Journal of Machine Learning Research 24 (2023) 1-13 Submitted 6/22; Revised 8/23; Published 9/23

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

Louis-Philippe Vignault louis-philippe.vignault.1@ulaval.ca
Audrey Durand audrey.durand@ift.ulaval.ca
Pascal Germain pascal.germain@ift.ulaval.ca
Department of Computer Science and Electrical Engineering
Université Laval

Editor: Pierre Alquier

Abstract

This work shows that the demonstration of Proposition 15 of Germain et al. (2015) is flawed and the proposition is false in a general setting. This proposition gave an inequality that upper-bounds the variance of the margin of a weighted majority vote classifier. Even though this flaw has little impact on the validity of the other results presented in Germain et al. (2015), correcting it leads to a deeper understanding of the $\mathcal{C}$-bound, which is a key inequality that upper-bounds the risk of a majority vote classifier by the moments of its margin, and to a new result, namely a lower-bound on the $\mathcal{C}$-bound. Notably, Germain et al.'s statement that "the $\mathcal{C}$-bound can be arbitrarily small" is invalid in presence of irreducible error in learning problems with label noise. In this erratum, we pinpoint the mistake present in the demonstration of the said proposition, we give a corrected version of the proposition, and we propose a new theoretical lower bound on the $\mathcal{C}$-bound.

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

1. Introduction

Germain et al. (2015) synthetize several papers on the PAC-Bayesian analysis of majority vote classifiers. A pivotal element of their analysis is a bound on the majority vote for binary classification linking the risk of such predictors to the first and second moments of the majority vote's *margins*. This result, coined as the $\mathcal{C}$-bound by Lacasse et al. (2006), is obtained through the one-sided Chebyshev inequality.

Section 4.3 of Germain et al. (2015) introduces formally the $\mathcal{C}$-bound as Theorem 11, and many of its mathematical properties are demonstrated in Section 4.4 therein. Upon further inspection, we have found one such property to be incorrect in general, namely, the claim that "The $\mathcal{C}$-bound can be arbitrarily small, even for large Gibbs risks" (Germain et al., 2015). The said property was derived from their Proposition 15 which had a flaw in its demonstration. In this erratum, after recalling the notation and definitions, we first present Proposition 15 of Germain et al. (2015) and point out the error in its demonstration. We then provide a concrete example where the result given by Proposition 15 of Germain et al. (2015) does not hold. We proceed to demonstrate that Proposition 15 holds in the specific case where the classification problem is devoid of label noise. Finally, we propose a new and more general alternative result which consists in an interesting property of the $\mathcal{C}$-bound in a general setting, namely that it is lower bounded by the expected value of the variance of the labels.

It is important to mention that in Germain et al. (2015) the said Proposition 15 was only used in the demonstration of Corollary 16 and that these results are not referenced again thereafter. Therefore, the current erratum does not affect the validity of other results presented in Germain et al. (2015), notably the PAC-Bayes guarantees and learning algorithms sourced from the $\mathcal{C}$-bound. Also, even if several subsequent papers cite or exploit the $\mathcal{C}$-bound, very few have mentioned the erroneous result. Up to our knowledge, it has been solely mentioned by Segev et al. (2017) and Du and Swamy (2019). We have found that the results presented in these two works remain valid despite the mistake we address in this erratum.

2. Notation and Definitions

Throughout this paper, we employ the same notation and definitions used by Germain et al. (2015) unless specified otherwise. That is, we consider a binary classification problem in which each example (x, y) consists of an input x and output y belonging to the spaces $\mathcal{X}$ and $\mathcal{Y} = \{-1, 1\}$ respectively. The examples are sampled i.i.d. according to a true unknown distribution D over $\mathcal{X} \times \{-1, 1\}$. We consider $\mathcal{H}$, a finite set of functions $f : \mathcal{X} \to [-1, 1]$. Each such f defines a "voter" in the context of majority vote rules. One may see the output of a given voter f as a measurement of the confidence of the voter toward a given label. For example, say we have two voters $f_1(x) = 0.1$ and $f_2(x) = 0.9$, then both voters lean towards the label 1, but the former less confidently then the latter.

A weighted majority vote is defined by a distribution Q over $\mathcal{H}$. Such a distribution defines the weight given to each voter, denoted $Q(f)$, which gives rise to the majority vote classifier $B_Q(x) = \operatorname{sgn}[\mathbf{E}_{f\sim Q}\, f(x)]$. This majority vote is also called the *Bayes classifier*[1] in the PAC-Bayesian literature, as opposed to the stochastic *Gibbs Classifier* G_Q; with $\Pr(G_Q(x){=}y) = \frac{1}{2}(1 + y\,\mathbf{E}_{f\sim Q}\, f(x))$ for $y \in \{-1, 1\}$. The probability that the majority vote will make an incorrect prediction of the output for a given example drawn from distribution D is called the risk of the majority vote, or Bayes risk[2], and is denoted $R_D(B_Q)$:

$$R_D(B_Q) = \Pr_{(x,y)\sim D}(B_Q(x) \neq y) = \mathop{\mathbf{E}}_{(x,y)\sim D} I\left(\mathop{\mathbf{E}}_{f\sim Q} y\, f(x) \leq 0\right),$$

where $I(\cdot)$ is the indicator function taking value 0 if the input is false and 1 otherwise. Correspondingly, the probability that the Gibbs classifier makes an error is called the Gibbs risk and is denoted $R_D(G_Q)$:

$$R_D(G_Q) = \Pr_{(x,y)\sim D}(G_Q(x) \neq y) = \mathop{\mathbf{E}}_{(x,y)\sim D}\left(\frac{1}{2} - \frac{1}{2}\mathop{\mathbf{E}}_{f\sim Q} y\, f(x)\right).$$

[1] The term *Bayes classifier*, used by Germain et al. (2015), is established in the PAC-Bayesian literature. It refers to a specific type of majority vote and is not to be confused with the prominent term *Bayes optimal classifier*.

[2] Similarly, the term *Bayes risk* is not to be confused with the more common *Bayes error* which refers to the error made by the *Bayes optimal classifier*.

The term $\mathbf{E}_{f\sim Q}\, y\, f(x)$ shared by both Bayes risk and Gibbs risk definitions is a commonly occurring expression in the context of majority votes. It is referred to as the *margin* of the majority vote and is denoted $M_Q(x,y)$. Given that $(x,y)\sim D$, Germain et al. (2015) study the margin as a random variable which they denote M_Q^D. A longstanding intermediate result in PAC-Bayes analyses is the *factor two bound* between the Bayes Risk and the Gibbs risk (Langford and Shawe-Taylor, 2002; McAllester, 2003). That is, for fixed distributions D and Q, we have $R_D(B_Q) \leq 2\, R_D(G_Q)$. The $\mathcal{C}$-bound (Lacasse et al., 2006; Germain et al., 2015) is a finer upper bound on the Bayes risk and is defined using the variance and first two moments of the margin M_Q^D (denoted $\mu_1(M_Q^D)$ and $\mu_2(M_Q^D)$ respectively, with $\mu_i(M_Q^D) = \mathbf{E}_{(x,y)\sim D}[M_Q(x,y)]^i$). Equivalently, the $\mathcal{C}$-bound can be expressed in terms of the Gibbs risk $R_D(G_Q)$ and the expected disagreement (denoted d_Q^D and defined below):

$$R_D(B_Q) \leq \mathcal{C}_Q^D = \frac{\mathbf{Var}(M_Q^D)}{\mu_2(M_Q^D)} = 1 - \frac{\left(\mu_1(M_Q^D)\right)^2}{\mu_2(M_Q^D)} = 1 - \frac{(1-2R_D(G_Q))^2}{1-2d_Q^D},$$

where

$$\mu_1(M_Q^D) = \mathop{\mathbf{E}}_{(x,y)\sim D}\, \mathop{\mathbf{E}}_{f\sim Q}\, y\, f(x)\,, \qquad \mu_2(M_Q^D) = \mathop{\mathbf{E}}_{x\sim D_{\mathcal{X}}} \left[\mathop{\mathbf{E}}_{f\sim Q} f(x)\right]^2,$$

$$\mathbf{Var}(M_Q^D) = \mu_2(M_Q^D) - \left(\mu_1(M_Q^D)\right)^2, \qquad d_Q^D = \frac{1}{2}\left(1 - \mu_2(M_Q^D)\right).$$

In these definitions, $D_{\mathcal{X}}$ denotes the marginal distribution on $\mathcal{X}$ for a given distribution D.[1] We adopt this notation throughout the paper, as well as the use of $D_{\mathcal{Y}|x}$ to denote the conditional distribution on the output space $\mathcal{Y}$ given a fixed input $x \in \mathcal{X}$.

An important note has to be made about the notions of *stochasticity* and *determinism* as they appear in the original paper. The generic use of these terms can be confusing to the reader as it is sometimes unclear what is meant to be stochastic or deterministic in some passages. To alleviate this ambiguity we differentiate three distinct sources of stochasticity that are of importance in the studied framework.

Label noise. The stochasticity of the labels arises when for a given value of $x \sim D_{\mathcal{X}}$, there is a non-zero probability of observing more than one value of $y \sim D_{\mathcal{Y}|x}$.

Stochastic voters. Of note, Germain et al. (2015) introduced *voters* as *functions* of the form $f : \mathcal{X} \to [-1,1]$, implying they were deterministic predictors. Nevertheless, the property that a given voter is deterministic remains unused in their analysis of majority votes. We may wish to consider stochastic voters, such as voters for which the output follows some distribution. Say we consider a stochastic voter f^* such that for any given input x, the output of the voter is a random variable $f^*(x) \sim W_x$, where W_x is a distribution over $[-1,1]$ that depends of the value x. To apply the work of Germain et al. (2015), one needs only to consider $f(x)$ as the *expected output* of the stochastic voter f^* given by $f(x) = \mathbf{E}_{f^*(x)\sim W_x}\, f^*(x)$. Doing so, the value $f(x)$ will be deterministic. Moreover, this is coherent with the idea that $f(x)$ is the confidence of

[1] The formula given above for μ_2 is the result of an easy calculation (cf. Germain et al., 2015, Eq. (8))

the voter towards a given label. The discussion in this erratum henceforth encompasses both deterministic and stochastic voters and we use the general term *voter* to refer to voters of both natures.

Stochastic majority vote. We specified that the Bayes classifier B_Q is deterministic while the Gibbs classifier G_Q is stochastic. This distinction, made by Germain et al. (2015), is meant to distinguish between the stochasticity that arises from sampling a single voter (Gibbs) as opposed to averaging over the value of each voter $f \in \mathcal{H}$ (Bayes). It is not meant to refer to the nature of the individual voters as we remind that both Germain et al. (2015) and this erratum consider deterministic voters only.

We remind that this erratum is valid when considering the deterministic majority vote classifier that embed the expected value of its voters for the Bayes risk of Equation (1). The interested reader may note that the recent work of Zantedeschi et al. (2021) provide PAC-Bayes bounds for stochastic majority votes where the voters weights are random variables. While this is a related line of work, it differs from ours and Germain et al. (2015) as we only consider majority votes for which the weight of each voter is fixed.

3. The Mistake

Let us recall the statement and proof of the problematic proposition from Germain et al. (2015). The structure of the following proof is the same as in the original paper. However, we chose to present each mathematical argument clearly according to the explanations given by Germain et al. (2015) following their proof as many details are absent from the proof itself. We believe this improves both the clarity of the proof and the subsequent discussion of its flaw. Note that a similar result, equally erroneous, is stated without proof in Lacasse et al. (2006, Proposition 3).

Erroneous result

Proposition 1 (Proposition 15 in Germain et al. (2015)) *For any countable set of voters $\mathcal{H}$, any distribution Q on $\mathcal{H}$, and any distribution D on $\mathcal{X} \times \{-1, 1\}$, we have*

$$\mathbf{Var}(M_Q^D) \leq \sum_{f\in\mathcal{H}} Q^2(f) + \sum_{f_1\in\mathcal{H}} \sum_{\substack{f_2\in\mathcal{H}\\ f_2\neq f_1}} Q(f_1)Q(f_2) \underset{(x,y)\sim D}{\mathbf{Cov}}(f_1(x), f_2(x)).$$

Proof We begin with the definition of the margin, which can be seen as a sum of random variables, and develop the variance of the sum into the sum of covariances:

$$\begin{aligned}\mathbf{Var}\left(M_Q^D\right) &= \underset{(x,y)\sim D}{\mathbf{Var}}\left(M_Q(x,y)\right)\\ &= \underset{(x,y)\sim D}{\mathbf{Var}}\left(\sum_{f\in\mathcal{H}} Q(f)yf(x)\right)\end{aligned}$$

Erroneous result

$$
\begin{aligned}
&= \sum_{f\in\mathcal{H}} Q^2(f) \underset{(x,y)\sim D}{\mathbf{Var}} yf(x) + \sum_{f_1\in\mathcal{H}} \sum_{\substack{f_2\in\mathcal{H}\\ f_2\neq f_1}} Q(f_1)Q(f_2) \underset{(x,y)\sim D}{\mathbf{Cov}} (yf_1(x), yf_2(x))\\
&= \sum_{f\in\mathcal{H}} Q^2(f) \underset{(x,y)\sim D}{\mathbf{Var}} yf(x) + \sum_{f_1\in\mathcal{H}} \sum_{\substack{f_2\in\mathcal{H}\\ f_2\neq f_1}} Q(f_1)Q(f_2) \underset{(x,y)\sim D}{\mathbf{Cov}} (f_1(x), f_2(x))\\
&\leq \sum_{f\in\mathcal{H}} Q^2(f) + \sum_{f_1\in\mathcal{H}} \sum_{\substack{f_2\in\mathcal{H}\\ f_2\neq f_1}} Q(f_1)Q(f_2) \underset{(x,y)\sim D}{\mathbf{Cov}} (f_1(x), f_2(x)).
\end{aligned}
$$

The inequality is given by the fact that the definitions of y and $f(x)$ implie $yf(x) \in [-1, 1]$, from which we infer that $\underset{(x,y)\sim D}{\mathbf{Var}} yf(x) \leq 1$ for all voters $f \in \mathcal{H}$. ■

The problem in this demonstration arises when we simplify $\mathbf{Cov}_{(x,y)\sim D}\,(yf_1(x), yf_2(x))$ to $\mathbf{Cov}_{(x,y)\sim D}\,(f_1(x), f_2(x))$, which could also be written as $\mathbf{Cov}_{x\sim D_\mathcal{X}}\,(f_1(x), f_2(x))$. As this final writing is independent of y, this equality would imply that the value y has no effect in the expression $\mathbf{Cov}_{(x,y)\sim D}\,(yf_1(x), yf_2(x))$. This simplification does not stand in the general case as it is possible to find various examples of functions f_i and distributions D that do not satisfy this equality. As an example, consider $f_1(x) = 1$, $f_2(x) = -1$ and $\Pr_{y\sim D_{\mathcal{Y}|x}}(y{=}1) = \Pr_{y\sim D_{\mathcal{Y}|x}}(y{=}-1) = \frac{1}{2}$ for any given value $x \in \mathcal{X}$. Using these values, we get the following, which disproves the equality assumed in the said simplification:

$$
\begin{aligned}
\mathbf{Cov}(f_1(x), f_2(x)) &= \mathbf{Cov}(1, -1) = 0\\
\mathbf{Cov}(yf_1(x), yf_2(x)) &= \mathbf{Cov}(y, -y) = -\mathbf{Cov}(y, y) = -1.
\end{aligned}
$$

4. Falsifying Proposition 1

Even though the demonstration of Proposition 1 is flawed, this does not necessarily imply that the proposition itself does not stand. In principle, the proposition might be valid due to a different argument. However, in order to show that Proposition 15 of Germain et al. (2015) is indeed false, we introduce the following corollary. If Proposition 1 were true, then so would Corollary 2 which follows directly from it.

Erroneous result

Corollary 2 (Corollary 16 in Germain et al. (2015)) *Given n independent voters under a uniform distribution Q, we have*

$$R_D(B_Q) \leq \mathcal{C}_Q^D \leq \frac{1}{n(1-2d_Q^D)} = \frac{1}{n(1-2R_D(G_Q))}.$$

To prove the above corollary, Germain et al. (2015) note that under the hypothesis that the n voters are independents under a uniform distribution, Proposition 1 implies that $\mathbf{Var}(M_Q^D) \leq 1/n$; the result then follows from the definition of the $\mathcal{C}$-bound.

At first glance, the corollary seems intuitive. Increasing the number of voters should improve the quality of the classification. Therefore, we would expect the $\mathcal{C}$-bound to decrease. The main problem with this result is that it implies that as the number of independent voters n goes to infinity, the right side goes to 0. This means that for any binary classification problem, the risk of the majority vote can be made arbitrarily small by using a sufficiently large set of independent voters. Though this may seem like a desirable property, it is not realistic in the presence of label noise. In such case, the noise associated with the output y given a certain x will induce a minimum amount of erroneous classification (*i.e.*, irreducible error). Notice that the label noise is an inherent property of a given distribution D, and the irreducible error (also known as *Bayes error* in the literature) is achieved by an optimal prediction rule for D. This irreducible error is not 0 in the presence of label noise, which contradicts the statement of Corollary 2. To clearly illustrate this contradiction, consider the following example.

Counter-example. Suppose a binary classification problem where each example $(x, y){\sim}D$ is such that x is drawn from the uniform distribution over a one dimensional input space $\mathcal{X} = [0, 1]$, and the output is either $y = 1$ with probability $p = 0.8$ or $y = -1$ with probability $1 - p = 0.2$ (that is, y is independent of x). Let us now consider a majority vote classifier made from the set of deterministic voters $\mathcal{H} := \{f_i(x)\}$ for $i \in \{1, 2, ..., n\}$ such that $f_i(x) = 1$ if the i^{th} decimal of x is greater or equal than 2, and $f_i(x) = -1$ otherwise. As x follows a uniform distribution over $[0, 1]$, every decimal value will be taken from $0, 1, 2, ..., 9$ with probability 0.1 for each value. Furthermore, the value of every decimal positions are independent from one another, rendering the output of each voter mutually independent. Finally, each voter $f_i \in \mathcal{H}$ is given an equal weight of $Q(f_i) = 1/n$.

According to Corollary 2, this majority vote should yield a $\mathcal{C}$-bound that decreases asymptotically to 0 as the number of voters n increases. Because $R_D(B_Q) \leq \mathcal{C}_Q^D$, this implies that given a large enough amount of voters, it would be possible to predict with almost certainty the noisy label y, which is uncorrelated to both the input x and the voters. However, this result is clearly incorrect, as there is always a minimal probability of $2/10$ to make a false prediction, *i.e.*, $R_D(B_Q) \geq 0.2$. Moreover, given the construction of data distribution and the voters, we obtain from the definition of the $\mathcal{C}$-bound $\mathcal{C}_Q^D = \frac{400+144n}{400+225n}$. It is easy to see that the value $\mathcal{C}_Q^D$ does not go to 0 as n grows as we have $\mathcal{C}_Q^D > 0.64$ for any value $n > 0$. Thus, the $\mathcal{C}$-bound does not possess the property stated by Corollary 2.

In other words, the result of said corollary is incorrect as we have now given a counterexample. Since the only argument in the demonstration of the corollary is the use of Proposition 1, this implies that Proposition 1 is also incorrect.

5. Proposition 1 in a Setting Devoid of Label Noise

We have now shown that Proposition 1 is incorrect as stated in the general case, namely when the labelling task includes label noise. The aim of the current section is to rectify Proposition 1 in a setting devoid of label noise, by introducing new assumptions on the data distribution and the voters. In the following section we build upon this knowledge a new, correct and improved result that stands in the general case.

Even though the assumption that a given problem is devoid of label noise might not be realistic in most applications, we can see this case as an easy classification problem where an error of 0 should be achievable for some classifier. Given this assumption, we would thus desire a majority vote made with "good enough" voters to come arbitrarily close to null classification error. Moreover, we would expect a good bound on the risk of majority votes to reflect this fact by being potentially arbitrarily small in this setting.

We begin by noting that the assumption that distribution D is devoid of label noise implies that there exists a function that maps each input x to its correct label.

Assumption 1: Deterministic labels. The data distribution D is such that there exists a function $g : \mathcal{X} \to \{-1, 1\}$ for which $y = g(x)$ for all $(x, y) \sim D$, or equivalently $\Pr_{y\sim D_{\mathcal{Y}|x}}\big(y{=}g(x)\big) = 1$ for all $x \in \mathcal{X}$.

The counter-example of the previous section also illustrates that uncorrelated voters do not necessarily result in an accurate majority vote, as it does not imply that the input-output correlation is taken into account. Thus, we need a second assumption, capturing the fact that the majority vote must be made of "good enough" voters. Namely, we want each voter to classify this input correctly with a probability larger than 1/2.

Assumption 2: Sufficiently accurate voters. For any given pair $(x, y) \sim D$, for any voter $f_i \in \mathcal{H}$, $yf_i(x) > 0$; in other words, all voters are leaning towards the right label (although with scores, i.e., varying confidence levels) in a non ambivalent manner (i.e., $f_i(x) \neq 0$).

Given the definition of a voter f_i, Assumption 2 implies that for any given input x in $\mathcal{X}$ and any given voter f_i in $\mathcal{H}$, the expected output of $f_i(x)$ will agree with the true value of y, meaning they will share the same sign. This assumption is very strong, as it will break when a single voter does not classify a given x with sufficient probability. It may seem trivial that under this assumption the majority vote will be correct every time and will thus have an actual risk of 0. Recall however that the majority vote risk $R_D(B_Q)$ only considers the sign of the margin $[y\,\mathbf{E}_{f\sim Q}\,f(x)]$ on $(x, y) \sim D$ while the computation of the $\mathcal{C}$-bound depends on the value of the margin itself. In other words, the risk $R_D(B_Q)$ depends on which label the majority vote is most confident in but the $\mathcal{C}$-bound further considers how confident the vote is. This implies that various majority votes which make the same prediction (same risk $R_D(B_Q)$) may have different $\mathcal{C}$-bound as they may have different margins. Therefore, the $\mathcal{C}$-bound is not trivially 0 under the above assumptions.

We now show that an updated Proposition 1 holds given these two new assumptions, namely that the labels are deterministic and the voters are sufficiently accurate.

Proposition 3 (Proposition 15 in Germain et al. (2015) corrected) *For any countable set of voters $\mathcal{H}$, any distribution Q on $\mathcal{H}$, and any distribution D on $\mathcal{X} \times \{-1, 1\}$ such that there exists a function $g : \mathcal{X} \to \{-1, 1\}$ for which $y = g(x)$ and $yf(x) > 0$ for all $(x, y) \sim D$ and $f \in \mathcal{H}$, we have*

$$\mathbf{Var}(M_Q^D) \leq \sum_{f \in \mathcal{H}} Q^2(f) + \sum_{f_1 \in \mathcal{H}} \sum_{\substack{f_2 \in \mathcal{H} \\ f_2 \neq f_1}} Q(f_1) Q(f_2) \mathop{\mathbf{Cov}}_{(x,y)\sim D} (f_1(x), f_2(x)) \,.$$

Proof Let $g(x)$ be the labelling function assigning each value x to its true, and noiseless, label y and let $f_i(x)$ be any voter taken from $\mathcal{H}$. Since $yf_i(x) > 0$ for any voter $f_i \in \mathcal{H}$ and for any given x, we have that the true labelling function $g(x)$ and the voter $f_i(x)$ are of the same sign. Thus, we always have $g(x)f_i(x) \geq f_i(x) \geq -g(x)f_i(x)$, which gives us the following inequality:

$$\begin{aligned}
& \mathop{\mathbf{E}}_{x\sim D_{\mathcal{X}}} (g(x)f_i(x)) \geq \mathop{\mathbf{E}}_{x\sim D_{\mathcal{X}}} (f_i(x)) \geq - \mathop{\mathbf{E}}_{x\sim D_{\mathcal{X}}} (g(x)f_i(x)) \\
\iff \quad & \mathop{\mathbf{E}}_{x\sim D_{\mathcal{X}}} (g(x)f_i(x)) \geq \left| \mathop{\mathbf{E}}_{x\sim D_{\mathcal{X}}} (f_i(x)) \right| \\
\implies \quad & \left[\mathop{\mathbf{E}}_{x\sim D_{\mathcal{X}}} (g(x)f_i(x)) \right]^2 \geq \left[\mathop{\mathbf{E}}_{x\sim D_{\mathcal{X}}} (f_i(x)) \right]^2 .
\end{aligned}$$

Since this inequality is true for each voter in $\mathcal{H}$, we consider the previous inequality using the functions f_i and f_j. Using the fact that for values $a, b, c, d > 0$ we have the implication $a < c$, $b < d \Rightarrow ab < cd$, we get the following:

$$\begin{aligned}
& \left[\mathop{\mathbf{E}}_{x\sim D_{\mathcal{X}}} (g(x)f_i(x)) \right]^2 \left[\mathop{\mathbf{E}}_{x\sim D_{\mathcal{X}}} (g(x)f_j(x)) \right]^2 \geq \left[\mathop{\mathbf{E}}_{x\sim D_{\mathcal{X}}} (f_i(x)) \right]^2 \left[\mathop{\mathbf{E}}_{x\sim D_{\mathcal{X}}} (f_j(x)) \right]^2 \\
\implies \quad & \mathop{\mathbf{E}}_{x\sim D_{\mathcal{X}}} (g(x)f_i(x)) \mathop{\mathbf{E}}_{x\sim D_{\mathcal{X}}} (g(x)f_j(x)) \geq \quad \mathop{\mathbf{E}}_{x\sim D_{\mathcal{X}}} (f_i(x)) \mathop{\mathbf{E}}_{x\sim D_{\mathcal{X}}} (f_j(x)).
\end{aligned}$$

Since $g(x)f_i(x) > 0$ for any x, we have that both terms on the left side of the inequality are positive prior to being squared. Knowing this guarantees that the inequality still holds as we take the square root on both sides. This inequality can then be used as follows:

$$\begin{aligned}
& \mathop{\mathbf{Cov}}_{(x,y)\sim D} (yf_i(x), yf_j(x)) \\
& = \mathop{\mathbf{E}}_{(x,y)\sim D} (y^2 f_i(x) f_j(x)) - \mathop{\mathbf{E}}_{(x,y)\sim D} (yf_i(x)) \mathop{\mathbf{E}}_{(x,y)\sim D} (yf_j(x)) \\
& = \mathop{\mathbf{E}}_{(x,y)\sim D} (1 f_i(x) f_j(x)) - \mathop{\mathbf{E}}_{(x,y)\sim D} (g(x)f_i(x)) \mathop{\mathbf{E}}_{(x,y)\sim D} (g(x)f_j(x)) \\
& \leq \mathop{\mathbf{E}}_{x\sim D_{\mathcal{X}}} (f_i(x) f_j(x)) - \mathop{\mathbf{E}}_{x\sim D_{\mathcal{X}}} (f_i(x)) \mathop{\mathbf{E}}_{x\sim D_{\mathcal{X}}} (f_j(x)) \\
& = \mathop{\mathbf{Cov}}_{x\sim D_{\mathcal{X}}} (f_i(x), f_j(x)) \\
& = \mathop{\mathbf{Cov}}_{(x,y)\sim D} (f_i(x), f_j(x)).
\end{aligned}$$

The last equality only serves to highlight that since y is absent from both $f_i(x)$ and $f_j(x)$, computing the covariance according to either the joint distribution D or the marginal distribution $D_{\mathcal{X}}$ does not affect the result. We have now proven that under our new set of hypothesis, the inequality $\mathbf{Cov}_{(x,y)\sim D}\left(yf_1(x), yf_2(x)\right) \leq \mathbf{Cov}_{(x,y)\sim D}\left(f_1(x), f_2(x)\right)$ always stands. We note that the original proof erroneously assumed this to be an equality. Nonetheless, we can apply this new inequality to the original proof as follows:

$$\begin{aligned}
\underset{(x,y)\sim D}{\mathbf{Var}}\left(M_Q(x,y)\right) &= \underset{(x,y)\sim D}{\mathbf{Var}}\left(\sum_{f\in\mathcal{H}} Q(f)yf(x)\right)\\
&= \sum_{f\in\mathcal{H}} Q^2(f)\underset{(x,y)\sim D}{\mathbf{Var}} yf(x) + \sum_{f_1\in\mathcal{H}}\sum_{\substack{f_2\in\mathcal{H}\\ f_2\neq f_1}} Q(f_1)Q(f_2)\underset{(x,y)\sim D}{\mathbf{Cov}}(yf_1(x), yf_2(x))\\
&\leq \sum_{f\in\mathcal{H}} Q^2(f)\underset{(x,y)\sim D}{\mathbf{Var}} yf(x) + \sum_{f_1\in\mathcal{H}}\sum_{\substack{f_2\in\mathcal{H}\\ f_2\neq f_1}} Q(f_1)Q(f_2)\underset{(x,y)\sim D}{\mathbf{Cov}}(f_1(x), f_2(x))\\
&\leq \sum_{f\in\mathcal{H}} Q^2(f) + \sum_{f_1\in\mathcal{H}}\sum_{\substack{f_2\in\mathcal{H}\\ f_2\neq f_1}} Q(f_1)Q(f_2)\underset{(x,y)\sim D}{\mathbf{Cov}}(f_1(x), f_2(x)).
\end{aligned}$$

The final inequality is given by the fact that $\underset{(x,y)\sim D}{\mathbf{Var}} yf(x) \leq 1$ for all voters $f \in \mathcal{H}$. ■

We first point out that under these additional assumptions, Corollary 2 is also valid as its proof is a direct consequence of Proposition 1. This implies that the risk of the majority vote can be made arbitrarily small by increasing the amount of "good enough" voters, which as stated previously, is not trivially true. Again, this proposition is intended as a confirmation that the majority vote and $\mathcal{C}$-bound behave well under the right (though very strong) assumptions and not as a practical result.

6. Optimal Value of the C-Bound

The counter-example given in Section 4 was based on the knowledge that the risk of the majority vote should never be inferior to minimal risk induced by the label noise in D. We have shown in Section 5 that the $\mathcal{C}$-bound can converge to the minimal risk in a setting deprived of label noise by proving that in such a setting, using voters of sufficient quality renders Corrolary 2 true. In this section, we aim to generalise this result to problems which contain label noise. To do so, we introduce in Proposition 4 a theoretical lower-bound for the $\mathcal{C}$-bound for a given distribution D as well as the majority vote which achieves this bound.

Proposition 4 *For any distribution D, there exist a majority vote B_Q such that*

$$R_D(B_Q) \leq \mathcal{C}_Q^D = \underset{x\sim D_{\mathcal{X}}}{\mathbf{E}}\left(\underset{y\sim D_{\mathcal{Y}|x}}{\mathbf{Var}}(y)\right).$$

Moreover, there is no majority vote which can produce a lower $\mathcal{C}$-bound.

Proof Let $h(x) = \mathbf{E}_{y \sim D_{\mathcal{Y}|x}}(y)$ and consider a majority vote B_Q defined by a set of voters $\mathcal{H}$ and a distribution Q over this set. The score given to any input x by said majority vote B_Q is given by $\sum_{f \in \mathcal{H}} Q(f) f(x)$. We first notice that this majority vote is equivalent to a majority vote comprised of a single voter $k(x) = \sum_{f \in \mathcal{H}} Q(f) f(x)$ as the margin of both majority vote would follow the same distribution. To see this, let $M_{Q'}(x, y)$ be the margin function of this new majority vote where Q' is a dirac distribution on the single voter $k(x)$. We then have

$$M_Q(x,y) = \mathop{\mathbf{E}}_{f\sim Q}[yf(x)] = y \mathop{\mathbf{E}}_{f\sim Q}[f(x)] = yk(x) = \mathop{\mathbf{E}}_{k\sim Q'}[yk(x)] = M_{Q'}(x,y)\,.$$

Since both functions $M_Q(x, y)$ and $M_{Q'}(x, y)$ are equal for any value (x, y), both margins ($M^D_{Q'}$ and M^D_Q) will have the same distribution. We remind that the $\mathcal{C}$-bound can be defined using only the moments of its margin. Therefore, as both majority votes lead to the same margin, they will possess the same $\mathcal{C}$-bound. We therefore adopt this equivalent majority vote which uses a single voter $k(x)$ and the margin M^D_Q for the remainder of the proof. We then consider the definition of the $\mathcal{C}$-bound and proceed as follows (recall that $y \in \{-1, 1\}$, and therefore $y^2 = 1$):

$$\begin{aligned}
\mathcal{C}^D_Q &= 1 - \frac{\Big(\mu_1(M^D_Q)\Big)^2}{\mu_2(M^D_Q)} \\
&= 1 - \frac{\Big(\mathop{\mathbf{E}}_{(x,y)\sim D}(yk(x))\Big)^2}{\mathop{\mathbf{E}}_{(x,y)\sim D}(y^2k(x)^2)} \\
&= 1 - \frac{\Big(\mathop{\mathbf{E}}_{(x,y)\sim D}(k(x)h(x))\Big)^2}{\mathop{\mathbf{E}}_{(x,y)\sim D}(k(x)^2)} \\
&\geq 1 - \frac{\mathop{\mathbf{E}}_{(x,y)\sim D}(k(x)^2)\mathop{\mathbf{E}}_{(x,y)\sim D}(h(x)^2)}{\mathop{\mathbf{E}}_{(x,y)\sim D}(k(x)^2)} \\
&= 1 - \mathop{\mathbf{E}}_{(x,y)\sim D}(h(x)^2)\,.
\end{aligned}$$

The above inequality is given by the Cauchy-Schwarz inequality. We can easily see that by choosing k such that $k(x) = h(x)$ for all x, the inequality becomes an equality. This means that the value of the $\mathcal{C}$-bound for a given distribution D is lower-bounded by

$1-\mathbf{E}_{(x,y)\sim D}(h(x)^2)$, which is always achievable. We can simplify this expression as follows:

$$\begin{aligned}
\mathcal{C}_Q^D &= 1 - \mathop{\mathbf{E}}_{(x,y)\sim D}(h(x)^2)\\
&= 1 - \mathop{\mathbf{E}}_{(x,y)\sim D}\left[\left(\mathop{\mathbf{E}}_{y\sim D_{\mathcal{Y}|x}}(y)\right)^2\right]\\
&= \mathop{\mathbf{E}}_{(x,y)\sim D}\left[1 - \left(\mathop{\mathbf{E}}_{y\sim D_{\mathcal{Y}|x}}(y)\right)^2\right]\\
&= \mathop{\mathbf{E}}_{(x,y)\sim D}\left[\mathop{\mathbf{E}}_{y\sim D_{\mathcal{Y}|x}}(y^2) - \left(\mathop{\mathbf{E}}_{y\sim D_{\mathcal{Y}|x}}(y)\right)^2\right]\\
&= \mathop{\mathbf{E}}_{(x,y)\sim D}\left(\mathop{\mathbf{Var}}_{y\sim D_{\mathcal{Y}|x}}(y)\right).
\end{aligned}$$

■

Proposition 4 can be seen as a rigorous demonstration that Proposition 1, namely that the $\mathcal{C}$-bound can be made arbitrarily small for any given problem, is incorrect. By definition, it is clear that the $\mathcal{C}$-bound is lower-bounded by the majority vote risk, which is an irreducible term and greater than zero in the presence of noise. Nonetheless, as the original erroneous proposition contradicted this claim, we wish to emphasise there is no inherent contradiction in the definition of the $\mathcal{C}$-bound. Indeed, we have proven that the $\mathcal{C}$-bound for any given majority vote is lower-bounded by a certain value, which will be greater than 0 if there is any label noise in the distribution D. This result also offers various interesting realisations about the $\mathcal{C}$-bound.

First, there exists a majority vote for any problem such that the $\mathcal{C}$-bound is equal to the mean value of the variance of y for a given input x taken over the input space. This hints towards the fact that the $\mathcal{C}$-bound may degrade quickly if there is significant label noise in the data. For example, consider the classification problem introduced as a counter-example in Section 3. In this case, we easily compute that the $\mathcal{C}$-bound is lower bounded by the value 0.64. This is coherent with the previously found value $\mathcal{C}_Q^D = \frac{400+144n}{400+225n}$ which converges to the optimal value asymptotically. This means that for this problem, no majority vote can produce a $\mathcal{C}$-bound lower than 0.64. This is true even for a majority vote that would always label $y = 1$, which would have true risk of 1/5. This suggests that the $\mathcal{C}$-bound may be too vacuous in settings where the label noise is significant. In fact, one may easily compute that in a context where the label noise is at least $\frac{1}{2} - \frac{1}{2\sqrt{2}} \approx 0.1464$ for any given input x, the $\mathcal{C}$-bound is lower bounded by 1/2 for any given majority vote and is thus uninformative as any voter with a risk greater than 1/2 behaves at least as poorly as a fair coin toss. This differs from the original (erroneous) result of Corollary 2 which suggested that the $\mathcal{C}$-bound could be an optimal tool in any setting.

Second, we observe that if y is assigned deterministically, the value of the lower bound of the $\mathcal{C}$-bound is 0 due to the fact that $\mathbf{Var}(y|x) = 0$. This is a very desirable property as it suggests the $\mathcal{C}$-bound could be arbitrarily small in a deterministic setting and is coherent with Proposition 3.

Finally, we notice that for the $\mathcal{C}$-bound to be minimal, the optimal value of the aggregated voters $\sum_{f \in \mathcal{H}} Q(f) f(x)$ given by the majority vote should be equal to $h(x) = \mathbf{E}_{y \sim D_{\mathcal{Y}|x}}(y)$ for any given input x. This means that according to the $\mathcal{C}$-bound, the confidence of the prediction of a good majority vote for any given input should be representative of the label noise in the distribution for said input. In other words, even though it is known that the optimal classifier always assigns a given x to its most likely label, many majority votes with vastly different margins can capture this behaviour. To minimise the $\mathcal{C}$-bound, a majority vote must also identify the exact label noise of the input x according to D. This suggests that minimising the $\mathcal{C}$-bound may lead to a model that is more informative about the true distribution D when compared to selecting a model that simply minimises the risk. Exploring such behaviour would however require further investigation which exceeds the scope of this erratum and may be explored in future works.

7. Conclusion

Throughout this paper, we have pointed out the fact that Proposition 1 as it is stated in Germain et al. (2015, Proposition 15) is flawed as well as the erroneous nature of the proposition in an applied setting. We were able to correct the statement of the said proposition in a noiseless setting based on the initial intuition given by the authors. This led to the elaboration of a new proposition, Proposition 4, which holds in the general setting. We believe this new property deepens the understanding of the $\mathcal{C}$-bound, notably by demonstrating its shortcomings in problems where a great amount of label noise is present as well as providing new insights regarding the potential properties of models for which the value of the $\mathcal{C}$-bound is optimal.

References

Ke-Lin Du and M. N. S. Swamy. *Combining multiple learners: Data fusion and ensemble learning*, pages 737–767. Springer London, 2019.

Pascal Germain, Alexandre Lacasse, Francois Laviolette, Mario Marchand, and Jean-Francis Roy. Risk bounds for the majority vote: From a PAC-Bayesian analysis to a learning algorithm. *Journal of Machine Learning Research*, 16(26):787–860, 2015. URL http://jmlr.org/papers/v16/germain15a.html.

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

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

David A. McAllester. Simplified PAC-Bayesian margin bounds. In *COLT*, volume 2777 of *Lecture Notes in Computer Science*, pages 203–215. Springer, 2003.

Noam Segev, Maayan Harel, Shie Mannor, Koby Crammer, and Ran El-Yaniv. Learn on source, refine on target: A model transfer learning framework with random forests. *IEEE Trans. Pattern Anal. Mach. Intell.*, 39(9):1811–1824, 2017.

Valentina Zantedeschi, Paul Viallard, Emilie Morvant, Rémi Emonet, Amaury Habrard, Pascal Germain, and Benjamin Guedj. Learning stochastic majority votes by minimizing a PAC-Bayes generalization bound. In *NeurIPS*, pages 455–467, 2021.