be very complicated. Finally, a large variety of kernels exists in the literature. For example, when the output is a string, we have the blended spectrum kernel, the N-gram kernel, the weighted degree, etc. When the output is a graph or a tree, we have the Tanimoto kernel and many other convolution/spectral kernels. See Gärtner (2003) for a survey.*

For every probability distribution ρ on $\boldsymbol{\mathcal{H}}$, we define the *ρ-weighted majority vote classifier* B_ρ such that:

$$\begin{aligned}
\forall x \in \mathcal{X}, \quad B_\rho(x) &\doteq \operatorname*{argmin}_{c\in\mathcal{Y}} \left\| \mathbf{Y}(c) - \mathop{\mathbf{E}}_{\mathbf{h}\sim\rho} \mathbf{h}(x) \right\|^2 \\
&= \operatorname*{argmin}_{c\in\mathcal{Y}} \|\mathbf{Y}(c)\|^2 + \left\| \mathop{\mathbf{E}}_{\mathbf{h}\sim\rho} \mathbf{h}(x) \right\|^2 - 2 \left\langle \mathbf{Y}(c), \mathop{\mathbf{E}}_{\mathbf{h}\sim\rho} \mathbf{h}(x) \right\rangle , \\
&= \operatorname*{argmin}_{c\in\mathcal{Y}} -2 \left\langle \mathbf{Y}(c), \mathop{\mathbf{E}}_{\mathbf{h}\sim\rho} \mathbf{h}(x) \right\rangle , \\
&= \operatorname*{argmax}_{c\in\mathcal{Y}} \left\langle \mathbf{Y}(c), \mathop{\mathbf{E}}_{\mathbf{h}\sim\rho} \mathbf{h}(x) \right\rangle , \qquad (2)
\end{aligned}$$

where the next-to-last equality comes from $\|\mathbf{Y}(c)\| = 1$, $\forall c \in \mathcal{Y}$ and from $\mathbf{E}_{\mathbf{h}\sim\rho}\, \mathbf{h}(x)$ being independent from c. As in the binary classification case, the learning objective in the present framework is to find a distribution ρ that minimizes the *true risk*

$$R_D(B_\rho) \;=\; \mathop{\mathbf{E}}_{(x,y)\sim D} \mathrm{I}\left[B_\rho(x) \neq y\right]$$

of the ρ-weighted majority vote. Inspired by the margin definition of Breiman (2001), we propose the following generalization of the binary margin, which measures the confidence of a prediction as the deviation between the voting weights received by the correct prediction and the largest voting weight received by any incorrect prediction.

Definition 1. *For any example* $(x, y) \in \mathcal{X} \times \mathcal{Y}$ *and any distribution* ρ *on a set of complex output voters* $\boldsymbol{\mathcal{H}}$*, we define the* margin $M_\rho(x, y)$ of the ρ-weighted majority vote *on* (x, y) *as*

$$M_\rho(x,y) \;\doteq\; \left\langle \mathop{\mathbf{E}}_{\mathbf{h}\sim\rho} \mathbf{h}(x)\,,\, \mathbf{Y}(y) \right\rangle - \max_{\substack{c\in\mathcal{Y} \\ c\neq y}} \left\langle \mathop{\mathbf{E}}_{\mathbf{h}\sim\rho} \mathbf{h}(x)\,,\, \mathbf{Y}(c) \right\rangle . \qquad (3)$$

With this definition at hand, it is obvious that the ρ-weighted majority vote errs on (x, y) if and only if the margin realized on (x, y) is negative. Therefore, we have:

$$R_D(B_\rho) \;=\; \mathop{\mathbf{Pr}}_{(x,y)\sim D} \left(M_\rho(x,y) \leq 0\right) . \qquad (4)$$

Remark 2. *One may retrieve the binary notion of majority vote from our general framework in various ways, by considering an appropriate feature map* $\mathbf{Y}(\cdot)$*. See A for more details.*

Using the proof technique of Theorem 1, we arrive at the following general $\mathcal{C}$-bound.

Theorem 2 (General $\mathcal{C}$-bound)**.** *For every probability distribution* ρ *on a set of voters* $\boldsymbol{\mathcal{H}}$ *from* $\mathcal{X}$ *to* $\mathrm{conv}(\mathrm{Im}\ \mathcal{Y})$*, and for every distribution* D *on* $\mathcal{X} \times \mathcal{Y}$*, if* $\mathbf{E}_{(x,y)\sim D}\, M_\rho(x, y) > 0$*, then we have:*

$$R_D(B_\rho) \;\leq\; 1 - \frac{\left(\mathop{\mathbf{E}}_{(x,y)\sim D} M_\rho(x,y) \right)^2}{\mathop{\mathbf{E}}_{(x,y)\sim D} M_\rho^2(x,y)} .$$

Proof. Thanks to Equation (4), the proof consists in bounding $\mathbf{Pr}_{(x,y)} \left(M_\rho(x, y) \leq 0\right)$ with the Cantelli-Chebyshev inequality as done for Theorem 1. □

Remark 3 (On the construction of the set of voters $\boldsymbol{\mathcal{H}}$)**.** *All our results hold for both extreme cases of weak voters, as usual in ensemble methods, and that of more expressive/highly-performing voters. Typical instantiations of the former situation are encountered when making use of a kernel function* $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ *that induces the set of voters* $\boldsymbol{\mathcal{H}} = \{z \mapsto k(x, z)\, \mathbf{Y}(y) \,|\, (x, y) \in S\}$*; the situation also arises when a set*

of structured prediction functions learned with different hyperparameters are considered; as evidenced in Section 4.2 for the multiclass setting, the weak voters may also be decision stumps or (more-or-less shallow) trees. Combining more expressive voters is a situation that may show up as a need to combine voters obtained from a primary mechanism. This is for instance the case in multiview learning (Sun, 2013; Yu et al., 2014) when one want to combine models learned from several data descriptions—note that the binary $\mathcal{C}$-bound has already shown its relevance in such a situation (Morvant et al., 2014).

3.2 A PAC-Bayesian Theorem to Estimate the General $\mathcal{C}$-bound

In this section, we briefly discuss how to estimate the previous bound from a sample S constituted by m examples drawn *i.i.d.* from D. To reach this goal, we derive a PAC-Bayesian theorem that upper-bounds the *true* risk $R_D(B_\rho)$ of the ρ-weighted majority vote by using the empirical estimation of the $\mathcal{C}$-bound of Theorem 2 on the sample S.

Theorem 3. *For any distribution D on $\mathcal{X} \times \mathcal{Y}$, for any set $\boldsymbol{\mathcal{H}}$ of voters from $\mathcal{X}$ to* $\text{conv}(\text{Im}\,\mathcal{Y})$*, for any prior distribution π on $\boldsymbol{\mathcal{H}}$ and any $\delta \in (0,1]$, with a probability at least $1-\delta$ over the choice of the m-sample $S \sim (D)^m$, for every posterior distribution ρ over $\boldsymbol{\mathcal{H}}$, if $\mathbf{E}_{(x,y)\sim D}\, M_\rho(x,y) > 0$, we have:*

$$R_D(B_\rho) \;\leq\; 1 - \frac{\max\left(0, \mu_1^2(S,\pi,\rho,\delta)\right)}{\min\left(1, \mu_2(S,\pi,\rho,\delta)\right)}\,,$$

where

$$\mu_1(S,\pi,\rho,\delta) \doteq \frac{1}{m}\sum_{(x,y)\in S} M_\rho(x,y) - B\sqrt{\frac{2}{m}\left[\text{KL}(\rho\|\pi) + \ln\frac{2\sqrt{m}}{\delta/2}\right]}\,,$$

$$\mu_2(S,\pi,\rho,\delta) \doteq \frac{1}{m}\sum_{(x,y)\in S} M_\rho^2(x,y) + B^2\sqrt{\frac{2}{m}\left[2\,\text{KL}(\rho||\pi) + \ln\frac{2\sqrt{m}}{\delta/2}\right]}\,,$$

and where $B \in (0,2]$ upper-bounds the absolute value of the margin $|M_\rho(x,y)|$, $\forall (x,y)$, and where $\text{KL}(\rho\|\pi) = \mathbf{E}_{\mathbf{h}\sim\rho} \ln \frac{\rho(\mathbf{h})}{\pi(\mathbf{h})}$ *is the Kullback-Leibler divergence between ρ and π.*

Proof. Since we have that $\mathbf{h}(x) \in \text{conv}(\text{Im}\,\mathcal{Y})$, $\forall x \in \mathcal{X}$ and $\|\mathbf{Y}(c)\| = 1$, $\forall c \in \mathcal{Y}$, then $\langle \mathbf{E}_{\mathbf{h}\sim\rho}\, \mathbf{h}(x), \mathbf{Y}(c)\rangle$ takes its value in $[-1,+1]$. It follows from Equation (3) that $B = 2$ is always an upper bound of $|M_\rho(x,y)|$.

The bound is obtained by deriving a PAC-Bayesian lower bound on $\mathbf{E}_{(x,y)\sim D}\, M_\rho(x,y)$ and a PAC-Bayesian upper bound on $\mathbf{E}_{(x,y)\sim D}\,(M_\rho(x,y))^2$. We then use a union bound argument to make these two bounds simultaneously valid, and the result follows from Theorem 2. These two bounds and their respective proof are provided in Appendix B, as Theorems 5 and 6. □

Unlike with classical PAC-Bayesian bounds and especially those provided for structured output prediction by McAllester (2009), our theorem has the advantage to directly upper bound the risk of the ρ-weighted majority vote thanks to the $\mathcal{C}$-bound of Theorem 2. Moreover, it allows us to deal with either the general notion of margin, or surrogate versions thereof, as illustrated in the following.

3.3 A Surrogate for the Margin

The general notion of margin can be hard to exploit in general because of the requirement to compute a max (the margin value of the runner-up) to be evaluated. We propose to define a simpler surrogate of the margin, by replacing the second term in Equation (3) by a threshold ω.

Definition 2 (The ω-margin)**.** *For any example $(x,y) \in \mathcal{X} \times \mathcal{Y}$, for any distribution ρ on $\boldsymbol{\mathcal{H}}$, we define the ω-margin $M_{\rho,\omega}(x,y)$* of the ρ-weighted majority vote *realized on (x,y) as*

$$M_{\rho,\omega}(x,y) \doteq \left\langle \mathop{\mathbf{E}}_{\mathbf{h}\sim\rho} \mathbf{h}(x)\,,\, \mathbf{Y}(y) \right\rangle - \omega\,.$$

Trivially, the ω-margin always upper-bounds the margin when $\omega = -1$. Moreover, since $\forall \mathbf{Y}(y) \in \text{Im}\,\mathcal{Y}, \|\mathbf{Y}(y)\| = 1$, and $\forall x \in \mathcal{X},\ \mathbf{E}_{\mathbf{h}\sim\rho}\, \mathbf{h}(x) \in \text{conv}(\text{Im}\,\mathcal{Y})$, then the ω-margin always lower-bounds the margin when $\omega = 1$. We

will see that in the multiclass setting it is also the case for $\omega = \frac{1}{2}$. When the ω-margin lower-bounds the margin, we can replace it in the $\mathcal{C}$-bound in the following way:

$$\mathcal{C}(M_{\rho,\omega}) \;\doteq\; 1 - \frac{\left(\mathop{\mathbf{E}}_{(x,y)\sim D} M_{\rho,\omega}(x,y)\right)^2}{\mathop{\mathbf{E}}_{(x,y)\sim D} M^2_{\rho,\omega}(x,y)}\,. \tag{5}$$

Indeed, in this situation we have:

$$R_D(B_\rho) \;=\; \Pr_{(x,y)\sim D}\left(M_\rho(x,y) \leq 0\right) \leq \Pr_{(x,y)\sim D}\left(M_{\rho,\omega}(x,y) \leq 0\right)\,.$$

Therefore, the proof process of Theorem 2 applies if $\mathop{\mathbf{E}}_{(x,y)\sim D} M_{\rho,\omega}(x,y) > 0$.

Note that even for values of ω for which $\mathcal{C}(M_{\rho,\omega})$ does not give rise to a valid upper bound of $R_D(B_\rho)$, it remains a value of interest as it still captures the behavior of $R_D(B_\rho)$ simultaneously for many different values of ρ. We provide some evidence about this in Section 4.2.

We now theoretically and empirically illustrate these results by studying multiclass classification from our general $\mathcal{C}$-bound perspective.

4 Specializations of the General $\mathcal{C}$-bound to Multiclass Prediction

4.1 From Multiclass Margins to $\mathcal{C}$-bounds

The input space at hand still is $\mathcal{X}$, but the output space $\mathcal{Y} = \{1, \ldots, k\}$ now is a finite set of classes (or categories) $k \geq 2$. We define the output feature map $\mathbf{Y}(\cdot)$ such that the image of $\mathcal{Y}$ is $\text{Im}\ \mathcal{Y} = \{0,1\}^k$. More precisely, the image of a label $c \in \mathcal{Y}$ under $\mathbf{Y}(\cdot)$ is the canonical k-dimensional vector $(0, \ldots, 1, \ldots 0)^\top$ whose only nonzero entry is a 1 at its c-th position. The set $\boldsymbol{\mathcal{H}}$ is a set of multiclass voters $\mathbf{h}$ from $\mathcal{X}$ to $\text{conv}(\text{Im}\ \mathcal{Y})$. We recall that given a prior distribution π over $\boldsymbol{\mathcal{H}}$ and an *i.i.d.* m-sample S (drawn from D), the goal of the PAC-Bayesian theory is to estimate the prediction ability of the ρ-weighted majority vote $B_\rho(\cdot)$ of Equation (2). In this multiclass setting, since for each class $c \in \mathcal{Y}$ only the c-th coordinate of $\mathbf{Y}(c)$ is equal to 1, the definitions of the majority vote classifier and the margin can respectively be rewritten as:

$$B_\rho(x) \;=\; \mathop{\text{argmax}}_{c\in\mathcal{Y}} \mathop{\mathbf{E}}_{\mathbf{h}\sim\rho} h_c(x)\,,$$

and

$$M_\rho(x,y) \;=\; \mathop{\mathbf{E}}_{\mathbf{h}\sim\rho} h_y(x) - \max_{c\in\mathcal{Y}, c\neq y} \mathop{\mathbf{E}}_{\mathbf{h}\sim\rho} h_c(x)\,,$$

where $h_c(x)$ is the c-th coordinate of $\mathbf{h}(x)$. The following theorem relates the risk of B_ρ and the ω-margin associated to the posterior distribution ρ over $\boldsymbol{\mathcal{H}}$.

Theorem 4. *Let $k \geq 2$ be the number of classes. For every distribution D over $\mathcal{X} \times \mathcal{Y}$ and for every distribution ρ over a set of multiclass voters $\boldsymbol{\mathcal{H}}$, we have:*

$$\mathop{\mathbf{Pr}}_{(x,y)\sim D}\left(M_{\rho,\frac{1}{k}}(x,y) \leq 0\right) \;\leq\; R_D(B_\rho) \;\leq\; \mathop{\mathbf{Pr}}_{(x,y)\sim D}\left(M_{\rho,\frac{1}{2}}(x,y) \leq 0\right).$$

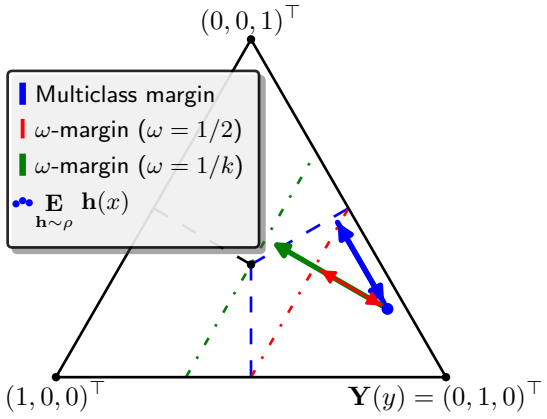

Figure 1: **Illustration of the multiclass margins.** Representation of the multiclass margins and the vote applied on an example $(x, y) \in \mathcal{X} \times \mathcal{Y}$ in the barycentric coordinate system defined by conv(Im $\mathcal{Y}$) when $\mathcal{Y} = \{1, 2, 3\}$ and the true class y is 2, *i.e.*, $(0, 1, 0)^\top$. We have $\mathbf{Y}(1) = (1, 0, 0)^\top$, $\mathbf{Y}(2) = (0, 1, 0)^\top$, and $\mathbf{Y}(3) = (0, 0, 1)^\top$. Each line is the decision boundary of a margin: the hyperplane where lies each example with a margin equals to 0. A vote correctly classifies an example if it lies on the same side of the hyperplane than the correct class.

Proof. First, let us prove the left-hand side inequality. We have:

$$
\begin{aligned}
R_D(B_\rho) &= \mathop{\mathbf{Pr}}_{(x,y)\sim D} \left(M_\rho(x,y) \leq 0\right) \\
&= \mathop{\mathbf{Pr}}_{(x,y)\sim D} \left(\mathop{\mathbf{E}}_{\mathbf{h}\sim\rho} h_y(x) \leq \max_{c\in\mathcal{Y}, c\neq y} \mathop{\mathbf{E}}_{\mathbf{h}\sim\rho} h_c(x) \right) \\
&\geq \mathop{\mathbf{Pr}}_{(x,y)\sim D} \left(\mathop{\mathbf{E}}_{\mathbf{h}\sim\rho} h_y(x) \leq \mathop{\mathbf{E}}_{c\in\mathcal{Y}, c\neq y} \mathop{\mathbf{E}}_{\mathbf{h}\sim\rho} h_c(x) \right) \\
&\geq \mathop{\mathbf{Pr}}_{(x,y)\sim D} \left(\mathop{\mathbf{E}}_{\mathbf{h}\sim\rho} h_y(x) \leq \frac{1}{k-1} \sum_{c=1, c\neq y}^{k} \mathop{\mathbf{E}}_{\mathbf{h}\sim\rho} h_c(x) \right) \\
&= \mathop{\mathbf{Pr}}_{(x,y)\sim D} \left(\mathop{\mathbf{E}}_{\mathbf{h}\sim\rho} h_y(x) \leq \frac{1}{k-1} \left[1 - \mathop{\mathbf{E}}_{\mathbf{h}\sim\rho} h_y(x) \right] \right) \\
&= \mathop{\mathbf{Pr}}_{(x,y)\sim D} \left(\mathop{\mathbf{E}}_{\mathbf{h}\sim\rho} h_y(x) - \frac{1}{k} \leq 0 \right) \\
&= \mathop{\mathbf{Pr}}_{(x,y)\sim D} \left(M_{\rho,\frac{1}{k}}(x,y) \leq 0 \right) .
\end{aligned}
$$

The right-hand side inequality is verified by observing that B_ρ necessarily makes a correct prediction if the weight $\mathbf{E}_{\mathbf{h}\sim\rho} h_y(x)$ given to the correct y is higher than $\frac{1}{2}$. □

Consequently, as illustrated in Figure 1, the ω-margin of the points that lie between the $\frac{1}{k}$-margin and the $\frac{1}{2}$-margin can be negative or positive according to ω. We thus have the following bound.

Corollary 1 (ω-margin multiclass $\mathcal{C}$-bound)**.** *For every probability distribution ρ on a set of multiclass voters $\boldsymbol{\mathcal{H}}$, and for every distribution D on $\mathcal{X} \times \mathcal{Y}$, if $\mathbf{E}_{(x,y)\sim D} M_{\rho,\frac{1}{2}}(x,y) > 0$, then we have:*

$$
R_D(B_\rho) \;\leq\; \mathcal{C}(M_{\rho,\frac{1}{2}}) \;=\; 1 - \frac{\left(\mathop{\mathbf{E}}_{(x,y)\sim D} M_{\rho,\frac{1}{2}}(x,y) \right)^2}{\mathop{\mathbf{E}}_{(x,y)\sim D} M^2_{\rho,\frac{1}{2}}(x,y)} ,
$$

where $\mathcal{C}(\cdot)$ is the function involved in the ω-margin-based $\mathcal{C}$-bound (Equation (5)).

The region of indecision when $\omega \in [\frac{1}{k}, \frac{1}{2}]$ implies there is possibly some value of ω to be chosen carefully to provide a good estimator of the true margin. If this is so, we can consider to make use of $\mathcal{C}(M_{\rho,\omega})$ for

Quantity	Pearson correlation
Trivial bound of Equation (1)	0.6709
$\mathcal{C}(M_\rho)$, the Multiclass (Theorem 2)	0.8758
$\mathcal{C}(M_{\rho,\omega})$ with $\omega = 1/2$ (Corollary 1)	0.5535
$\mathcal{C}(M_{\rho,\omega})$ with $\omega = 1/3 + 1/(3k)$	0.8811
$\mathcal{C}(M_{\rho,\omega})$ with $\omega = 1/6 + 2/(3k)$	**0.8950**
$\mathcal{C}(M_{\rho,\omega})$ with $\omega = 1/k$	0.8627

Table 1: Pearson correlations of the bounds with the risk of the majority vote. All values are evaluated on the test set T, averaged for 10 random train/test splits.

that particular value of ω to improve the analysis of the majority vote's behavior. Obviously, in such a situation, $\mathcal{C}(M_{\rho,\omega})$ is no longer a bound on $R_D(B_\rho)$. However, due to the linearity of the ω-margin, this could open the way to a generalization of the MinCq algorithm of Laviolette et al. (2011) to the multiclass setting.

4.2 Experimental Evaluation of the Bounds in the Multiclass Setting

The binary $\mathcal{C}$-bound is known to be well-suited to characterize the behavior of the risk of the ρ-weighted majority vote, as their respective values are correlated (Lacasse et al., 2007). We extend this analysis by empirically evaluating the behavior of the multiclass $\mathcal{C}$-bounds introduced above on natural data. We generate multiclass ρ-weighted majority votes by running a multiclass version of AdaBoost (Freund and Schapire, 1997)—known as AdaBoost-SAMME[2] (Zhu et al., 2009)—on multiclass datasets from the UCI dataset repository (Blake and Merz, 1998). We split each dataset in two halves: a training set S and a test set T. We train the algorithm on S, using 100, 250, 500 and 1,000 decision trees of depth 2, 3, 4 and 5 as base voters, for a total of 16 majority votes per dataset. We repeat the process for 10 random train/test splits, and the reported values are all computed on the test set. Figure 2 shows, using 3 of these splits, how the values of different upper bounds relate with the risk of the majority vote, and how the choice of ω for various values of $\mathcal{C}(M_{\rho,\omega})$ affects the correlation with the risk. We finally report in Table 1 the Pearson product-moment correlation coefficients for all computed values, using the 10 train/test splits.

As pointed out before, we notice from Figure 2 and Table 1 that for some values ω, the values of $\mathcal{C}(M_{\rho,\omega})$ are highly correlated with the risk of the majority vote. Unfortunately, the only one that is an upper bound of the latter ($\omega = \frac{1}{2}$) does not show the same predictive power. Thus, these results also give some empirical evidence that a wise choice of ω can improve the correlation between the $\mathcal{C}$-bound based on the ω-margin and the risk of the vote.

These experiments confirm the usefulness of the $\mathcal{C}$-bounds based on a notion of margin to upper-bound the true risk of the ρ-weighted majority vote. Taking into account the first and second statistical moments of such margins seems effectively very informative. This property is interesting from an algorithmic viewpoint: one may derive a multiclass optimization algorithm generalizing the algorithm MinCq (Laviolette et al., 2011) by minimizing $\mathcal{C}(M_{\rho,\omega})$, with ω considered as a hyperparameter to tune (*e.g.*, by cross-validation).

5 Conclusion

In the context of binary classification, it is well known that the PAC-Bayesian $\mathcal{C}$-bound offers a tight bound over the risk of the ρ-weighted majority vote by taking into account the first two statistical moments of its margin. Moreover, from a practical standpoint, minimizing the $\mathcal{C}$-bound leads to a well-performing algorithm called MinCq (Laviolette et al., 2011). This paper fills the gap between this binary PAC-Bayesian bound and more complex tasks by generalizing the $\mathcal{C}$-bound for majority vote over complex output voters, and by proposing a new surrogate of the margin that is easier to manipulate; we also explain how an empirical estimation of the $\mathcal{C}$-bound may be related to its expectation thanks to PAC-Bayesian results. In addition, we show how to specialize our result to multiclass and we provide in C insights as to how the bound we propose may be instantiated for multilabel classification—thoroughly studying the case of multilabel classification from the standpoint of the results we have provided, together with accompanying empirical results would deserve a whole paper.

[2]We use of the implementation provided in the Scikit-Learn Python library (Pedregosa et al., 2011).

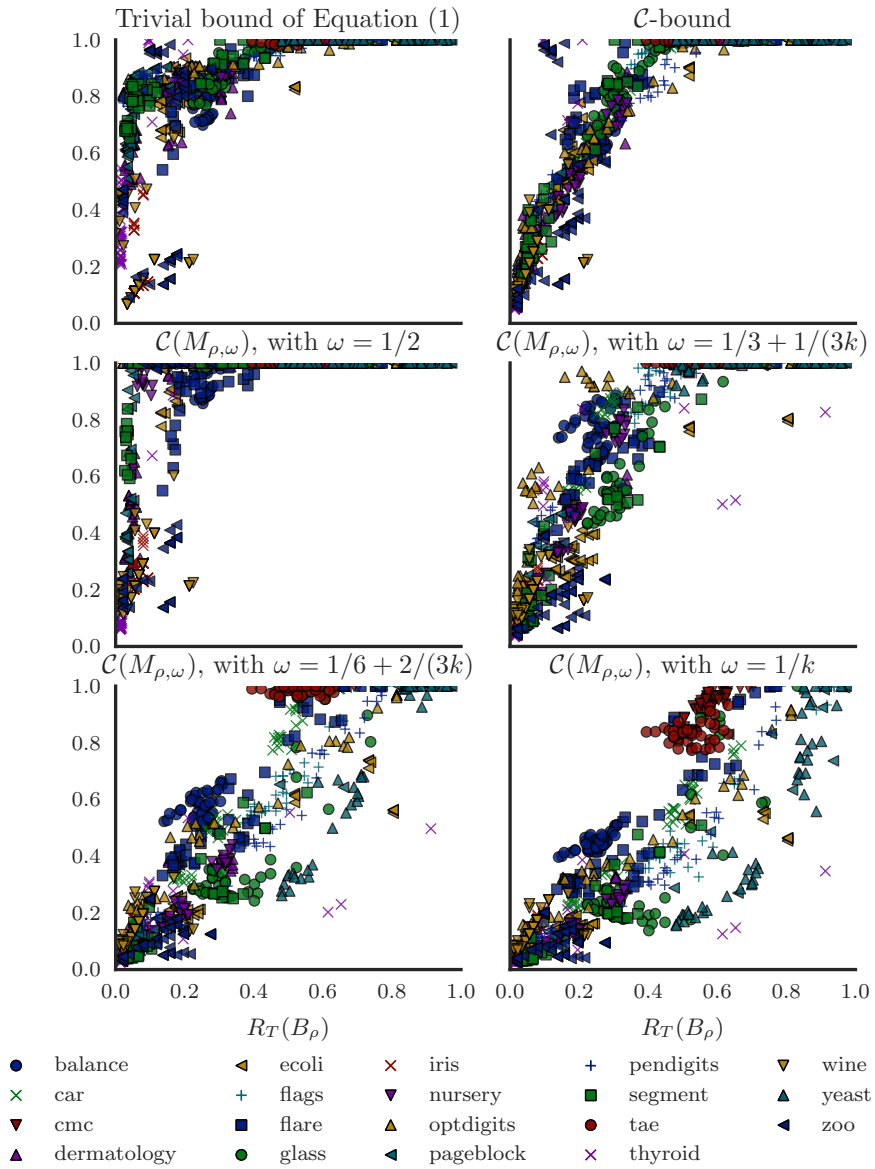

Figure 2: **Empirical results.** Comparison of the *true* risk of the ρ-weighted majority with: the trivial bound of Equation (1), the $\mathcal{C}$-bound, and $\mathcal{C}(M_{\rho,\omega})$ for various values of ω. All the values were calculated on a test set disjoint from the one used to learn ρ.

Concretely, we think that the theoretical $\mathcal{C}$-bounds provided here are a first step towards developing ensemble methods to learn ρ-weighted majority vote for complex outputs through the minimization of a $\mathcal{C}$-bound, or of a surrogate of it. A first solution for deriving such a method could be to study the general weak learning conditions necessary and sufficient to define an ensemble of structured output voters, as done by Mukherjee and Schapire (2013) for multiclass boosting. From a theoretical standpoint, we would like to study how much our generalization bound is robust to label noise as done for example by (Liu and Tao, 2016).

Acknwoledgment

This work is partially supported by the French project LIVES ANR-15-CE23-0026-03.

A Recovering Binary classification from the General Framework

From the general framework of Section 3, many choices of feature maps $\mathbf{Y}(\cdot)$ lead back to binary classification. Perhaps the most intuitive choice would be to consider $\mathbf{Y} : \{-1,+1\} \to \mathbb{R}$, with $\mathbf{Y}(+1) = 1$ and $\mathbf{Y}(-1) = -1$, but this choice would give us a margin that do not lie in $[-1,1]$. To directly recover the binary framework of Section 2, one must use the feature map described below.

Let us consider $\mathbf{Y} : \{-1,+1\} \to \mathbb{R}^2$, with $\mathbf{Y}(+1) = \left(\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}\right)^\top$ and $\mathbf{Y}(-1) = \left(-\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}\right)^\top$. In this case, each voter $\mathbf{h}$ outputs a vector of $\mathbb{R}^2$ whose first coordinate h_1 is an element of $\left[-\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}\right]$, and the second coordinate is always $\frac{1}{\sqrt{2}}$. In this case, we have

$$\begin{aligned}
B_\rho(x) &= \operatorname*{argmax}_{c\in\{-1,+1\}} \left\langle \mathbf{Y}(c)\,,\, \mathop{\mathbf{E}}_{\mathbf{h}\sim\rho} \mathbf{h}(x) \right\rangle \\
&= \operatorname*{argmax}_{c\in\{-1,+1\}} \left\langle \left(\tfrac{1}{\sqrt{2}}\,c,\, \tfrac{1}{\sqrt{2}}\right)\,,\, \left(\tfrac{1}{\sqrt{2}} \mathop{\mathbf{E}}_{\mathbf{h}\sim\rho} \sqrt{2}\,h_1(x)\,,\, \tfrac{1}{\sqrt{2}}\right) \right\rangle \\
&= \operatorname*{argmax}_{c\in\{-1,+1\}} \left[\tfrac{1}{2}\,c \mathop{\mathbf{E}}_{\mathbf{h}\sim\rho} \sqrt{2}\,h_1(x) \,+\, \tfrac{1}{2} \right] \\
&= \operatorname*{argmax}_{c\in\{-1,+1\}} \left[c \mathop{\mathbf{E}}_{\mathbf{h}\sim\rho} \sqrt{2}\,h_1(x) \right] \\
&= \operatorname{sign}\left[\mathop{\mathbf{E}}_{\mathbf{h}\sim\rho} \sqrt{2}\,h_1(x) \right] .
\end{aligned}$$

and

$$\begin{aligned}
M_\rho(x,y) &= \left\langle \mathop{\mathbf{E}}_{\mathbf{h}\sim\rho} \mathbf{h}(x)\,,\, \mathbf{Y}(y) \right\rangle - \max_{\substack{c\in\{-1,+1\} \\ c\neq y}} \left\langle \mathop{\mathbf{E}}_{\mathbf{h}\sim\rho} \mathbf{h}(x)\,,\, \mathbf{Y}(c) \right\rangle \\
&= \left\langle \mathop{\mathbf{E}}_{\mathbf{h}\sim\rho} \mathbf{h}(x)\,,\, \mathbf{Y}(y) \right\rangle - \left\langle \mathop{\mathbf{E}}_{\mathbf{h}\sim\rho} \mathbf{h}(x)\,,\, \mathbf{Y}(-y) \right\rangle \\
&= \left(\tfrac{1}{2}\,y \mathop{\mathbf{E}}_{\mathbf{h}\sim\rho} \sqrt{2}\,h_1(x) \,+\, \tfrac{1}{2} \right) \,-\, \left(-\tfrac{1}{2}\,y \mathop{\mathbf{E}}_{\mathbf{h}\sim\rho} \sqrt{2}\,h_1(x) \,+\, \tfrac{1}{2} \right) \\
&= y \mathop{\mathbf{E}}_{\mathbf{h}\sim\rho} \sqrt{2}\,h_1(x)\,.
\end{aligned}$$

As $\mathbf{E}_{\mathbf{h}\sim\rho} \sqrt{2}\,h_1(x)$ represents the "binary" margin that lies in $[-1,1]$ and $y \in \{-1,1\}$ is the binary label, we recover the usual definitions of Section 2.

B The Bounds Required to Prove Theorem 3

Theorem 5. *For any distribution D on $\mathcal{X}\times\mathcal{Y}$, for any set $\mathcal{H}$ of voters from $\mathcal{X}$ to $\mathrm{conv}(\mathrm{Im}\,\mathcal{Y})$, for any prior distribution π on $\mathcal{H}$ and any $\delta\in(0,1]$, with a probability at least $1-\delta$ over the choice of the m-sample $S\sim(D)^m$, for every posterior distribution ρ over $\mathcal{H}$ we have :*

$$\mathop{\mathbf{E}}_{(x,y)\in D} M_\rho(x,y) \;\geq\; \frac{1}{m} \sum_{(x,y)\in S} M_\rho(x,y) - B\sqrt{\frac{2}{m}\left[\mathrm{KL}(\rho\|\pi) + \ln\frac{2\sqrt{m}}{\delta}\right]},$$

where $B\in(0,2]$ bounds the absolute value of the margin $|M_\rho(x,y)|$ for all (x,y), and $\mathrm{KL}(\rho\|\pi) = \mathbf{E}_{\mathbf{h}\sim\rho} \ln\frac{\rho(\mathbf{h})}{\pi(\mathbf{h})}$ is the Kullback-Leibler divergence between ρ and π.

Proof. The following proof shows how to obtain the lower bound on the first moment of $M_\rho(x,y)$, and uses the same notions as the classical PAC-Bayesian proofs.[3]

Given a distribution D' on $\mathcal{X}\times\mathcal{Y}$, for any distribution ρ' over $\mathcal{H}$, we can rewrite $\mathbf{E}_{(x,y)\sim D'} M_{\rho'}(x,y)$ as an expectation over ρ'. We denote $M_{\mathbf{h}}^{D'}$ the random variable such that $\mathbf{E}_{\mathbf{h}\sim\rho'} M_{\mathbf{h}}^{D'} = \mathbf{E}_{(x,y)\sim D'} M_{\rho'}(x,y)$.

[3] The reader can refer to (Germain et al., 2009; Seeger, 2003; Catoni, 2007; McAllester, 2003; Germain et al., 2015) for examples of classical PAC-Bayesian analyses.

First, we have that $\mathbf{E}_{\mathbf{h}\sim\pi} \exp\left[\frac{m}{2B^2}\left(M^S_{\mathbf{h}} - M^D_{\mathbf{h}}\right)^2\right]$ is a non-negative random variable. Applying Markov's inequality yields that with probability at least $1-\delta$ over the choice of $S \sim (D)^m$, we have:

$$\mathop{\mathbf{E}}_{\mathbf{h}\sim\pi} \exp\left[\frac{m}{2B^2}\left(M^S_{\mathbf{h}} - M^D_{\mathbf{h}}\right)^2\right] \;\leq\; \frac{1}{\delta} \mathop{\mathbf{E}}_{S\sim D^m} \mathop{\mathbf{E}}_{\mathbf{h}\sim\pi} \exp\left[\frac{m}{2B^2}\left(M^S_{\mathbf{h}} - M^D_{\mathbf{h}}\right)^2\right]. \tag{6}$$

We upper-bound the right-hand side of the inequality:

$$\begin{aligned}
\mathop{\mathbf{E}}_{S\sim D^m} \mathop{\mathbf{E}}_{\mathbf{h}\sim\pi} \exp\Big[\frac{m}{2B^2}\left(M^S_{\mathbf{h}} - M^D_{\mathbf{h}}\right)^2\Big] &= \mathop{\mathbf{E}}_{\mathbf{h}\sim\pi} \mathop{\mathbf{E}}_{S\sim D^m} \exp\Big[\frac{m}{2B^2}\left(M^S_{\mathbf{h}} - M^D_{\mathbf{h}}\right)^2\Big] &(7)\\
&= \mathop{\mathbf{E}}_{\mathbf{h}\sim\pi} \mathop{\mathbf{E}}_{S\sim D^m} \exp\left[m\, 2\left(\frac{1}{2}\Big(1-\frac{1}{B}M^S_{\mathbf{h}}\Big) - \frac{1}{2}\Big(1-\frac{1}{B}M^D_{\mathbf{h}}\Big)\right)^2\right]\\
&\leq \mathop{\mathbf{E}}_{\mathbf{h}\sim\pi} \mathop{\mathbf{E}}_{S\sim D^m} \exp\left[m\,\mathrm{kl}\left(\frac{1}{2}\Big(1-\frac{M^S_{\mathbf{h}}}{B}\Big)\middle\|\frac{1}{2}\Big(1-\frac{M^D_{\mathbf{h}}}{B}\Big)\right)\right] &(8)\\
&\leq \mathop{\mathbf{E}}_{\mathbf{h}\sim\pi} 2\sqrt{m} \;=\; 2\sqrt{m}\,. &(9)
\end{aligned}$$

Line (7) comes from the fact that the distribution π is defined a priori. Since B is an upper bound of the possible absolute values of the margin, both $\frac{1}{2}\Big(1-\frac{M^S_{\mathbf{h}}}{B}\Big)$ and $\frac{1}{2}\Big(1-\frac{M^D_{\mathbf{h}}}{B}\Big)$ are between 0 and 1. Thus Line (8) is an application of Pinsker's inequality $2(q-p)^2 \leq \mathrm{kl}(q\|p) = q\ln\frac{q}{p} + (1-q)\ln\frac{1-q}{1-p}$. Finally, Line (9) is an application of (**?**)Theorem 5]m-04.

By applying this upper bound in Inequality (6) and by taking the logarithm on each side, with probability at least $1-\delta$ over the choice of $S \sim D^m$, we have:

$$\ln\left(\mathop{\mathbf{E}}_{\mathbf{h}\sim\pi} \exp\left[\frac{m}{2B^2}\left(M^S_{\mathbf{h}} - M^D_{\mathbf{h}}\right)^2\right]\right) \;\leq\; \ln\left(\frac{2\sqrt{m}}{\delta}\right).$$

Now, by applying the change of measure inequality proposed by Seldin *et al.* (**?**)Lemma 4]seldin-tishby-10 with $\phi(\mathbf{h}) = \frac{m}{2B^2}\left(M^S_{\mathbf{h}} - M^D_{\mathbf{h}}\right)^2$, and by using Jensen's inequality exploiting the convexity of $\phi(\mathbf{h})$, we obtain that for all distributions ρ on $\boldsymbol{\mathcal{H}}$:

$$\begin{aligned}
\ln\left(\mathop{\mathbf{E}}_{\mathbf{h}\sim\pi} \exp\left[\frac{m}{2B^2}\left(M^S_{\mathbf{h}} - M^D_{\mathbf{h}}\right)^2\right]\right) &\geq \mathop{\mathbf{E}}_{\mathbf{h}\sim\rho} \frac{m}{2B^2}\left(M^S_{\mathbf{h}} - M^D_{\mathbf{h}}\right)^2 - \mathrm{KL}(\rho\|\pi)\\
&\geq \frac{m}{2B^2}\left(\mathop{\mathbf{E}}_{\mathbf{h}\sim\rho} M^S_{\mathbf{h}} - \mathop{\mathbf{E}}_{\mathbf{h}\sim\rho} M^D_{\mathbf{h}}\right)^2 - \mathrm{KL}(\rho\|\pi)\,.
\end{aligned}$$

From all what precedes, we have that with probability at least $1-\delta$ over the choice of $S \sim (D)^m$, for every posterior distribution ρ on $\boldsymbol{\mathcal{H}}$, we have:

$$\frac{m}{2B^2}\left(\mathop{\mathbf{E}}_{(x,y)\sim S} M_\rho(x,y) - \mathop{\mathbf{E}}_{(x,y)\sim D} M_\rho(x,y)\right)^2 - \mathrm{KL}(\rho\|\pi) \;\leq\; \ln\left(\frac{2\sqrt{m}}{\delta}\right).$$

The result follows from algebraic calculations. □

Theorem 6. *For any distribution D on $\mathcal{X}\times\mathcal{Y}$, for any set $\boldsymbol{\mathcal{H}}$ of voters from $\mathcal{X}$ to $\mathrm{conv}(\mathrm{Im}\,\mathcal{Y})$, for any prior distribution π on $\boldsymbol{\mathcal{H}}$ and any $\delta\in(0,1]$, with a probability at least $1-\delta$ over the choice of the m-sample $S\sim(D)^m$, for every posterior distribution ρ over $\boldsymbol{\mathcal{H}}$ we have :*

$$\mathop{\mathbf{E}}_{(x,y)\in D}\left(M_\rho(x,y)\right)^2 \;\leq\; \frac{1}{m}\sum_{(x,y)\in S}\left(M_\rho(x,y)\right)^2 + B^2\sqrt{\frac{2}{m}\left[2\mathrm{KL}(\rho||\pi) + \ln\frac{2\sqrt{m}}{\delta}\right]},$$

where $B\in(0,2]$ bounds the absolute value of the margin $|M_\rho(x,y)|$ for all (x,y), and $\mathrm{KL}(\rho\|\pi) = \mathbf{E}_{\mathbf{h}\sim\rho}\ln\frac{\rho(\mathbf{h})}{\pi(\mathbf{h})}$ is the Kullback-Leibler divergence between ρ and π.

Proof. This proof uses many notions that are usual in classical PAC-Bayesian proofs, but the expectation over single voters is replaced with an expectation over pairs of voters. Given a distribution D' on $\mathcal{X}\times\mathcal{Y}$, for any distribution ρ'^2 over $\boldsymbol{\mathcal{H}}$, we rewrite $\mathbf{E}_{(x,y)\sim D'}(M_{\rho'}(x,y))^2$ as an expectation over ρ'^2. Let $M^{D'}_{\mathbf{h},\mathbf{h}'}$ be the r.v. such that $\mathbf{E}_{(\mathbf{h},\mathbf{h}')\sim\rho'^2} M^{D'}_{\mathbf{h},\mathbf{h}'} = \mathbf{E}_{(x,y)\sim D'}(M_{\rho'}(x,y))^2$. First, we apply the Markov's inequality on the

non-negative r.v. $\mathbf{E}_{(\mathbf{h},\mathbf{h}')\sim\pi^2} \exp\left[\frac{m}{2B^4}\left(M^S_{\mathbf{h},\mathbf{h}'}-M^D_{\mathbf{h},\mathbf{h}'}\right)^2\right]$. Thus, we have that with probability at least $1-\delta$ over the choice of $S\sim D^m$:

$$\mathop{\mathbf{E}}_{(\mathbf{h},\mathbf{h}')\sim\pi^2} \exp\left[\frac{m}{2B^4}\left(M^S_{\mathbf{h},\mathbf{h}'}-M^D_{\mathbf{h},\mathbf{h}'}\right)^2\right] \leq \frac{1}{\delta} \mathop{\mathbf{E}}_{S\sim(D)^m} \mathop{\mathbf{E}}_{(\mathbf{h},\mathbf{h}')\sim\pi^2} \exp\left[\frac{m}{2B^4}\left(M^S_{\mathbf{h},\mathbf{h}'}-M^D_{\mathbf{h},\mathbf{h}'}\right)^2\right]. \tag{10}$$

Then, we upper-bound the right-hand side of the inequality:

$$\begin{aligned}
&\mathop{\mathbf{E}}_{S\sim D^m} \mathop{\mathbf{E}}_{(\mathbf{h},\mathbf{h}')\sim\pi^2} \exp\left[\frac{m}{2B^4}\left(M^S_{\mathbf{h},\mathbf{h}'} - M^D_{\mathbf{h},\mathbf{h}'}\right)^2\right] \\
&= \mathop{\mathbf{E}}_{(\mathbf{h},\mathbf{h}')\sim\pi^2} \mathop{\mathbf{E}}_{S\sim D^m} \exp\left[\frac{m}{2B^4}\left(M^S_{\mathbf{h},\mathbf{h}'} - M^D_{\mathbf{h},\mathbf{h}'}\right)^2\right] &(11)\\
&= \mathop{\mathbf{E}}_{(\mathbf{h},\mathbf{h}')\sim\pi^2} \mathop{\mathbf{E}}_{S\sim D^m} \exp\left[m\,2\left(\frac{1}{2}\left(1-\frac{1}{B^2}M^S_{\mathbf{h},\mathbf{h}'}\right)-\frac{1}{2}\left(1-\frac{1}{B^2}M^D_{\mathbf{h},\mathbf{h}'}\right)\right)^2\right] \\
&\leq \mathop{\mathbf{E}}_{(\mathbf{h},\mathbf{h}')\sim\pi^2} \mathop{\mathbf{E}}_{S\sim D^m} \exp\left[m\,\mathrm{kl}\left(\frac{1}{2}\left(1-\frac{M^S_{\mathbf{h},\mathbf{h}'}}{B^2}\right)\middle\|\frac{1}{2}\left(1-\frac{M^D_{\mathbf{h},\mathbf{h}'}}{B^2}\right)\right)\right] &(12)\\
&\leq \mathop{\mathbf{E}}_{(\mathbf{h},\mathbf{h}')\sim\pi^2} 2\sqrt{m} \;=\; 2\sqrt{m}\,. &(13)
\end{aligned}$$

Line 11 comes from the fact that the distribution π is defined *a priori*, *i.e.*, before observing S. Since B upper-bounds the absolute value of the margin, both $\frac{1}{2}\left(1-\frac{M^S_{\mathbf{h},\mathbf{h}'}}{B^2}\right)$ and $\frac{1}{2}\left(1-\frac{M^D_{\mathbf{h},\mathbf{h}'}}{B^2}\right)$ lie between 0 and 1. Line 12 is then an application of Pinsker's inequality[4]. Finally, Line 13 is an application of (**?**)Theorem 5]m-04, which is stated to be valid for $m \geq 8$, but is also valid for any $m \geq 1$.
By applying this upper bound in Inequality (10) and by taking the logarithm on each side, with probability at least $1-\delta$ over the choice of $S\sim(D)^m$, we have:

$$\ln\left(\mathop{\mathbf{E}}_{(\mathbf{h},\mathbf{h}')\sim\pi^2} \exp\left[\frac{m}{2B^4}\left(M^S_{\mathbf{h},\mathbf{h}'} - M^D_{\mathbf{h},\mathbf{h}'}\right)^2\right]\right) \;\leq\; \ln\left(\frac{2\sqrt{m}}{\delta}\right).$$

Now, we need the change of measure inequality[5] of Lemma 1 (stated below) that has the novelty to use pairs of voters. By applying this lemma with $\phi(\mathbf{h},\mathbf{h}') = \frac{m}{2B^4}\left(M^S_{\mathbf{h},\mathbf{h}'}-M^D_{\mathbf{h},\mathbf{h}'}\right)^2$, and by using Jensen's inequality exploiting the convexity of $\phi(\mathbf{h},\mathbf{h}')$, we obtain that for all distributions ρ on $\boldsymbol{\mathcal{H}}$:

$$\begin{aligned}
\ln\left(\mathop{\mathbf{E}}_{(\mathbf{h},\mathbf{h}')\sim\pi^2} \exp\left[\frac{m}{2B^4}\left(M^S_{\mathbf{h},\mathbf{h}'}-M^D_{\mathbf{h},\mathbf{h}'}\right)^2\right]\right) &\geq \mathop{\mathbf{E}}_{(\mathbf{h},\mathbf{h}')\sim\rho^2} \frac{m}{2B^4}\left(M^S_{\mathbf{h},\mathbf{h}'}-M^D_{\mathbf{h},\mathbf{h}'}\right)^2 - 2\mathrm{KL}(\rho\|\pi) \\
&\geq \frac{m}{2B^4}\left(\mathop{\mathbf{E}}_{(\mathbf{h},\mathbf{h}')\sim\rho^2} M^S_{\mathbf{h},\mathbf{h}'} - \mathop{\mathbf{E}}_{(\mathbf{h},\mathbf{h}')\sim\rho^2} M^D_{\mathbf{h},\mathbf{h}'}\right)^2 - 2\mathrm{KL}(\rho\|\pi)\,.
\end{aligned}$$

From all what precedes, with probability at least $1-\delta$ on the choice of $S\sim D^m$, for every posterior distribution ρ on $\boldsymbol{\mathcal{H}}$, we have:

$$\frac{m}{2B^4}\left(\mathop{\mathbf{E}}_{(x,y)\sim S}(M_\rho(x,y))^2 - \mathop{\mathbf{E}}_{(x,y)\sim D}(M_\rho(x,y))^2\right)^2 - 2\mathrm{KL}(\rho\|\pi) \;\leq\; \ln\left(\frac{2\sqrt{m}}{\delta}\right).$$

The result follows from algebraic calculations. □

The change of measure used in the previous proof is stated below.

Lemma 1 (Change of measure inequality for pairs of voters)**.** *For any set of voters $\boldsymbol{\mathcal{H}}$, for any distributions π and ρ on $\boldsymbol{\mathcal{H}}$, and for any measurable function $\phi : \boldsymbol{\mathcal{H}}\times\boldsymbol{\mathcal{H}} \mapsto \mathbb{R}$, we have:*

$$\ln\left(\mathop{\mathbf{E}}_{(\mathbf{h},\mathbf{h}')\sim\pi^2} \exp\left[\phi(\mathbf{h},\mathbf{h}')\right]\right) \;\geq\; \mathop{\mathbf{E}}_{(\mathbf{h},\mathbf{h}')\sim\rho^2} \phi(\mathbf{h},\mathbf{h}') - 2\mathrm{KL}(\pi\|\rho)\,.$$

Proof. The proof is very similar to the one of Seldin et al.(**?**)Lemma 4]seldin-tishby-10, but is defined using pairs of voters. The first inequality below is given by using Jensen's inequality on the concave

[4]The Pinkser inequality is: $2(q-p)^2 \leq \mathrm{kl}(q\|p) = q\ln\frac{q}{p} + (1-q)\ln\frac{1-q}{1-p}$

[5]The change of measure is an important step in most PAC-Bayesian proofs (Seldin and Tishby, 2010).

function $\ln(\cdot)$.

$$\begin{aligned}
\mathop{\mathbf{E}}_{(\mathbf{h},\mathbf{h}')\sim\rho^2} \phi(\mathbf{h},\mathbf{h}') &= \mathop{\mathbf{E}}_{(\mathbf{h},\mathbf{h}')\sim\rho^2} \ln\Big(\mathrm{e}^{\phi(\mathbf{h},\mathbf{h}')}\Big) = \mathop{\mathbf{E}}_{(\mathbf{h},\mathbf{h}')\sim\rho^2} \ln\left(\mathrm{e}^{\phi(\mathbf{h},\mathbf{h}')}\,\frac{\rho^2(\mathbf{h},\mathbf{h}')}{\pi^2(\mathbf{h},\mathbf{h}')}\frac{\pi^2(\mathbf{h},\mathbf{h}')}{\rho^2(\mathbf{h},\mathbf{h}')}\right)\\
&= \mathrm{KL}(\rho^2\|\pi^2) + \mathop{\mathbf{E}}_{(\mathbf{h},\mathbf{h}')\sim\rho^2} \ln\left(\mathrm{e}^{\phi(\mathbf{h},\mathbf{h}')}\,\frac{\pi^2(\mathbf{h},\mathbf{h}')}{\rho^2(\mathbf{h},\mathbf{h}')}\right)\\
&\le \mathrm{KL}(\rho^2\|\pi^2) + \ln\left(\mathop{\mathbf{E}}_{(\mathbf{h},\mathbf{h}')\sim\rho^2} \mathrm{e}^{\phi(\mathbf{h},\mathbf{h}')}\,\frac{\pi^2(\mathbf{h},\mathbf{h}')}{\rho^2(\mathbf{h},\mathbf{h}')}\right)\\
&\le \mathrm{KL}(\rho^2\|\pi^2) + \ln\left(\mathop{\mathbf{E}}_{(\mathbf{h},\mathbf{h}')\sim\pi^2} \mathrm{e}^{\phi(\mathbf{h},\mathbf{h}')}\right)\\
&= 2\mathrm{KL}(\rho\|\pi) + \ln\left(\mathop{\mathbf{E}}_{(\mathbf{h},\mathbf{h}')\sim\pi^2} \mathrm{e}^{\phi(\mathbf{h},\mathbf{h}')}\right).
\end{aligned}$$

Note that the last inequality becomes an equality if ρ and π share the same support. The last equality comes from the definition of the KL-divergence, and from the fact that $\pi^2(\mathbf{h},\mathbf{h}') = \pi(\mathbf{h})\pi(\mathbf{h}')$ and $\rho^2(\mathbf{h},\mathbf{h}') = \rho(\mathbf{h})\rho(\mathbf{h}')$. □

C Specializations of the $\mathcal{C}$-bound to Multilabel Prediction

We instantiate the general $\mathcal{C}$-bound approach to multilabel classification. To do so, we stand in the following setting, where the space of possible labels is $\{1,\ldots,k\}$ with a finite number of classes $k \ge 2$, but we consider the *multilabel* output space $\boldsymbol{\mathcal{Y}} = \{0,1\}^k$ that contains vectors $\mathbf{y} = (y_1,\ldots,y_k)^\top$. In other words we consider multiple binary labels. Given an example $(x,\mathbf{y}) \in \mathcal{X}\times\boldsymbol{\mathcal{Y}}$, the output vector $\mathbf{y}$ is then defined as follows:

$$\forall j \in \{1,\ldots,k\}, \quad y_j = \begin{cases} 1 \text{ if } x \text{ is labeled with } j \\ 0 \text{ otherwise.}\end{cases}$$

In this specific case, we define the output feature map $\mathbf{Y}(\cdot)$ such that the image of $\boldsymbol{\mathcal{Y}}$ is $\mathrm{Im}\,\boldsymbol{\mathcal{Y}} = \left\{-\frac{1}{\sqrt{k}}, \frac{1}{\sqrt{k}}\right\}^k$, and:

$$\forall j \in \{1,\ldots,k\}, \quad Y_j(\mathbf{y}) = \begin{cases} \frac{1}{\sqrt{k}} \text{ if } y_j = 1 \ (x \text{ is labeled with } j) \\ \frac{-1}{\sqrt{k}} \text{ otherwise,}\end{cases}$$

where $Y_j(\mathbf{y})$ is the j-th coordinate of $\mathbf{Y}(\mathbf{y})$. According to this definition, we have that: $\forall \mathbf{c} \in \boldsymbol{\mathcal{Y}},\ \|\mathbf{Y}(\mathbf{c})\| = 1$. The set $\boldsymbol{\mathcal{H}}$ is made of *multilabel voters* $\mathbf{h}$ from $\mathcal{X}$ to $\mathrm{conv}(\mathrm{Im}\,\boldsymbol{\mathcal{Y}})$. In the light of the feature output map $\mathbf{Y}(\cdot)$, the definition of the ρ-weighted majority vote classifier and the margin can respectively be rewritten as:

$$B_\rho(x) \;=\; \mathop{\mathrm{argmax}}_{\mathbf{c}\in\boldsymbol{\mathcal{Y}}} \left\{ \sum_{j=1}^{k} \mathop{\mathbf{E}}_{\mathbf{h}\sim\rho} h_j(x) Y_j(\mathbf{c}) \right\},$$

and

$$M_\rho(x,\mathbf{y}) \;=\; \sum_{j=1}^{k} \mathop{\mathbf{E}}_{\mathbf{h}\sim\rho} h_j(x) Y_j(\mathbf{y}) \;-\; \max_{\substack{\mathbf{c}\in\boldsymbol{\mathcal{Y}}\\ \mathbf{c}\neq\mathbf{y}}} \left[\sum_{j=1}^{k} \mathop{\mathbf{E}}_{\mathbf{h}\sim\rho} h_j(x) Y_j(\mathbf{c}) \right],$$

where $h_j(x)$ is the j-th coordinate of $\mathbf{h}(x)$.

The following theorem relates the risk of $B_\rho(\cdot)$ and the ω-margin associated to the distribution ρ on $\boldsymbol{\mathcal{H}}$.

Theorem 7. *Let $k \ge 2$ be the number of labels. For every distribution D over $\mathcal{X}\times\boldsymbol{\mathcal{Y}}$ and for every distribution ρ over a set of multilabel voters $\boldsymbol{\mathcal{H}}$, we have:*

$$R_D(B_\rho) \;\le\; \mathop{\mathbf{Pr}}_{(x,\mathbf{y})\sim D} \left(M_{\rho,\frac{k-1}{k}}(x,\mathbf{y}) \le 0 \right).$$

Proof. We have to show:

$$\mathop{\mathbf{Pr}}_{(x,\mathbf{y})\sim D} \left(M_\rho(x,\mathbf{y}) \le 0 \right) \le \mathop{\mathbf{Pr}}_{(x,\mathbf{y})\sim D} \left(M_{\rho,\frac{k-1}{k}}(x,\mathbf{y}) \le 0 \right).$$

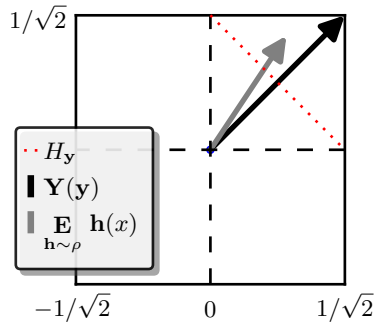

Figure 3: **Representation of the multilabel ω-margin.** Graphical representation of the $\frac{k-1}{k}$-margin and the vote applied on an example $(x, \mathbf{y})$ for multilabel classification when $k = 2$ and the true $\mathbf{y}$ is $(\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}})^\top$. The angles of the cube corresponds to the different multilabels, that are: $\mathbf{Y}(\boldsymbol{\mathcal{Y}}) = \left\{(\frac{-1}{\sqrt{2}}, \frac{-1}{\sqrt{2}})^\top, (\frac{1}{\sqrt{2}}, \frac{-1}{\sqrt{2}})^\top, (\frac{-1}{\sqrt{2}}, \frac{1}{\sqrt{2}})^\top, (\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}})^\top\right\}$. Each line represents the decision boundary of a margin: the hyperplane where lies each example with a margin equal to 0. A vote correctly classifies an example if it lies on the same side of the hyperplane than the correct label.

To do so we will prove that:

$$M_{\rho,\frac{k-1}{k}}(x, \mathbf{y}) > 0 \implies M_\rho(x, \mathbf{y}) > 0\,.$$

Recall that conv(Im $\boldsymbol{\mathcal{Y}}$) is a hypercube whose vertices are exactly the $\mathbf{Y}(\mathbf{c})$'s with $\mathbf{c} \in \boldsymbol{\mathcal{Y}}$. Given a vertex $\mathbf{Y}(\mathbf{y})$, we denote $H_{\mathbf{y}}$ the hyperplane which passes through all the points $\mathbf{Y}^{(j)}(\mathbf{y})$, where $\mathbf{Y}^{(j)}(\mathbf{y})$ is the point of the hypercube that has exactly the same coordinates as $\mathbf{Y}(\mathbf{y})$, excepting for the j^{th} that has been put to 0.

Now, consider the region $R_{\mathbf{y}}$ of the hypercube conv(Im $\boldsymbol{\mathcal{Y}}$) that consists of all the points that correspond to $M_{\rho,\frac{k-1}{k}}(x, \mathbf{y}) > 0$, that is, the points that are on the same side of hyperplane $H_{\mathbf{y}}$ than $\mathbf{Y}(\mathbf{y})$. Clearly, for any $k \geq 2$, the point $\mathbf{Y}(\mathbf{y})$ is strictly closer to the point $\mathbf{E}_{\mathbf{h}\sim\rho}\, \mathbf{h}(x)$ than any other $\mathbf{Y}(\mathbf{c})$'s if the vector $\mathbf{E}_{\mathbf{h}\sim\rho}\, \mathbf{h}(x)$ lies in $R_{\mathbf{y}}$. This in turn implies that the margin $M_\rho(x, \mathbf{y})$ is strictly positive. Figure 3 shows an example with $k = 2$ and $\mathbf{Y}(\mathbf{y}) = (\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}})^\top$, where $H_{\mathbf{y}}$ is represented by a red dotted line, and $R_{\mathbf{y}}$ is the region delimited by the top-right corner and $H_{\mathbf{y}}$.

To finish the proof, we have to show that $R_{\mathbf{y}}$ is exactly the region where $M_{\rho,\frac{k-1}{k}}(x, \mathbf{y}) > 0$. Equivalently, we must show that the intersection of $H_{\mathbf{y}}$ and the hypercube conv(Im $\boldsymbol{\mathcal{Y}}$) is exactly the points for which $M_{\rho,\frac{k-1}{k}}(x, \mathbf{Y}(\mathbf{y})) = 0$, *i.e.,* the vectors $\mathbf{E}_{\mathbf{h}\sim\rho}\, \mathbf{h}(x)$ for which $\langle \mathbf{E}_{\mathbf{h}\sim\rho}\, \mathbf{h}(x)\,,\, \mathbf{y}\rangle - \frac{k-1}{k} = 0$. We know from basic linear algebra that the points P that lie on hyperplane $H_{\mathbf{y}}$ must satisfy the following equation: $(P - P_0) \cdot N = 0$, where N is the normal of the hyperplane and P_0 is any point in P. It is easy to see that $\mathbf{Y}(\mathbf{y})$ is the normal of $H_{\mathbf{y}}$ and that we can take $P_0 = \mathbf{Y}^{(1)}(\mathbf{y})$. Hence, the equation becomes $(P - \mathbf{Y}^{(1)}(\mathbf{y})) \cdot \mathbf{Y}(\mathbf{y}) = 0$.

Since all coordinates of $\mathbf{Y}(\mathbf{y})$ are either $\frac{1}{\sqrt{k}}$ or $\frac{-1}{\sqrt{k}}$, and all coordinates of $\mathbf{Y}^{(1)}(\mathbf{y})$ are the same as the ones of $\mathbf{Y}(\mathbf{y})$ except the first one being 0 in $\mathbf{Y}^{(1)}(\mathbf{y})$, we have that $\mathbf{Y}^{(1)}(\mathbf{y}) \cdot \mathbf{Y}(\mathbf{y}) = \frac{k-1}{k}$. The result then follows from

$$(P - \mathbf{Y}^{(1)}(\mathbf{y})) \cdot \mathbf{Y}(\mathbf{y}) \;=\; P \cdot \mathbf{Y}(\mathbf{y}) - \mathbf{Y}^{(1)}(\mathbf{y}) \cdot \mathbf{Y}(\mathbf{y}) \;=\; \langle P\,,\, \mathbf{Y}(\mathbf{y})\rangle - \frac{k-1}{k}\,.$$

□

Finally, according to the same arguments as in Corollary 1, one can derive the following multilabel $\mathcal{C}$-bound.

Corollary 2 (ω-margin multilabel $\mathcal{C}$-bound)**.** *For every probability distribution ρ on a set of multilabel voters $\boldsymbol{\mathcal{H}}$, for every distribution D on $\mathcal{X} \times \boldsymbol{\mathcal{Y}}$, if $\mathbf{E}_{(x,\mathbf{y})\sim D}\, M_{\rho,\frac{k-1}{k}}(x, \mathbf{y}) > 0$, we have:*

$$R_D(B_\rho) \leq\; \mathcal{C}(M_{\rho,\frac{k-1}{k}}) \;=\; 1 - \frac{\left(\mathop{\mathbf{E}}_{(x,\mathbf{y})\sim D} M_{\rho,\frac{k-1}{k}}(x, \mathbf{y})\right)^2}{\mathop{\mathbf{E}}_{(x,\mathbf{y})\sim D}\left(M_{\rho,\frac{k-1}{k}}(x, \mathbf{y})\right)^2}\,.$$

References

Allwein, E., Schapire, R., Singer, Y., 2001. Reducing multiclass to binary: A unifying approach for margin classifiers. Journal of Machine Learning Research 1, 113–141.

Blake, C., Merz, C., 1998. UCI Repository of machine learning databases. Dpt. of Information & Computer Science, Univ. of California, archive.ics.uci.edu/ml.

Blanchard, G., 2004. Different paradigms for choosing sequential reweighting algorithms. Neural Computation 16 (4), 811–836.

Boser, B., Guyon, I., Vapnik, V., 1992. A training algorithm for optimal margin classifiers. In: Workshop on Computational learning theory. pp. 144–152.

Breiman, L., 1996. Bagging predictors. Machine Learning 24 (2), 123–140.

Breiman, L., 2001. Random Forests. Machine Learning 45 (1), 5–32.

Brouard, C., d'Alché Buc, F., Szafranski, M., 2011. Semi-supervised penalized output kernel regression for link prediction. In: International Conference on Machine Learning (ICML-11). pp. 593–600.

Catoni, O., 2007. PAC-Bayesian supervised classification: the thermodynamics of statistical learning. Vol. 56. IMS Lecture Notes Monograph Series.

Cortes, C., Kuznetsov, V., Mohri, M., 2014. Ensemble methods for structured prediction. In: Proceedings of the 31st International Conference on Machine Learning (ICML-14). pp. 1134–1142.

Cortes, C., Mohri, ., Weston, J., 2007. A general regression framework for learning string-to-string mappings. Predicting Structured Data, 143–168.

Cortes, C., Vapnik, V., 1995. Support-vector networks. Machine Learning 20 (3), 273–297.

Dietterich, T., Bakiri, G., 1995. Solving multiclass learning problems via error-correcting output codes. Journal of Artificial Intelligence Research 2 (263), 286.

Dietterich, T. G., 2000. Ensemble methods in machine learning. In: Multiple Classifier Systems. pp. 1–15.

Domingos, P., 2000. Bayesian averaging of classifiers and the overfitting problem. In: International Conference on Machine Learning. pp. 223–230.

Freund, Y., Schapire, R., 1997. A decision-theoretic generalization of on-line learning and an application to boosting. Journal of Computer and System Sciences 55, 119–139.

Gärtner, T., 2003. A survey of kernels for structured data. ACM SIGKDD Explorations Newsletter 5 (1), 49–58.

Gelman, A., Carlin, J., Stern, H., Rubin, D., 2004. Bayesian data analysis. Chapman & Hall/CRC.

Germain, P., Lacasse, A., Laviolette, F., Marchand, M., 2009. PAC-Bayesian learning of linear classifiers. In: International Conference on Machine Learning. pp. 353–360.

Germain, P., Lacasse, A., Laviolette, F., Marchand, M., Roy, J.-F., 2015. Risk bounds for the majority vote: From a PAC-Bayesian analysis to a learning algorithm. Journal of Machine Learning Research 16, 787–860.

Giguere, S., Laviolette, F., Marchand, M., Rolland, A., 2014. Pac-bayesian risk bounds and learning algorithms for the regression approach to structured output prediction. Advanced Structured Prediction, 239.

Haussler, D., Kearns, M., Schapire, R., 1994. Bounds on the sample complexity of bayesian learning using information theory and the VC dimension. Machine Learning 14 (1), 83–113.

Kuznetsov, V., Mohri, M., Syed, U., 2014. Multi-class deep boosting. In: Advances in Neural Information Processing Systems. pp. 2501–2509.

Lacasse, A., Laviolette, F., Marchand, M., Germain, P., Usunier, N., 2007. PAC-Bayes bounds for the risk of the majority vote and the variance of the Gibbs classifier. In: Advances in Neural Information Processing Systems. pp. 769–776.

Langford, J., Shawe-Taylor, J., 2002. PAC-Bayes & margins. In: Advances in Neural Information Processing Systems. pp. 423–430.

Laviolette, F., Marchand, M., Roy, J.-F., 2011. From PAC-Bayes bounds to quadratic programs for majority votes. In: International Conference on Machine Learning. pp. 649–656.

Li, L., Zou, B., Hu, Q., Wu, X., Yu, D., 2013. Dynamic classifier ensemble using classification confidence. Neurocomputing 99, 581–591.

Liu, T., Tao, D., 2016. Classification with noisy labels by importance reweighting. IEEE Transactions on pattern analysis and machine intelligence 38 (3), 447–461.

McAllester, D., 1999. Some PAC-Bayesian theorems. Machine Learning 37, 355–363.

McAllester, D., 2003. Simplified PAC-Bayesian margin bounds. In: Learning Theory and Kernel Machines. Springer, pp. 203–215.

McAllester, D., 2009. Generalization bounds and consistency for structured labeling. In: Predicting Structured Data. MIT Press, pp. 247–262.

Morvant, E., Habrard, A., Ayache, S., 2014. Majority vote of diverse classifiers for late fusion. In: IAPR Joint International Workshops on Statistical Techniques in Pattern Recognition and Structural and Syntactic Pattern Recignition. pp. 153–162.

Morvant, E., Koço, S., Ralaivola, L., 2012. PAC-Bayesian generalization bound on confusion matrix for multi-class classification. In: International Conference on Machine Learning.

Mroueh, Y., Poggio, T., Rosasco, L., Slotine, J.-J., 2012. Multiclass learning with simplex coding. In: Advances in Neural Information Processing Systems. pp. 2789–2797.

Mukherjee, I., Schapire, R., 2013. A theory of multiclass boosting. Journal of Machine Learning Research 14 (1), 437–497.

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

Re, M., Valentini, G., 2012. Ensemble methods: a review. Advances in machine learning and data mining for astronomy, 563–582.

Read, J., Pfahringer, B., Holmes, G., Frank, E., 2011. Classifier chains for multi-label classification. Machine learning 85 (3), 333–359.

Schapire, R., Singer, Y., 1999. Improved boosting algorithms using confidence-rated predictions. In: Machine Learning. pp. 80–91.

Seeger, M., 2003. PAC-Bayesian generalisation error bounds for gaussian process classification. Journal of Machine Learning Research 3, 233–269.

Seldin, Y., Tishby, N., 2010. PAC-Bayesian analysis of co-clustering and beyond. Journal of Machine Learning Research 11, 3595–3646.

Sun, S., 2013. A survey of multi-view machine learning. Neural Computing and Applications 23 (7-8), 2031–2038.

Tsoumakas, G., Vlahavas, I., 2007. Random k-labelsets: An ensemble method for multilabel classification. In: European Conference on Machine Learning. pp. 406–417.

Yu, J., Rui, Y., Tao, D., 2014. Click prediction for web image reranking using multimodal sparse coding. IEEE Transactions on Image Processing 23 (5), 2019–2032.

Zhang, Y., Schneider, J., 2012. Maximum margin output coding. In: International Conference on Machine Learning. pp. 1575–1582.

Zhu, J., Zou, H., Rosset, S., Hastie, T., 2009. Multi-class adaboost. Statistics and its Interface 2 (3), 349–360.