On the Generalization of the C-Bound to Multiclass Setting

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

On the Generalization of the C-Bound to Multiclass Setting

Emilie Morvant[1*] Jean-Francis Roy[2] François Laviolette[2] Liva Ralaivola[3]
[1] Institute of Science and Technology (IST) Austria, Klosterneuburg, Austria
[2] Département d'informatique et de génie logiciel, Université Laval, Québec, Canada
[3] Aix-Marseille Univ., LIF-QARMA, CNRS, UMR 7279, F-13013, Marseille, France

August 6, 2014

Abstract

PAC-Bayesian theory provides generalization bounds for weighted majority vote. However, these bounds do not directly focus on the risk of the majority vote, but on the risk of the Gibbs classifier. Indeed, it is well-known that the Gibbs classifier and the majority vote are related. To the best of our knowledge the tightest relation is the C-bound proposed by Lacasse et al. (2007); Laviolette et al. (2011) for binary classification. In this paper, we provide three generalizations of the C-bound to multiclass setting.

Keywords: Majority vote, Multiclass Classification, PAC-Bayesian Theory, C-Bound

1 Introduction

The PAC-Bayesian framework, first introduced by McAllester (2003), is an important field in machine learning theory. Given a family $\mathcal{H}$ of models, called voters, given a *prior* distribution over $\mathcal{H}$, and given a learning sample S, the PAC-Bayesian approach aims at learning a *posterior* distribution ρ over $\mathcal{H}$ that leads to a well-performing ρ-weighted majority vote[1] (over $\mathcal{H}$). An important result in PAC-Bayesian theory known as "the PAC-Bayes theorem" (McAllester, 2003; Langford & Shawe-Taylor, 2002; Seeger, 2002; Catoni, 2007; Germain et al., 2009) offers generalization bound for the ρ-majority vote by bounding the risk of the stochastic Gibbs classifier associated to ρ, which predicts the label of an example $\mathbf{x}$ by first picking h in $\mathcal{H}$ according to ρ, and then by returning $h(\mathbf{x})$. However, in binary classification this bound is based upon that the risk of the majority vote is bounded by twice the error of the Gibbs classifier. Note that, in multiclass classification the risk of the majority vote is bounded by the number of classes times the error of the Gibbs classifier (Morvant et al., 2012). To tackle this drawback in binary classification setting, (Lacasse et al., 2007; Laviolette et al., 2011) proposed a tighter relation between the risk of the majority vote and the one of the Gibbs classifier: the C-bound. This bound involves the first two statistical moments of the margin of the majority vote. In this work we generalize the C-bound for multiclass weighted majority vote.

This paper is organized as follows. Section 2 recalls the C-bound in binary classification. We present our generalizations to the multiclass setting in Section **??**. We conclude in Section 4.

2 The C-Bound for Binary Classification

In this section, we recall the C-bound of Lacasse et al. (2007); Laviolette et al. (2011) in the binary classification setting, which stands in the PAC-Bayesian framework (first introduced by McAllester (2003)).

Let $X \subseteq \mathbb{R}^d$ be the input space of dimension d, and let $Y = \{-1, +1\}$ be the output space. The learning sample $S = \{(\mathbf{x}_i, y_i)\}_{i=1}^m$ is constituted by m examples *i.i.d.* from a fixed but unknown distribution P over $X \times Y$. Let $\mathcal{H}$ be a set of real-valued voters from X to Y. Given a prior distribution π over $\mathcal{H}$ and

*This work was carried out while Emilie Morvant was affiliated with Aix-Marseille Univ., LIF-QARMA, CNRS, UMR 7279, F-13013, Marseille, France.

[1] Note that the ρ-majority vote is sometimes called the Bayes classifier. However to avoid any confusion with the Bayes classifier in Bayesian classification we rather use the term majority vote.

given a sample S, the objective in the PAC-Bayesian approach is to find the posterior distribution ρ on $\mathcal{H}$ which minimizes the true risk of the ρ-weighted majority vote $B_\rho(\cdot)$ where:

$$\forall \mathbf{x} \in X,\ B_\rho(\mathbf{x}) = \text{sign}\left[\mathop{\mathbf{E}}_{h\sim\rho} h(\mathbf{x})\right].$$

Its true risk $\mathbf{R}_P(B_\rho)$ on the distribution P is:

$$\mathbf{R}_P(B_\rho) = \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim P} \mathbf{I}\left(B_\rho(\mathbf{x}) \neq y\right),$$

where $\mathbf{I}(a) = 1$ if predicate a is true and 0 otherwise.

It is well-know that minimizing $\mathbf{R}_P(B_\rho)$ is NP-hard. To get around this problem, one solution is to make use of the C-bound which is a tight bound over $\mathbf{R}_P(B_\rho)$. This bound is based on the notion of margin of $B_\rho(\cdot)$ defined as follows.

Definition 1 (ρ-margin)**.** *The ρ-margin of $B_\rho(\cdot)$ realized on an example $(\mathbf{x}, y) \sim P$ is:*

$$\mathcal{M}^\rho(\mathbf{x}, y) = y \mathop{\mathbf{E}}_{h\sim\rho} h(\mathbf{x}).$$

According to this definition, $B_\rho(\cdot)$ correctly classifies an example $(\mathbf{x}, y)$ when its ρ-margin is strictly positive:

$$\mathbf{R}_P(B_\rho) = \mathop{\mathbf{Pr}}_{(\mathbf{x},y)\sim P} \left(\mathcal{M}^\rho(\mathbf{x}, y) \leq 0\right). \tag{1}$$

This equality allows to prove the following theorem.

Theorem 1 (C-bound of Laviolette et al. (2011))**.** *For every distribution ρ on a set of real-valued functions $\mathcal{H}$, and for every distribution P on $X \times Y$, if $\mathcal{M}_P^\rho > 0$, then we have :*

$$\begin{aligned}\mathbf{R}_P(B_\rho) &\leq \frac{\mathbf{Var}_{(\mathbf{x},y)\sim P}\, \mathcal{M}^\rho(\mathbf{x}, y)}{\mathbf{E}_{(\mathbf{x},y)\sim P}\, \left(\mathcal{M}^\rho(\mathbf{x}, y)\right)^2}. \\ &= 1 - \frac{\left(\mathcal{M}_P^\rho\right)^2}{\mathcal{M}_P^{\rho^2}},\end{aligned}$$

where $\mathcal{M}_P^\rho$ and $\mathcal{M}_P^{\rho^2}$ are respectively the first and the second statistical moments of the ρ-margin on P, and are defined by:

$$\begin{aligned}\mathcal{M}_P^\rho &= \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim P} \mathcal{M}^\rho(\mathbf{x}, y) \\ &= \mathop{\mathbf{E}}_{h\sim\rho} \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim P} yh(\mathbf{x}), \\ \mathcal{M}_P^{\rho^2} &= \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim P} \left(\mathcal{M}^\rho(\mathbf{x}, y)\right)^2 \\ &= \mathop{\mathbf{E}}_{(h,h')\sim\rho^2} \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim P} h(\mathbf{x})h'(\mathbf{x}).\end{aligned}$$

Proof. We first introduce the Cantelli-Chebyshev inequality, that is the main tool of the proof.

Theorem 2 (Cantelli-Chebyshev Inequality)**.** *Let Z be a random variable. Then, we have:*

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

We prove the C-bound by applying the Cantelli-Chebyshev inequality on the random variable $\mathcal{M}^\rho(\mathbf{x}, y)$, and with $a = \mathcal{M}_P^\rho$. According to Definition 1, we have:

$$\mathop{\mathbf{Var}}_{(\mathbf{x},y)\sim P} \mathcal{M}^\rho(\mathbf{x}, y) = \mathcal{M}_P^{\rho^2} - \left(\mathcal{M}_P^\rho\right)^2.$$

Then:

$$\begin{aligned}\mathbf{R}_P(B_\rho) &= \Pr_{(\mathbf{x},y)\sim P}\left(\mathcal{M}^\rho(\mathbf{x},y)\leq 0\right)\\ &\leq \frac{\mathbf{Var}_{(\mathbf{x},y)\sim P}\,\mathcal{M}^\rho(\mathbf{x},y)}{\mathbf{Var}_{(\mathbf{x},y)\sim P}\,\mathcal{M}^\rho(\mathbf{x},y)+\left(\mathcal{M}_P^\rho\right)^2}\\ &= \frac{\mathbf{Var}_{(\mathbf{x},y)\sim P}\,\mathcal{M}^\rho(\mathbf{x},y)}{\mathbf{E}_{(\mathbf{x},y)\sim P}\,\left(\mathcal{M}^\rho(\mathbf{x},y)\right)^2}\\ &= 1-\frac{\left(\mathcal{M}_P^\rho\right)^2}{\mathcal{M}_P^{\rho^2}}.\end{aligned}$$

□

Note that the minimization of the empirical counterpart of the C-bound is a natural solution for learning a distribution ρ that leads to a ρ-weighted majority vote $B_\rho(\cdot)$ with a low error. This strategy is justified thanks to an elegant PAC-Bayesian generalization bound over the C-bound, and have lead to a simple and elegant algorithm called MinCq (See (Laviolette et al., 2011) for more details).

In the following, we propose to generalize this important theoretical result in the PAC-Bayesian literature to multiclass setting.

3 Generalization of the C-Bound for Multiclass Classification

3.1 Multiclass Setting

In this section, we stand in the multiclass classification setting where $X \subseteq \mathbb{R}^d$ is still the input space, but $Y = \{1,\ldots,Q\}$ is the output space with a finite number of classes $Q \geq 2$. Let $\mathcal{H}$ be a set of multiclass voters from X to Y. We recall that given a prior distribution π over $\mathcal{H}$ and given a learning sample S, *i.i.d.* from P, the PAC-Bayesian approach looks for the ρ distribution which minimizes the true risk of the ρ-weighted majority vote $B_\rho(\cdot)$. In the multiclass classification setting, $B_\rho(\cdot)$ is defined as follow:

$$B_\rho(\mathbf{x}) = \operatorname*{argmax}_{c\in Y}\left[\mathop{\mathbf{E}}_{h\sim\rho}\mathbf{I}(h(\mathbf{x})=c)\right]. \tag{2}$$

Given a distribution ρ on a set $\mathcal{H}$ of multiclass voters, we recall that the risk of the ρ-weighted majority vote $\mathbf{R}_P(B_\rho)$ is defined as the probability that it commits an error on an example coming from P:

$$\mathbf{R}_P(B_\rho) = \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim P}\mathbf{I}\Big(B_\rho(\mathbf{x})\neq y\Big).$$

An important notion related to the majority vote is the notion of margin realized on an example $(\mathbf{x},y)$. For multiclass classification, such a notion can be expressed in a variety of manners. In the next section, we present three versions of multiclass margin that are equivalent in binary classification.

3.1.1 Margins in Multiclass Classification

Firstly, we make use of the multiclass margin proposed by Breiman (2001) for the random forests, which can be seen as the usual notion of margin.

Definition 2 (ρ-margin)**.** *Let P be a distribution over $X\times Y$, let $\mathcal{H}$ be a set of multiclass voters. Given a distribution ρ on $\mathcal{H}$, the ρ-margin of the majority vote $B_\rho(\cdot)$ realized on a example $(\mathbf{x},y)\sim P$ is:*

$$\mathcal{M}^\rho(\mathbf{x},y) = \mathop{\mathbf{E}}_{h\sim\rho}\mathbf{I}(h(\mathbf{x})=y) - \max_{c\in Y, c\neq y}\left(\mathop{\mathbf{E}}_{h\sim\rho}\mathbf{I}(h(\mathbf{x})=c)\right).$$

Like in the binary classification framework presented in Section 2, the majority vote $B_\rho(\cdot)$ correctly classifies an example if its ρ-margin is strictly positive:

$$\mathbf{R}_P(B_\rho) = \Pr_{(\mathbf{x},y)\sim P}\left(\mathcal{M}^\rho(\mathbf{x},y)\leq 0\right). \tag{3}$$

Note that when $Y = \{-1, +1\}$, we find the usual notion of margin:

$$\begin{aligned}
\mathcal{M}^\rho(\mathbf{x}, y) &= \mathop{\mathbf{E}}_{h\sim\rho} \mathbf{I}(h(\mathbf{x}) = y) - \max_{c\in Y, c\neq y} \left(\mathop{\mathbf{E}}_{h\sim\rho} \mathbf{I}(h(\mathbf{x}) = c) \right) \\
&= \mathop{\mathbf{E}}_{h\sim\rho} \mathbf{I}(h(\mathbf{x}) = y) - \mathop{\mathbf{E}}_{h\sim\rho} \mathbf{I}(h(\mathbf{x}) \neq y) \\
&= \mathop{\mathbf{E}}_{h\sim\rho} \left[\mathbf{I}(h(\mathbf{x}) = y) - \mathbf{I}(h(\mathbf{x}) \neq y) \right] \\
&= \mathop{\mathbf{E}}_{h\sim\rho} y h(\mathbf{x}) \\
&= y \mathop{\mathbf{E}}_{h\sim\rho} h(\mathbf{x}) \,.
\end{aligned}$$

Since the ρ-margin is defined with a maximum, it could be hard to deal with it in an algorithmic perspective.

We then consider the relaxation proposed by Breiman (2001) that is based on a notion of strength for a given example $(\mathbf{x}, y)$. This notion is related to the deviation between the correct classification and the errors independently.

Definition 3 (ρ-strength)**.** *Let P be a distribution over $X \times Y$, let $\mathcal{H}$ be a set of multiclass voters from X to Y. Given a distribution ρ on $\mathcal{H}$, the ρ-strength of the majority vote $B_\rho(\cdot)$ realized on an example $(\mathbf{x}, y) \sim P$ for a class $c \in Y$ is:*

$$\mathcal{S}^\rho(c, (\mathbf{x}, y)) = \mathop{\mathbf{E}}_{h\sim\rho} \mathbf{I}(h(\mathbf{x}) = y) - \mathop{\mathbf{E}}_{h\sim\rho} \mathbf{I}(h(\mathbf{x}) = c).$$

From this definition and Equation (3) we have:

$$\begin{aligned}
\mathbf{R}_P(B_\rho) &= \mathop{\mathbf{Pr}}_{(\mathbf{x},y)\sim P} \left(\mathcal{M}^\rho(\mathbf{x}, y) \leq 0 \right) \\
&= \mathop{\mathbf{Pr}}_{(\mathbf{x},y)\sim P} \left(\exists c \in Y, c \neq y : \mathop{\mathbf{E}}_{h\sim\rho} \mathbf{I}(h(\mathbf{x}) = y) \leq \mathop{\mathbf{E}}_{h\sim\rho} \mathbf{I}(h(\mathbf{x}) = c) \right) \\
&= \mathop{\mathbf{Pr}}_{(\mathbf{x},y)\sim P} \left(\bigvee_{c=1,\ c\neq y}^{Q} \mathop{\mathbf{E}}_{h\sim\rho} \mathbf{I}(h(\mathbf{x}) = y) \leq \mathop{\mathbf{E}}_{h\sim\rho} \mathbf{I}(h(\mathbf{x}) = c) \right) \\
&= \mathop{\mathbf{Pr}}_{(\mathbf{x},y)\sim P} \left(\bigvee_{c=1}^{Q} \left[\mathop{\mathbf{E}}_{h\sim\rho} \mathbf{I}(h(\mathbf{x}) = y) \leq \mathop{\mathbf{E}}_{h\sim\rho} \mathbf{I}(h(\mathbf{x}) = c) \wedge c \neq y \right] \right) \\
&\leq \sum_{c=1}^{Q} \mathop{\mathbf{Pr}}_{(\mathbf{x},y)\sim P} \left(\mathop{\mathbf{E}}_{h\sim\rho} \mathbf{I}(h(\mathbf{x}) = y) \leq \mathop{\mathbf{E}}_{h\sim\rho} \mathbf{I}(h(\mathbf{x}) = c) \wedge c \neq y \right) \\
&= \sum_{c=1}^{Q} \left[\mathop{\mathbf{Pr}}_{(\mathbf{x},y)\sim P} \left(\mathop{\mathbf{E}}_{h\sim\rho} \mathbf{I}(h(\mathbf{x}) = y) \leq \mathop{\mathbf{E}}_{h\sim\rho} \mathbf{I}(h(\mathbf{x}) = c) \right) + \mathop{\mathbf{Pr}}_{(\mathbf{x},y)\sim P} \left(c \neq y \right) \right. \\
&\qquad \left. - \mathop{\mathbf{Pr}}_{(\mathbf{x},y)\sim P} \left(\mathop{\mathbf{E}}_{h\sim\rho} \mathbf{I}(h(\mathbf{x}) = y) \leq \mathop{\mathbf{E}}_{h\sim\rho} \mathbf{I}(h(\mathbf{x}) = c) \vee c \neq y \right) \right] \\
&= \sum_{c=1}^{Q} \left[\mathop{\mathbf{Pr}}_{(\mathbf{x},y)\sim P} \left(\mathop{\mathbf{E}}_{h\sim\rho} \mathbf{I}(h(\mathbf{x}) = y) \leq \mathop{\mathbf{E}}_{h\sim\rho} \mathbf{I}(h(\mathbf{x}) = c) \right) + \mathop{\mathbf{Pr}}_{(\mathbf{x},y)\sim P} \left(c \neq y \right) - 1 \right] \qquad (4) \\
&= \sum_{c=1}^{Q} \mathop{\mathbf{Pr}}_{(\mathbf{x},y)\sim P} \left(\mathop{\mathbf{E}}_{h\sim\rho} \mathbf{I}(h(\mathbf{x}) = y) \leq \mathop{\mathbf{E}}_{h\sim\rho} \mathbf{I}(h(\mathbf{x}) = c) \right) + \sum_{c=1}^{Q} \mathop{\mathbf{Pr}}_{(\mathbf{x},y)\sim P} \left(c \neq y \right) - \sum_{c=1}^{Q} 1 \\
&= \sum_{c=1}^{Q} \mathop{\mathbf{Pr}}_{(\mathbf{x},y)\sim P} \left(\mathop{\mathbf{E}}_{h\sim\rho} \mathbf{I}(h(\mathbf{x}) = y) \leq \mathop{\mathbf{E}}_{h\sim\rho} \mathbf{I}(h(\mathbf{x}) = c) \right) + (Q - 1) - Q \\
&= \sum_{c=1}^{Q} \mathop{\mathbf{Pr}}_{(\mathbf{x},y)\sim P} \left(\mathop{\mathbf{E}}_{h\sim\rho} \mathbf{I}(h(\mathbf{x}) = y) \leq \mathop{\mathbf{E}}_{h\sim\rho} \mathbf{I}(h(\mathbf{x}) = c) \right) - 1 \\
&= \sum_{c=1}^{Q} \mathop{\mathbf{Pr}}_{(\mathbf{x},y)\sim P} \left(\mathcal{S}^\rho(c, (\mathbf{x}, y)) \leq 0 \right) - 1. \qquad (5)
\end{aligned}$$

Line 4 comes from the fact that either $c \neq y$ or $\mathop{\mathbf{E}}_{h\sim\rho} \mathbf{I}(h(\mathbf{x}) = y) \leq \mathop{\mathbf{E}}_{h\sim\rho} \mathbf{I}(h(\mathbf{x}) = c)$.

When $Y = \{-1, +1\}$ it is trivial to prove: $\mathbf{R}_P(B_\rho) = \sum_{c=1}^{Q} \mathop{\mathbf{Pr}}_{(\mathbf{x},y)\sim P} \left(\mathcal{S}^\rho(c, (\mathbf{x}, y)) \leq 0\right) - 1$.

Lastly, in order to relax the ρ-margin, one solution is to consider the following loss.

Definition 4 (the ω-loss)**.** *Let P be a distribution over $X \times Y$, let $\mathcal{H}$ be a set of multiclass voters from X to Y, and let $\omega \in [0, 1]$ be a constant. For every distribution ρ on $\mathcal{H}$, we define the ω-loss, associated to ρ on an example $(\mathbf{x}, y) \sim P$, as the loss function $\ell^\rho(\omega, (\mathbf{x}, y))$:*

$$\ell^\rho(\omega, (\mathbf{x}, y)) = \mathbf{I}\left[\mathop{\mathbf{E}}_{h\sim\rho} \mathbf{I}(h(\mathbf{x}) = y) \leq \omega\right]. \tag{6}$$

The true value of the ω-loss of ρ on P is:

$$\begin{aligned} \ell_P^\rho(\omega) &= \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim P} \ell^\rho(\omega, (\mathbf{x}, y)) \\ &= \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim P} \mathbf{I}\left[\mathop{\mathbf{E}}_{h\sim\rho} \mathbf{I}(h(\mathbf{x}) = y) \leq \omega\right]. \end{aligned} \tag{7}$$

This loss can be seen as a linear relaxation of the ρ-margin. Moreover, for every distribution ρ on $\mathcal{H}$, the following theorem relates the risk of $B_\rho(\cdot)$ and the ω-loss associated to ρ.

Theorem 3. *Let $Q \geq 2$ be the quantity of classes. For every distribution P over $X \times Y$, for every example $(\mathbf{x}, y)$ i.i.d according to P, and for every distribution ρ over a set of multiclass voters $\mathcal{H}$, we have:*

$$\ell_P^\rho(\tfrac{1}{Q}) \leq \mathbf{R}_P(B_\rho) \leq \ell_P^\rho(\tfrac{1}{2}). \tag{8}$$

Proof. Firstly, we prove the left-hand side of Inequality (8):

$$\ell_P^\rho(\tfrac{1}{Q}) \leq \mathbf{R}_P(B_\rho).$$

In fact, we have:

$$\begin{aligned} \mathbf{R}_P(B_\rho) &= \mathop{\mathbf{Pr}}_{(\mathbf{x},y)\sim P} \left[\mathcal{M}^\rho(\mathbf{x}, y) \leq 0\right] \\ &= \mathop{\mathbf{Pr}}_{(\mathbf{x},y)\sim P} \left[\exists c \in Y, c \neq y : \mathop{\mathbf{E}}_{h\sim\rho} \mathbf{I}(h(\mathbf{x}) = y) \leq \mathop{\mathbf{E}}_{h\sim\rho} \mathbf{I}(h(\mathbf{x}) = c)\right] \\ &= \mathop{\mathbf{Pr}}_{(\mathbf{x},y)\sim P} \left[\mathop{\mathbf{E}}_{h\sim\rho} \mathbf{I}(h(\mathbf{x}) = y) \leq \max_{c\in Y, c\neq y} \mathop{\mathbf{E}}_{h\sim\rho} \mathbf{I}(h(\mathbf{x}) = c)\right] \\ &\geq \mathop{\mathbf{Pr}}_{(\mathbf{x},y)\sim P} \left[\mathop{\mathbf{E}}_{h\sim\rho} \mathbf{I}(h(\mathbf{x}) = y) \leq \frac{1 - \mathop{\mathbf{E}}_{h\sim\rho} \mathbf{I}(h(\mathbf{x}) = y)}{Q - 1}\right] \qquad (9) \\ &= \mathop{\mathbf{Pr}}_{(\mathbf{x},y)\sim P} \left[\mathop{\mathbf{E}}_{h\sim\rho} \mathbf{I}(h(\mathbf{x}) = y) \leq \frac{1}{Q}\right] \\ &= \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim P} \mathbf{I}\left[\mathop{\mathbf{E}}_{h\sim\rho} \mathbf{I}(h(\mathbf{x}) = y) \leq \frac{1}{Q}\right] \\ &= \ell_P^\rho(\tfrac{1}{Q}). \end{aligned}$$

Line (9) comes from :

$$\begin{aligned} \max_{c\in Y, c\neq y} \mathop{\mathbf{E}}_{h\sim\rho} \mathbf{I}(h(\mathbf{x}) = c) &\geq \mathop{\mathbf{E}}_{\substack{c\in Y \\ c\neq y}} \mathop{\mathbf{E}}_{h\sim\rho} \mathbf{I}(h(\mathbf{x}) = c) \\ &= \frac{1}{Q-1} \sum_{\substack{c=1 \\ c\neq y}}^{Q} \mathop{\mathbf{E}}_{h\sim\rho} \mathbf{I}(h(\mathbf{x}) = c) \\ &= \frac{1}{Q-1}\left[1 - \mathop{\mathbf{E}}_{h\sim\rho} \mathbf{I}(h(\mathbf{x}) = y)\right] \end{aligned}$$

Secondly, we prove the right-hand side of Inequality (8):

$$\mathbf{R}_P(B_\rho) \leq \ell_P^\rho(\tfrac{1}{2}).$$

This is verified if:

$$\begin{aligned} R(B_\rho) &= \mathop{\mathbf{Pr}}_{(\mathbf{x},y)\sim P} \left[\mathcal{M}^\rho(\mathbf{x},y) \leq 0\right] \\ &= \mathop{\mathbf{Pr}}_{(\mathbf{x},y)\sim P} \left[\mathop{\mathbf{E}}_{h\sim\rho} \mathbf{I}(h(\mathbf{x})=y) \leq \max_{c\in Y, c\neq y} \mathop{\mathbf{E}}_{h\sim\rho} \mathbf{I}(h(\mathbf{x})=c)\right] \\ &\leq \mathop{\mathbf{Pr}}_{(\mathbf{x},y)\sim P} \left[\mathop{\mathbf{E}}_{h\sim\rho} \mathbf{I}(h(\mathbf{x})=y) \leq \frac{1}{2}\right] \\ &= \ell_P^\rho(\tfrac{1}{2}). \end{aligned}$$

It is equivalent to prove that:

$$\forall (\mathbf{x},y)\sim P,\ \max_{c\in Y, c\neq y} \mathop{\mathbf{E}}_{h\sim\rho} \mathbf{I}(h(\mathbf{x})=c) \leq \frac{1}{2}.$$

(i) According to the definition of $B_\rho(\cdot)$, if $\mathop{\mathbf{E}}_{h\sim\rho} \mathbf{I}(h(\mathbf{x})=y) \leq 1/2$ for $(\mathbf{x},y)\sim P$, then $\mathcal{M}^\rho(\mathbf{x},y) \leq 0$.

(ii) Moreover, from the definition of the ω-loss, if we fix $\omega = \frac{1}{2}$, then we have:

$$\begin{aligned} \mathbf{R}_P(B_\rho) &= \mathop{\mathbf{Pr}}_{(\mathbf{x},y)\sim P} \left[\mathop{\mathbf{E}}_{h\sim\rho} \mathbf{I}(h(\mathbf{x})=y) \leq \max_{c\in Y, c\neq y} \mathop{\mathbf{E}}_{h\sim\rho} \mathbf{I}(h(\mathbf{x})=c)\right] \\ &\leq \mathop{\mathbf{Pr}}_{(\mathbf{x},y)\sim P} \left[\mathop{\mathbf{E}}_{h\sim\rho} \mathbf{I}(h(\mathbf{x})=y) \leq \frac{1}{2}\right]. \end{aligned}$$

The result comes from (i) and (ii). □

There exists a region of indecision when $\omega \in \left[\frac{1}{Q}, \frac{1}{2}\right]$ (see Figure 1) which implies that ω should be chosen carefully. Note that when $Y = \{-1, +1\}$, it is trivial to prove: $\mathbf{R}_P(B_\rho) = \ell_P^\rho(\frac{1}{2})$.

The three notions of margin presented above are then equivalent if we stand in the binary classification setting. However, they differ on the considered information.

- The ρ-margin is associated to the true decision function in multiclass and is independent from the true class of the example.
- The ρ-strength depends on the true class y of $\mathbf{x}$ and corresponds to a combination of the binary margin (one class versus another class) for $y' \neq y$.
- The ω-loss depends also on the true class y of $\mathbf{x}$, but it does not consider the other classes. This loss is a linear measure regarding to y, easier to manipulate, but implies a higher region of indecision (see Theorem 3).

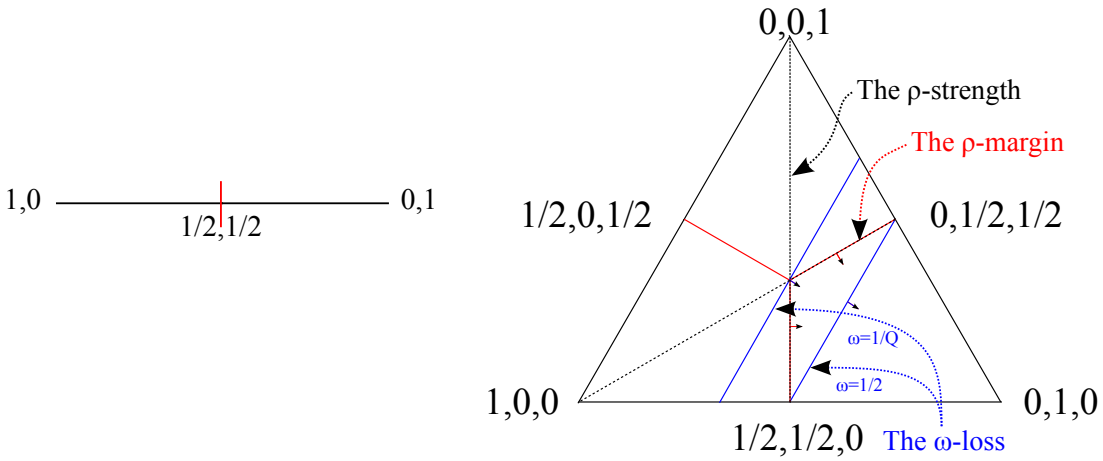
These properties are illustrated on Figure 1, and lead to the three generalizations of the C-bound presented bellow.

3.1.2 Generalizations of the C-bound in the Multiclass Setting

The following bound is based on the definition of ρ-margin in multiclass (Definition 2).

Theorem 4 (the multiclass C-bound)**.** *For every distribution ρ on a set of multiclass voters $\mathcal{H}$, and for every distribution P on $X \times Y$, such that $\mathcal{M}_P^\rho > 0$, we have:*

$$R(B_\rho) \ = \ \mathop{\mathbf{Pr}}_{(\mathbf{x},y)\sim P} \left(\mathcal{M}^\rho(\mathbf{x},y) \leq 0\right)$$



(a) The three margins are equivalent when $Y = \{-1, +1\}$. The coding of the class -1 is $(1, 0)$, the one of $+1$ is $(0, 1)$.

(b) The three margins when $Y = \{1, 2, 3\}$, and when the true class of $\mathbf{x}$ is 2. The coding of the class 1 is $(1, 0, 0)$, the one of the class 2 is $(0, 1, 0)$, and the one of the class 3 is $(0, 0, 1)$.

Figure 1: Given an example $(\mathbf{x}, y)$, we can represent the vote $B_\rho(\mathbf{x})$ by the convex combination with the barycentric coordinates where each angle corresponds to a class of $Y = \{1, \ldots, Q\}$. The coordinates of $B_\rho(\mathbf{x})$ are then $(\mathbf{E}_{h\sim\rho}\, \mathbf{I}(h(\mathbf{x}) = 1), \ldots, \mathbf{E}_{h\sim\rho}\, \mathbf{I}(h(\mathbf{x}) = Q))$.

$$\leq \frac{\mathbf{Var}_{(\mathbf{x},y)\sim P}\, \mathcal{M}^\rho(\mathbf{x}, y)}{\mathbf{E}_{(\mathbf{x},y)\sim P}\, \left(\mathcal{M}^\rho(\mathbf{x}, y)\right)^2}$$
$$= 1 - \frac{\left(\mathcal{M}_P^\rho\right)^2}{\mathcal{M}_P^{\rho^2}},$$

where $\mathcal{M}_P^\rho$ *and* $\mathcal{M}_P^{\rho^2}$ *are respectively the first and second statistical moments of the* ρ*-margin* $\mathcal{M}^\rho(\mathbf{x}, y)$*, and are defined by:*

$$\begin{aligned}
\mathcal{M}_P^\rho &= \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim D'} \mathcal{M}^\rho(\mathbf{x}, y) \\
&= \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim P} \left[\mathop{\mathbf{E}}_{h\sim\rho} \mathbf{I}(h(\mathbf{x}) = y) - \max_{c\in Y, c\neq y} \left(\mathop{\mathbf{E}}_{h\sim\rho} \mathbf{I}(h(\mathbf{x}) = c) \right) \right] \\
&= \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim P} \mathop{\mathbf{E}}_{h\sim\rho} \mathbf{I}(h(\mathbf{x}) = y) - \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim P} \max_{c\in Y, c\neq y} \left(\mathop{\mathbf{E}}_{h\sim\rho} \mathbf{I}(h(\mathbf{x}) = c) \right),
\end{aligned}$$

$$\begin{aligned}
\mathcal{M}_P^{\rho^2} &= \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim P} \left(\mathcal{M}^\rho(\mathbf{x}, y)\right)^2 \\
&= \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim P} \left[\mathop{\mathbf{E}}_{h\sim\rho} \mathbf{I}(h(\mathbf{x}) = y) - \max_{c\in Y, c\neq y} \left(\mathop{\mathbf{E}}_{h\sim\rho} \mathbf{I}(h(\mathbf{x}) = c) \right) \right]^2.
\end{aligned}$$

Proof. Same proof process than the binary C-bound (see Theorem 1) □

This bound offers an accurate relation between the risk of the majority vote and the margin without the number of classes Q. However, the term $\max_{c\in Y, c\neq y} \left(\mathop{\mathbf{E}}_{h\sim\rho} \mathbf{I}(h(\mathbf{x}) = c) \right)$ makes the derivation of an algorithm to minimize this bound harder than in binary classification.

We now present the approaches related to the relaxations stated in the previous section.

Thanks to the definition of the ρ-strength (Definition 3), and according to the proof process of the C-bound we obtain the following relation.

Theorem 5. *For every distribution ρ on a set of multiclass voters $\mathcal{H}$, and for every distribution P over $X \times Y$, such that $\forall c \in Y, \mathcal{S}_P^\rho(c) > 0$, we have:*

$$\begin{aligned} R(B_\rho) &\leq \sum_{c=1}^{Q} \mathop{\mathbf{Pr}}_{(\mathbf{x},y)\sim P} \left(\mathcal{S}^\rho(c,(\mathbf{x},y)) \leq 0\right) - 1 \\ &\leq \sum_{c=1}^{Q} \frac{\mathbf{Var}_{(\mathbf{x},y)\sim P}\left(\mathcal{S}^\rho(c,(\mathbf{x},y))\right)}{\mathbf{Var}_{(\mathbf{x},y)\sim D'}\left(\mathcal{S}^\rho(c,(\mathbf{x},y))\right) - \mathbf{E}_{(\mathbf{x},y)\sim P}\left(\mathcal{S}^\rho(c,(\mathbf{x},y))\right)^2} - 1 \\ &= \sum_{c=1}^{Q} \left(1 - \frac{\left(\mathcal{S}_P^\rho(c)\right)^2}{\mathcal{S}_P^{\rho^2}(c)}\right) - 1 \\ &= (Q-1) - \sum_{c=1}^{Q} \frac{\left(\mathcal{S}_P^\rho(c)\right)^2}{\mathcal{S}_P^{\rho^2}(c)}, \end{aligned}$$

where $\mathcal{S}_P^\rho(c)$ and $\mathcal{S}_P^{\rho^2}(c)$ are respectively the first and the second statistical moments of the ρ-strength of the class c, and are defined by:

$$\begin{aligned} \mathcal{S}_P^\rho(c) &= \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim P} \mathcal{S}^\rho(c,(\mathbf{x},y)) \\ &= \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim P} \left(\mathop{\mathbf{E}}_{h\sim\rho} \mathbf{I}(h(\mathbf{x}) = y) - \mathop{\mathbf{E}}_{h\sim\rho} \mathbf{I}(h(\mathbf{x}) = c)\right) \\ &= \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim P} \mathop{\mathbf{E}}_{h\sim\rho} \mathbf{I}(h(\mathbf{x}) = y) - \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim D'} \mathop{\mathbf{E}}_{h\sim\rho} \mathbf{I}(h(\mathbf{x}) = c) \\ \mathcal{S}_P^{\rho^2}(c) &= \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim P} \left(\mathcal{S}^\rho(c,(\mathbf{x},y))\right)^2 \\ &= \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim P} \left(\mathop{\mathbf{E}}_{h\sim\rho} \mathbf{I}(h(\mathbf{x}) = y) - \mathop{\mathbf{E}}_{h\sim\rho} \mathbf{I}(h(\mathbf{x}) = c)\right)^2 \end{aligned}$$

Proof. Comes for Inequality (5). □

This relation does not depend on the computation of a maximum, but explicitly depends on the quantity of classes Q. This result can be seen as a sum of C-bounds for every class. A practical drawbacks of this bound is then that we have to minimize Q binary C-bounds.

Finally, the bound related to the ω-loss is:

Theorem 6. *For every distribution ρ on a set of multiclass voters $\mathcal{H}$, and for every distribution P on $X \times Y$, if $\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim P}\left(\mathop{\mathbf{E}}_{h\sim\rho} \mathbf{I}(h(\mathbf{x}) = y) - \omega\right) > 0$, then:*

$$\begin{aligned} \ell_P^\rho(\omega) &\leq \frac{\mathbf{Var}_{(\mathbf{x},y)\sim P}\left(\mathop{\mathbf{E}}_{h\sim\rho} \mathbf{I}(h(\mathbf{x}) = y) - \omega\right)}{\mathbf{E}_{(\mathbf{x},y)\sim P}\left(\mathop{\mathbf{E}}_{h\sim\rho} \mathbf{I}(h(\mathbf{x}) = y) - \omega\right)^2} \\ &= 1 - \frac{\left[\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim P}\left(\mathop{\mathbf{E}}_{h\sim\rho} \mathbf{I}(h(\mathbf{x}) = y) - \omega\right)\right]^2}{\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim P}\left(\mathop{\mathbf{E}}_{h\sim\rho} \mathbf{I}(h(\mathbf{x}) = y) - \omega\right)^2}. \end{aligned} \tag{10}$$

Proof. According to Definition 4 of $\ell_P^\rho(\omega)$, we have:

$$\begin{aligned} \ell_P^\rho(\omega) &= \mathop{\mathbf{Pr}}_{(\mathbf{x},y)\sim P}\left[\mathop{\mathbf{E}}_{h\sim\rho} \mathbf{I}(h(\mathbf{x}) = y) \leq \omega\right] \\ &= \mathop{\mathbf{Pr}}_{(\mathbf{x},y)\sim P}\left[\mathop{\mathbf{E}}_{h\sim\rho} \mathbf{I}(h(\mathbf{x}) = y) - \omega \leq 0\right]. \end{aligned} \tag{11}$$

We apply the Cantelli-Chebyshev inequality on the line (11) on the random variable $\mathop{\mathbf{E}}_{h\sim\rho} \mathbf{I}(h(\mathbf{x}) = y)$. We obtain:

$$\ell_P^\rho(\omega) \leq \frac{\mathbf{Var}_{(\mathbf{x},y)\sim P}\left(\mathop{\mathbf{E}}_{h\sim\rho} \mathbf{I}(h(\mathbf{x}) = y) - \omega\right)}{\mathbf{E}_{(\mathbf{x},y)\sim P}\left(\mathop{\mathbf{E}}_{h\sim\rho} \mathbf{I}(h(\mathbf{x}) = y) - \omega\right)^2}.$$

Finally, since:

$$\begin{aligned}&\mathbf{Var}_{(\mathbf{x},y)\sim P}\left(\mathop{\mathbf{E}}_{h\sim\rho} \mathbf{I}(h(\mathbf{x}) = y) - \omega\right)\\ &= \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim P}\left(\mathop{\mathbf{E}}_{h\sim\rho} \mathbf{I}(h(\mathbf{x}) = y) - \omega\right)^2 - \left[\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim P}\left(\mathop{\mathbf{E}}_{h\sim\rho} \mathbf{I}(h(\mathbf{x}) = y) - \omega\right)\right]^2,\end{aligned}$$

we have:

$$\ell_P^\rho(\omega) \leq 1 - \frac{\left[\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim P}\left(\mathop{\mathbf{E}}_{h\sim\rho} \mathbf{I}(h(\mathbf{x}) = y) - \omega\right)\right]^2}{\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim P}\left(\mathop{\mathbf{E}}_{h\sim\rho} \mathbf{I}(h(\mathbf{x}) = y) - \omega\right)^2}.$$

□

Since the term $\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim P}\left(\mathop{\mathbf{E}}_{h\sim\rho} \mathbf{I}(h(\mathbf{x}) = y) - \omega\right)$ is linear, the derivation of an algorithm seems easier.

4 Conclusion and Perspectives

In this paper, we extend to the multiclass setting an important theoretical result in the PAC-Bayesian literature. Concretely, we prove three versions of the C-bound, a bound over the risk of the majority vote, based on three generalizations of the notion of margin in multiclass classification. These results opens the door to develop new algorithms for multiclass classification when we desire to learn a weighted majority vote over a set of multiclass voters.

References

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

Catoni, O. *PAC-Bayesian supervised classification: the thermodynamics of statistical learning*, volume 56. Institute of Mathematical Statistic, 2007.

Germain, P., Lacasse, A., Laviolette, F., and Marchand, M. PAC-Bayesian Learning of Linear Classifiers. In *Proceedings of International Conference on Machine Learning*, 2009.

Lacasse, A., Laviolette, F., Marchand, M., Germain, P., and Usunier, N. PAC-bayes bounds for the risk of the majority vote and the variance of the Gibbs classifier. In *Proceedings of annual conference on Neural Information Processing Systems*, 2007.

Langford, J. and Shawe-Taylor, J. PAC-bayes & margins. In *Proceedings of Annual Conference on Advances in Neural Information Processing Systems*, pp. 439–446. MIT Press, 2002.

Laviolette, F., Marchand, M., and Roy, J.-F. From PAC-Bayes Bounds to Quadratic Programs for Majority Votes. In *Proceedings of International Conference on Machine Learning*, June 2011.

McAllester, D. A. Simplified PAC-bayesian margin bounds. In *Proceedings of annual conference on Computational learning theory*, pp. 203–215, 2003.

Morvant, E., Koço, S., and Ralaivola, L. PAC-Bayesian Generalization Bound on Confusion Matrix for Multi-Class Classification. In *Proceedings of International Conference on Machine Learning*, pp. 815–822. Omnipress, 2012. (Full Paper, Acceptance rate: 27%).

Seeger, M. PAC-bayesian generalization error bounds for gaussian process classification. *Journal of Machine Learning Research*, 3:233–269, 2002.