PAC-Bayesian Theory for Transductive Learning

Luc Bégin[1] **Pascal Germain**[2] **François Laviolette**[2] **Jean-Francis Roy**[2]
luc.begin@umoncton.ca {pascal.germain, francois.laviolette, jean-francis.roy}@ift.ulaval.ca
[1] Campus d'Edmundston, Université de Moncton, Nouveau-Brunswick, Canada
[2] Département d'informatique et de génie logiciel, Université Laval, Québec, Canada
All authors contributed equally to this work.

Abstract

We propose a PAC-Bayesian analysis of the transductive learning setting, introduced by Vapnik [1998], by proposing a family of new bounds on the generalization error. Some of them are derived from their counterpart in the inductive setting, and others are new. We also compare their behavior.

1 INTRODUCTION

In classification, most learning algorithms are designed for what is defined as the *inductive learning* setting, that corresponds to the following experiment: A learner is given a finite sample of examples (the *training set*), generated independently and identically distributed (*i.i.d.*) from an unknown distribution D, and is asked to produce a classifier having a low probability of misclassifying an example drawn from D (*i.e.*, a low *generalization risk*). It is important to point out that some learning tasks cannot be satisfactorily modelized by this framework. The *i.i.d.* setting implies namely that there is no correlation between the entries, which is a strong assumption. For example, consider the experiment where one collects a finite set of examples Z (possibly non-*i.i.d.*), asks an expert to label a subset S of examples drawn from Z (without replacement), runs a learning algorithm on Z (on labeled S and unlabeled $Z \setminus S$), and finally uses the obtained classifier to label remaining examples $Z \setminus S$. The transductive setting introduced by Vapnik [1998] proposes a paradigm in which one can obtain a generalization guarantee in such non-*i.i.d.* situations.

Since its introduction by McAllester [1999], the PAC-Bayesian theory succeeds to provide tight generalization guarantees for the inductive setting. It has been extended to the transductive setting, namely by Derbeko et al. [2004]. In this paper, we continue their pioneer work by giving tighter bounds. Our proposed bounds do not suffer of the major drawback of Derbeko's bound, which is the fact that its value diverges to infinity as the number of unlabeled examples grows. Also, inspired from Germain et al. [2009], we propose a general transductive PAC-Bayesian theorem that is a tool to derive various bounds by choosing a convex function $\mathcal{D}$. Interestingly, unlike in the inductive setting, our general transductive theorem has no real limitation in the choice of $\mathcal{D}$. Thus, we derive a family of transductive PAC-Bayesian bounds, and compare their behavior on empirical data. We also propose a bound that takes into account the unlabeled part of the data, while usual transductive bounds only consider the number of unlabeled examples.

1.1 Inductive versus Transductive Learning

We consider binary classification problems with an arbitrary input space $\mathcal{X}$ and output space $\mathcal{Y} = \{-1, 1\}$. An example $(x, y) \in \mathcal{X} \times \mathcal{Y}$ is an input-output pair, where x is a description, and y is a label.

In inductive learning, we consider that each example (x, y) is drawn *i.i.d.* from a fixed, but unknown, probability distribution D on $\mathcal{X} \times \mathcal{Y}$. The *training set* of m examples is denoted by $S = \{(x_1, y_1), (x_2, y_2), \ldots, (x_m, y_m)\} \sim D^m$. The task of an inductive learner is to, given the training set S, learn a *classifier* $h : \mathcal{X} \mapsto \mathcal{Y}$ that will be capable of classifying new examples drawn according to distribution D. We can now define the *risk* of a classifier h on a distribution D' by being the probability that h misclassifies an example generated by D',

$$R_{D'}(h) \stackrel{\text{def}}{=} \mathop{\mathbf{E}}_{(x,y)\sim D'} I\Big[h(x) \neq y\Big] ,$$

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

Appearing in Proceedings of the 17^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2014, Reykjavik, Iceland. JMLR: W&CP volume 33.

If we calculate a risk on a discrete set S' instead of a probability distribution, we consider the uniform distribution on S', thus calculating the mean :

$$R_{S'}(h) = \frac{1}{|S'|} \sum_{(x,y)\in S'} I\Big[h(x) \neq y\Big] .$$

In transductive learning, we consider a set Z of N examples, $Z = \{(x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)\}$, often named the *full sample*, which contains all examples of interest.[1] We then obtain a training set S by *drawing m examples from Z without replacement*. The remaining examples from Z form a set U of $N-m$ examples. In the transductive setting, a learning algorithm is given the training set S, and $U_{\mathcal{X}} = \{x_{m+1}, x_{m+2}, \ldots, x_N\}$, the set of unlabeled examples of U. The task of a transductive learner is to learn a classifier $h : Z_{\mathcal{X}} \mapsto \mathcal{Y}$ that correctly classifies the unlabeled examples of the set U.[2] Thus, the goal is to minimize the risk of the classifier on U.

We also define $R_Z(h)$ and $R_S(h)$ similarly. Note that one can recover $R_Z(h)$ from $R_S(h)$ and $R_U(h)$, as

$$R_Z(h) \;=\; \frac{1}{N}\Big(mR_S(h) + (N-m)R_U(h)\Big) . \qquad (1)$$

In both inductive and transductive learning, the goal is to find the classifier with the lowest possible risk on a distribution or a set that is not completely known to the learner : the data-generating distribution D in inductive learning, and the set U in transductive learning, from which the learner does not know the labels. Fortunately, PAC-Bayesian theorems will allow us to upper bound these risks by using their *empirical counterpart* plus some *complexity term*.

1.2 Inductive PAC-Bayesian Theory

Consider a training set S of m examples drawn *i.i.d.* from a data-generating distribution D, a hypothesis space $\mathcal{H}$ of classifiers, a *prior* distribution P on $\mathcal{H}$, and a *posterior* distribution Q on $\mathcal{H}$. The *prior* encodes some knowledge about the problem (before exploiting the information contained in S), while the *posterior* is obtained by running a learning algorithm on S. The PAC-Bayesian theory studies the *Gibbs classifier* G_Q which, given a distribution Q on $\mathcal{H}$, classifies an example x by drawing at random a classifier h according to Q, and returns $h(x)$. Thus, G_Q is a stochastic classifier whose risk on a set S' is given by

$$R_{S'}(G_Q) \;=\; \frac{1}{|S'|} \sum_{(x,y)\in S'} \mathop{\mathbf{E}}_{h\sim Q} I\Big[h(x) \neq y\Big] . \qquad (2)$$

As discussed in Section 2.5, the Gibbs classifier's risk is closely related to the risk of the Q-weighted deterministic majority vote classifier.

Inductive PAC-Bayesian bounds give guarantees on the generalization risk $R_D(G_Q)$, that corresponds to the probability that G_Q makes an error on an example generated by D. Typically, these bounds rely on the empirical risk $R_S(G_Q)$ and the Kullback-Leibler divergence between the prior and posterior distributions. The following PAC-Bayesian theorem originally comes from Germain et al. [2009], but is presented in a slightly different form to ease the comparison with the transductive bounds of Section 2. Given any convex function $\mathcal{D} : [0,1]\times[0,1] \to \mathbb{R}$, often named *$\mathcal{D}$-function* in this paper, Theorem 1 allows one to obtain an interval in which lies the risk $R_D(G_Q)$ with high probability. Thus, the extremities of this interval give both a lower bound and an upper bound of $R_D(G_Q)$.

Theorem 1. *For any distribution D, for any set $\mathcal{H}$ of classifiers, for any prior distribution P on $\mathcal{H}$, for any $\delta \in (0,1]$, and for any convex function $\mathcal{D} : [0,1] \times [0,1] \to \mathbb{R}$, with probability at least $1-\delta$ over the choice of $S \sim D^m$, we have*

$$\forall Q \text{ on } \mathcal{H}:$$
$$\mathcal{D}\big(R_S(G_Q), R_D(G_Q)\big) \;\leq\; \frac{1}{m}\left[\mathrm{KL}(Q\|P) + \ln\frac{\mathcal{I}_{\mathcal{D}}(m)}{\delta}\right],$$
$$\text{where } \mathcal{I}_{\mathcal{D}}(m) \stackrel{\text{def}}{=} \sup_{r\in[0,1]}\left[\sum_{k=0}^{m}\binom{m}{k} r^k(1-r)^{m-k} e^{m\mathcal{D}(\frac{k}{m}, r)}\right],$$
$$\text{and where } \mathrm{KL}(Q\|P) \stackrel{\text{def}}{=} \mathbf{E}_{h\sim Q} \ln\tfrac{Q(h)}{P(h)}.$$

Proof. We start from Theorem 2.1 of Germain et al. [2009], that is, with probability at least $1-\delta$ over the choice of $S \sim D^m$, for all distributions Q on $\mathcal{H}$,

$$\mathcal{D}(R_S(G_Q), R_D(G_Q)) \leq \frac{1}{m}\left[\mathrm{KL}(Q\|P) + \ln\frac{1}{\delta} \mathop{\mathbf{E}}_{S\sim D^m} \mathop{\mathbf{E}}_{h\sim P} e^{m\mathcal{D}(R_S(h), R_D(h))}\right].$$

Because the choice of P is independent of S, and given that the number of errors $mR_S(h)$ follows a binomial distribution with parameters m and $R_D(h)$, we have

$$\begin{aligned}
&\mathop{\mathbf{E}}_{S\sim D^m} \mathop{\mathbf{E}}_{h\sim P} e^{m\mathcal{D}(R_S(h), R_D(h))} \\
&= \mathop{\mathbf{E}}_{h\sim P} \sum_{k=0}^{m} \Pr_{S\sim D^m}\Big(R_S(h) = \tfrac{k}{m}\Big) e^{m\mathcal{D}(\frac{k}{m}, R_D(h))} \\
&= \mathop{\mathbf{E}}_{h\sim P} \sum_{k=0}^{m} \binom{m}{k} \big(R_D(h)\big)^k \big(1-R_D(h)\big)^{m-k} e^{m\mathcal{D}(\frac{k}{m}, R_D(h))} \\
&\leq \mathcal{I}_{\mathcal{D}}(m) .
\end{aligned}$$

□

[1] In the transductive "Setting 1" of Vapnik [1998], the full sample is drawn *i.i.d.* from an unknown distribution on $\mathcal{X}\times\mathcal{Y}$. This assumption is not needed here.

[2] In Vapnik [1998], the transductive classifier is defined using $U_{\mathcal{X}}$ as an input space, i.e., $h : U_{\mathcal{X}} \mapsto \mathcal{Y}$. However, PAC-Bayesian bounds need the classifier to be defined on S as well, as they require the computation of $R_S(h)$.

As discussed in Germain et al. [2009], Theorem 1 is a generic tool to derive various inductive PAC-Bayesian bounds, as $\mathcal{D}$ can be any convex function. However, one needs to calculate (or upper bound) the value of $\mathcal{I}_{\mathcal{D}}(m)$ to express a computable bound. A common choice is $\mathcal{D} = \mathcal{D}_{\mathrm{KL}}$, the Kullback-Leibler divergence between two Bernoulli distributions of probability of success p and q, defined by

$$\mathcal{D}_{\mathrm{KL}}(q,p) \stackrel{\text{def}}{=} q \ln \tfrac{q}{p} + (1-q) \ln \tfrac{1-q}{1-p} \,. \tag{3}$$

We can rewrite $\mathcal{D}_{\mathrm{KL}}(q,p) = H(q,p) - H(q)$, using usual definitions of entropy and cross-entropy,

$$\begin{aligned} H(q) &\stackrel{\text{def}}{=} -q \ln q - (1-q)\ln(1-q) \,, \\ H(q,p) &\stackrel{\text{def}}{=} -q \ln p - (1-q)\ln(1-p) \,. \end{aligned} \tag{4}$$

With these definitions, it is easy to see that the r's cancel out in each term of the inner sum of $\mathcal{I}_{\mathcal{D}_{\mathrm{KL}}}(m)$, giving the following simplification:

$$\mathcal{I}_{\mathcal{D}_{\mathrm{KL}}}(m) = \sup_{r\in[0,1]} \left[\sum_{k=0}^{m} \tbinom{m}{k} e^{-m H(\frac{k}{m})} \right] = \sum_{k=0}^{m} \alpha(k,m) \,, \tag{5}$$

where $\alpha(a,b) \stackrel{\text{def}}{=} \binom{b}{a} \left(\frac{a}{b}\right)^a \left(1-\frac{a}{b}\right)^{b-a}$.

Furthermore, when one wants to avoid the computational burden needed to compute the sum of Equation (5), it is also possible to upper bound the value of $\mathcal{I}_{\mathcal{D}_{\mathrm{KL}}}(m)$ by simpler expressions, using bounds on the function $\alpha(\cdot,\cdot)$, expressed by Lemmas 2 and 3 below.

Lemma 2. *Given any integers a, b such that $0 \le a \le b$,*

$$\frac{1}{b+1} \;\le\; \alpha(a,b) \;\le\; 1 \,.$$

Proof. $\alpha(a,b)$ corresponds to the probability mass function of a Bernoulli trial of b experiments with probability of success $\frac{a}{b}$, evaluated at point a (the most probable event among $b+1$ possible outcomes). □

From Lemma 2, we trivially obtain that $\mathcal{I}_{\mathcal{D}_{\mathrm{KL}}}(m) \le m+1$. However, Maurer [2004] shows that $\mathcal{I}_{\mathcal{D}_{\mathrm{KL}}}(m) \le 2\sqrt{m}$ is a tight upper-bound. Lemma 3 presents one key step of the proof leading to this result. We reuse this lemma to obtain new transductive guarantees in the next section.

Lemma 3. *Given any integers a, b such that $0 < a < b$,*

$$\sqrt{\tfrac{b}{2\pi a(b-a)}} e^{-\frac{1}{12a} - \frac{1}{12(b-a)}} \;<\; \alpha(a,b) \;<\; \sqrt{\tfrac{b}{2\pi a(b-a)}} e^{\frac{1}{12b}} .$$

Proof. The result follows from straightforward calculations using Stirling bounds of the factorial, *i.e.*, $\sqrt{2\pi n}\left(\frac{n}{e}\right)^n < n! < \sqrt{2\pi n}\left(\frac{n}{e}\right)^n e^{\frac{1}{12n}}$. □

We conclude this section by stating two inductive PAC-Bayesian bounds in Corollary 4, below. Bound (a) is similar to the bound of Seeger [2002], and Bound (b) is similar to the one of McAllester [2003a]. Note that Bound (a) is tighter than Bound (b), but the latter is easier to compute since it has an explicit form.

Corollary 4. *For any distribution D, for any set $\mathcal{H}$ of classifiers, for any prior distribution P on $\mathcal{H}$, for any $\delta \in (0,1]$, with probability at least $1-\delta$ over the choice of $S \sim D^m$, we have*

$\forall Q$ *on* $\mathcal{H}$:

a) $\mathcal{D}_{\mathrm{KL}}\big(R_S(G_Q), R_D(G_Q)\big) \le \frac{1}{m}\left[\mathrm{KL}(Q\|P) + \ln \frac{2\sqrt{m}}{\delta}\right]$,

b) $R_D(G_Q) \;\le\; R_S(G_Q) + \sqrt{\frac{1}{2m}\left[\mathrm{KL}(Q\|P) + \ln \frac{2\sqrt{m}}{\delta}\right]}$.

Proof. Bound (a) is obtained from Theorem 1, with $\mathcal{D}(q,p) = \mathcal{D}_{\mathrm{KL}}(q,p)$, and $\mathcal{I}_{\mathcal{D}_{\mathrm{KL}}}(m) \le 2\sqrt{m}$ [Maurer, 2004]. Bound (b) is obtained from Bound (a), using Pinsker's inequality: $\mathcal{D}_{\mathrm{KL}}(q,p) \ge 2(q-p)^2$. □

2 TRANSDUCTIVE PAC-BAYESIAN THEORY

2.1 Transductive General Theorem

In the inductive setting, an important assumption used to derive PAC-Bayesian guarantees is that the m examples of the training set S are drawn *i.i.d.* from the data-generating distribution D. In the proof of Theorem 1, we use this fact to express the probability of misclassifying k among m examples as a binomial distribution. This assumption does not hold in the transductive setting. Indeed, in the transductive setting, the set of labeled examples S is a subset of a finite set Z. Therefore, the number of errors observed in S follows a hypergeometric distribution, as S contains m draws *without replacement* from Z. This idea is exploited in the proof of Theorem 5, below.

Theorem 5. *For any set Z of N examples, for any set $\mathcal{H}$ of classifiers, for any prior distribution P on $\mathcal{H}$, for any $\delta \in (0,1]$, and for any convex function $\mathcal{D} : [0,1] \times [0,1] \to \mathbb{R}$, with probability at least $1-\delta$ over the choice S of m examples among Z, we have*

$\forall Q$ *on* $\mathcal{H}$:

$$\mathcal{D}\big(R_S(G_Q), R_Z(G_Q)\big) \le \frac{1}{m}\left[\mathrm{KL}(Q\|P) + \ln \frac{\mathcal{T}_{\mathcal{D}}(m,N)}{\delta}\right],$$

where

$$\mathcal{T}_{\mathcal{D}}(m,N) \stackrel{\text{def}}{=} \max_{K=0\ldots N} \left[\sum_{k \in \mathcal{K}_{mNK}} \frac{\binom{K}{k}\binom{N-K}{m-k}}{\binom{N}{m}} e^{m\mathcal{D}(\frac{k}{m}, \frac{K}{N})} \right], \tag{6}$$

and $\mathcal{K}_{mNK} \stackrel{\text{def}}{=} \{\max[0, K+m-N], \ldots, \min[m, K]\}$.

Proof. Let denote $[Z]^m$ the uniform distribution over all subsets of Z of size m. Consider the non-negative random variable $\mathbf{E}_{h\sim P}\, e^{m\mathcal{D}(R_S(h),R_Z(h))}$. Markov's inequality gives that, with probability at least $1-\delta$ over the choice of m examples among Z,

$$\mathop{\mathbf{E}}_{h\sim P} e^{m\mathcal{D}(R_S(h),R_Z(h))} \leq \frac{1}{\delta} \mathop{\mathbf{E}}_{S\sim[Z]^m} \mathop{\mathbf{E}}_{h\sim P} e^{m\mathcal{D}(R_S(h),R_Z(h))}.$$

By taking the logarithm on each side of the above equation, and applying the change of measure inequality $\mathbf{E}_{f\sim Q}[f] \leq \mathrm{KL}(Q\|P) + \ln(\mathbf{E}_{f\sim P}[e^f])$ [Donsker and Varadhan, 1975] we obtain that, for all choices of Q,

$$\mathop{\mathbf{E}}_{h\sim Q} \mathcal{D}\big(R_S(h), R_Z(h)\big) \leq \frac{1}{m}\left[\mathrm{KL}(Q\|P) + \ln\frac{1}{\delta} \mathop{\mathbf{E}}_{S\sim[Z]^m} \mathop{\mathbf{E}}_{h\sim P} e^{m\mathcal{D}(R_S(h),R_Z(h))}\right].$$

By Jensen's inequality on convex function $\mathcal{D}$, we have $\mathbf{E}_{h\sim Q}\, \mathcal{D}\big(R_S(h), R_Z(h)\big) \geq \mathcal{D}\big(R_S(G_Q), R_Z(G_Q)\big)$.

Because the choice of P is independent of S, and given that the number of errors $mR_S(h)$ follows a hypergeometric distribution of m draws among a population of size N containing $NR_Z(h)$ successes, we have

$$\begin{aligned}
&\mathop{\mathbf{E}}_{S\sim[Z]^m} \mathop{\mathbf{E}}_{h\sim P} e^{m\mathcal{D}(R_S(h),R_Z(h))} \\
&= \mathop{\mathbf{E}}_{h\sim P} \sum_{k=0}^{m} \Pr_{S\sim[Z]^m}\left(R_S(h) = \tfrac{k}{m}\right) e^{m\mathcal{D}(\frac{k}{m}, R_Z(h))} \\
&= \mathop{\mathbf{E}}_{h\sim P} \sum_{k=\max[0,NR_Z(h)-N+m]}^{\min[m,NR_Z(h)]} \frac{\binom{NR_Z(h)}{k}\binom{N-NR_Z(h)}{m-k}}{\binom{N}{m}} e^{m\mathcal{D}(\frac{k}{m}, R_Z(h))} \\
&\leq \mathcal{T}_{\mathcal{D}}(m,N)\,. \qquad \square
\end{aligned}$$

There is a correspondence between the expression $\mathcal{I}_{\mathcal{D}}(m)$ in Theorem 1 (inductive case) and the expression $\mathcal{T}_{\mathcal{D}}(m,N)$ in Theorem 5 (transductive case). However, to compute the value $\mathcal{I}_{\mathcal{D}}(m)$, one needs to find the supremum value of a possibly neither convex nor concave expression over a continuous variable $r \in [0,1]$. Instead, the value of $\mathcal{T}_{\mathcal{D}}(m,N)$ is given by a maximum value of the inner sum over a discrete variable $K \in \{0,1,2,\ldots,N\}$. This value can be computed directly for any $\mathcal{D}$-function (*i.e.*, any convex function defined on $[0,1]\times[0,1]$), provided that m and N are not unreasonably large. It opens the way to the use of many $\mathcal{D}$-functions.

For instance, a natural choice of $\mathcal{D}$-function is the divergence $\mathcal{D}_{\mathrm{KL}}(q,p)$, as it leads to one of the best known bound in the inductive setting. By Equations (3) and (6), we obtain

$$\mathcal{T}_{\mathcal{D}_{\mathrm{KL}}}(m,N) = \max_{K=0\ldots N}\left[\sum_{k\in\mathcal{K}_{mNK}} \frac{\binom{K}{k}\binom{N-K}{m-k}}{\binom{N}{m}} \left(\frac{k/m}{K/N}\right)^k \left(\frac{1-k/m}{1-K/N}\right)^{m-k}\right]. \tag{7}$$

Unfortunately, the obtained expression does not simplify itself as in the inductive case (see Equation (5)). Hence, the time needed to compute the value of $\mathcal{T}_{\mathcal{D}_{\mathrm{KL}}}(m,N)$ depends on the magnitude of N. To overcome this issue, we design a $\mathcal{D}$-function tailored for the transductive setting in the following section.

2.2 A $\mathcal{D}$-function for the Transductive Case

In the inductive setting, we express $\mathcal{I}_{\mathcal{D}_{\mathrm{KL}}}(m)$ by a sum of terms $\alpha(k,m)$ that can be bounded using either Lemma 2 or 3. To recover the same phenomenon in the transductive setting, we suggest to use the following $\mathcal{D}$-function that pairs each of the three binomial coefficients of $\mathcal{T}_{\mathcal{D}}(m,N)$ (defined by Equation (6)) with an appropriate entropy term (defined by Equation (4)).

$$\mathcal{D}^{\star}_{\beta}(q,p) \stackrel{\mathrm{def}}{=} \frac{H(\beta) - pH(\beta\frac{q}{p}) - (1-p)H(\beta\frac{1-q}{1-p})}{\beta}. \tag{8}$$

The β parameter will typically be set to $\frac{m}{N}$. As shown by Lemma S9[3], Equation (8) can be rewritten as

$$\mathcal{D}^{\star}_{\beta}(q,p) \;=\; \mathcal{D}_{\mathrm{KL}}(q,p) + \tfrac{1-\beta}{\beta}\,\mathcal{D}_{\mathrm{KL}}\Big(\tfrac{p-\beta q}{1-\beta}, p\Big)\,. \tag{9}$$

The latter equation highlights that, when $N\to\infty$ and m is finite, $\mathcal{D}^{\star}_{m/N}(q,p)$ converges to $\mathcal{D}_{\mathrm{KL}}(q,p)$. That is, we recover the KL-divergence used in inductive learning whereas the full sample cardinality is infinite.

Interestingly, the formulation of $\mathcal{D}^{\star}_{\beta}(q,p)$ of Equation (9) appears in the proof of the transductive theorem of Derbeko et al. [2004]. However, as we discuss in Section 2.4, tighter bounds can be derived using this same $\mathcal{D}$-function. Indeed, when we introduce Equation (8) in Equation (6), with $\beta = \frac{m}{N}$, we have

$$\mathcal{T}_{\mathcal{D}^{\star}_{m/N}}(m,N) = \max_{K=0\ldots N}\left[\sum_{k\in\mathcal{K}_{mNK}} \frac{\alpha(k,K)\,\alpha(m-k,N-K)}{\alpha(m,N)}\right]. \tag{10}$$

By Lemma 2, we trivially obtain that

$$\mathcal{T}_{\mathcal{D}^{\star}_{m/N}}(m,N) \leq \max_{K=0\ldots N} \sum_{k=0}^{m} N{+}1 = (m{+}1)(N{+}1)\,. \tag{11}$$

However, this upper-bound on $\mathcal{T}_{\mathcal{D}^{\star}_{m/N}}(m,N)$ is far from being tight, as shown by the next theorem.

Theorem 6. *Let m and N be any integers such that $20 \leq m \leq N-20$, we have*

$$\mathcal{T}_{\mathcal{D}^{\star}_{m/N}}(m,N) \leq t(m,N) \stackrel{\mathrm{def}}{=} 3\ln(m)\sqrt{m(1-\tfrac{m}{N})}\,. \tag{12}$$

[3] Throughout the paper, a lemma prefixed by 'S' refers to a result whose proof is in the supplementary material.

Proof. Given fixed m, N, K, let $k^- = \max[0, K+m-N]$, $k^+ = \min[m, K]$, $\mathcal{K}^*_{mNK} = \mathcal{K}_{mNK} \setminus \{k^-, k^+\}$, and

$$F(k) = \frac{\alpha(k,K)\,\alpha(m-k,N-K)}{\alpha(m,N)}\,.$$

Lemma S10 shows that

$$F(k^-) + F(k^+) \;\leq\; 2\,e^{\frac{1}{6\times 20}}\sqrt{2\pi\, m(1-\tfrac{m}{N})}\,.$$

Moreover, using Lemma 3 and algebraic manipulations, we obtain for any $k \in \mathcal{K}^*_{mNK}$,

$$F(k) < \frac{\gamma}{\sqrt{2\pi}}\sqrt{m(1-\tfrac{m}{N})\Big(\tfrac{1}{k}+\tfrac{1}{K-k}\Big)\Big(\tfrac{1}{m-k}+\tfrac{1}{N-K-m+k}\Big)}\,,$$

where $\gamma = e^{\frac{1}{12}[\frac{1}{K}+\frac{1}{N-K}+\frac{1}{m}+\frac{1}{N-m}]} \leq e^{\frac{1}{12}[2+\frac{2}{20}]}$ as $m \geq 20$ and $N-m \geq 20$. Furthermore, Lemma S11 shows that

$$\sum_{k\in\mathcal{K}^*_{mNK}} \sqrt{\Big(\tfrac{1}{k}+\tfrac{1}{K-k}\Big)\Big(\tfrac{1}{m-k}+\tfrac{1}{N-K-m+k}\Big)} \;\leq\; 2[1+\ln(m)].$$

Then,
$$\sum_{k\in\mathcal{K}_{mNK}} F(k) \;\leq\; \sqrt{m(1-\tfrac{m}{N})}\times C(m)\,, \qquad (13)$$

where $C(m) = 2\,e^{\frac{1}{6\times 20}}\sqrt{2\pi} + \frac{\gamma}{\sqrt{2\pi}}2[1+\ln(m)]$. For $m \geq 20$, we have $C(m) \leq 3\ln(m)$. Since Equation (13) is independent of K, we are done. □

From Theorem 6, we conclude $\mathcal{T}_{\mathcal{D}_{\mathrm{KL}}}(m,N) \leq t(m,N)$. Indeed, Eq. (6) and (9) give $\mathcal{D}_{\mathrm{KL}}(q,p) \leq \mathcal{D}^\star_\beta(q,p)$, and then $\mathcal{T}_{\mathcal{D}_{\mathrm{KL}}}(m,N) \leq \mathcal{T}_{\mathcal{D}^\star_\beta}(m,N)$. However, one should not conclude from latter inequality that the bounds obtained from Theorem 5 are tighter using $\mathcal{D}_{\mathrm{KL}}$ instead of $\mathcal{D}^\star_\beta$ as $\mathcal{D}$-function, as the choice of $\mathcal{D}$ impacts both sides of the inequality of Theorem 5.

2.3 New Explicit PAC-Bayesian Bounds

The next result presents two transductive bounds derived from Theorems 5 and 6. Bound (a) is the tightest, while Bound (b) has an explicit form.

Corollary 7. *For any set Z of $N \geq 40$ examples, for any set $\mathcal{H}$ of classifiers, for any prior distribution P on $\mathcal{H}$, for any $\delta \in (0,1]$, with probability at least $1-\delta$ over the choice S of m examples among Z (such that $20 \leq m \leq N-20$), we have*

$\forall Q$ *on* $\mathcal{H}$:

a) $\mathcal{D}^\star_{m/N}\big(R_S(G_Q), R_Z(G_Q)\big) \leq \frac{1}{m}\Big[\mathrm{KL}(Q\|P) + \ln\frac{t(m,N)}{\delta}\Big]$,

b) $R_Z(G_Q) \leq R_S(G_Q) + \sqrt{\frac{1-\frac{m}{N}}{2m}\Big[\mathrm{KL}(Q\|P) + \ln\frac{t(m,N)}{\delta}\Big]}$,

where $t(m,N)$ is defined by Equation (12).

Proof. Bound (a) is obtained from Theorem 5, with $\mathcal{D}(q,p) = \mathcal{D}^\star_{m/N}(q,p)$, and from Theorem 6. From Bound (a), using Equation (9) and Pinsker's inequality ($\mathcal{D}_{\mathrm{KL}}(q,p) \geq 2(q-p)^2$) twice, we get

$$\begin{aligned}\mathcal{D}^\star_{m/N}(R_S,R_Z) \qquad & (14)\\ \geq\;& 2(R_S-R_Z)^2 + 2(\tfrac{N}{m}-1)\Big(\tfrac{R_Z-\frac{m}{N}R_S}{1-\frac{m}{N}} - R_Z\Big)^2\\ =\;& \tfrac{2(R_S-R_Z)^2}{1-\frac{m}{N}}\,,\end{aligned}$$

and Bound (b) is then obtained by isolating R_Z in $2(R_S-R_Z)^2 \leq (1-\frac{m}{N})\frac{1}{m}\Big[\mathrm{KL}(Q\|P) + \ln\frac{t(m,N)}{\delta}\Big]$. □

The two bounds presented by Corollary 7 are analogous to the inductive bounds of Corollary 4. Moreover, these transductive bounds converge toward their inductive counterpart as $N \to \infty$ (up to a small factor due to the term $\ln(t(m,N))$ which tends to $\ln(3\ln(m)\sqrt{m})$ instead of $\ln(2\sqrt{m})$, but this has a little effect since divided by m). More precisely, Bound (a) can be regarded as a generalization of the bound of Seeger [2002]. Indeed, as shown by Equation (9), $\mathcal{D}^\star_{m/N}(R_S(G_Q), R_Z(G_Q)) \simeq \mathcal{D}_{\mathrm{KL}}(R_S(G_Q), R_Z(G_Q))$ whereas $N \gg m$ (that is, the ratio $\frac{m}{N}$ tends to 0). Also, Bound (b) generalizes the PAC-Bayesian bound of McAllester [2003a], as the multiplicative factor $\frac{1-\frac{m}{N}}{2m}$ reduces to $\frac{1}{2m}$ whereas $N \gg m$.

2.4 Relation with Previous Works

The transductive bounds presented in this paper can be seen as an improvement of the prior work of Derbeko et al. [2004]. First, let us reveal a small error in the proof of Theorem 18 of Derbeko et al. At the end of their proof, *to obtain an explicit bound*, they lower bound the divergence $\mathcal{D}^\star_{m/N}(R_S,R_Z)$, applying the inequality $\mathcal{D}_{\mathrm{KL}}(q,p) \geq \frac{(q-p)^2}{2p}$ twice (see Equation (17) of Derbeko et al.). However, as stated in McAllester [2003b], this inequality only holds when $q < p$, and therefore cannot be applied on the term $\mathcal{D}_{\mathrm{KL}}(\frac{R_Z-\frac{m}{N}R_S}{1-\frac{m}{N}}, R_Z)$ when $R_S < R_Z$, because we necessarily have $\frac{R_Z-\frac{m}{N}R_S}{1-\frac{m}{N}} \geq R_Z$. The error can be fixed by using Pinsker's inequality instead, like in Equation (14) of the current paper. We present the fixed result, and a detailed proof, in the supplementary material (see Theorem S12). The obtained (fixed) bound states that an upper bound on $R_Z(G_Q)$ is given by

$$R_S(G_Q) + \sqrt{\frac{1-\frac{m}{N}}{2(m-1)}\Big[\mathrm{KL}(Q\|P) + \ln\tfrac{m}{\delta} + 7\ln(N+1)\Big]}. \qquad (15)$$

Hence, the major difference between Bound (b) of Corollary 7 and the expression of Equation (15) is that the complexity term $\ln(m) + 7\ln(N+1)$ of the latter replaces the term $\ln(t(m,N))$ of the former.

Our result therefore leads to much tighter bounds. Indeed, we already obtain a tighter bound if we loosely upper bound $\ln(t(m,N))$ by $\ln((m+1)(N+1))$, using Equation (11) instead of using the much tighter Equation (12) of Theorem 6. The major drawback the expression of Derbeko et al. [2004] is that the bound's value (both in its original and fixed versions) diverge to infinity as N grows, unless m, the number of labeled examples also goes to infinity. This is clearly an unwanted behavior. Conversely, as discussed in section 2.3, both bounds of Corollary 7 converge to their inductive counterpart as the ratio $\frac{m}{N}$ tends to 0.

2.5 Bounds on the Risk of Majority Votes

Many machine learning algorithms, like Ensemble Methods, construct a posterior distribution Q over a hypothesis class $\mathcal{H}$. However, the Gibbs classifier is not commonly used, as we generally prefer the deterministic behavior of the Q-weighted majority vote classifier (also called the *Bayes classifier*) defined as

$$B_Q(x) \stackrel{\text{def}}{=} \underset{c\in\mathcal{Y}}{\operatorname{argmax}} \left[\mathop{\mathbf{E}}_{h\sim Q} I(h(x)=c) \right].$$

Nevertheless, the output of the majority vote classifier is closely related to the output of the stochastic *Gibbs classifier* G_Q. Any upper-bound for the Gibbs classifier's risk $R_{S'}(G_Q)$ can straightforwardly be turned into a bound of the majority vote classifier's risk $R_{S'}(B_Q)$, as $R_{S'}(B_Q) \leq 2R_{S'}(G_Q)$ [Langford and Shawe-Taylor, 2002]. This *factor-of-two law* can be misleading, since in many situations, for a same posterior distribution Q, $R_{S'}(G_Q)$ is greater $R_{S'}(B_Q)$. To capture more effectively the *community effect* of the majority vote, Lacasse et al. [2006] suggested to exploit the relation exposed by Theorem 8, below.

Theorem 8 (Lacasse et al. [2006])**.** *For any distribution Q on $\mathcal{H}$ and any dataset S', if $R_{S'}(G_Q) \leq \frac{1}{2}$, then*

$$R_{S'}(B_Q) \;\leq\; \mathcal{C}_Q^{S'} \;\stackrel{\text{def}}{=}\; 1 - \frac{\left(1-2\,R_{S'}(G_Q)\right)^2}{1-2\,d_Q^{S'}}\,, \quad (16)$$

where $d_Q^{S'}$ is the expected disagreement*:*

$$d_Q^{S'} \stackrel{\text{def}}{=} \frac{1}{|S'|} \sum_{x\in S'_{\mathcal{X}}} \mathop{\mathbf{E}}_{h_1\sim Q} \mathop{\mathbf{E}}_{h_2\sim Q} I\Big[h_1(x) \neq h_2(x)\Big]. \quad (17)$$

Note that the value $\mathcal{C}_Q^{S'}$ is a better bound on $R_{S'}(B_Q)$ than the factor-of-two law if and only if the expected disagreement is greater than the Gibbs classifier's risk. (*i.e.*, $R_{S'}(G_Q) < d_Q^{S'} \Leftrightarrow \mathcal{C}_Q^{S'} < 2\,R_{S'}(G_Q)$). As this situation often occurs when learning from empirical data, Lacasse et al. [2006] suggested PAC-Bayesian bounds giving a generalization guarantee on $\mathcal{C}_Q^D$ (that is, an upper bound on the risk of the Q-weighted majority vote on the data-generating distribution D) from an empirical estimate $\mathcal{C}_Q^S$ (computed on a training set S). Unfortunately, the quotient in Equation (16) causes a rapid degradation of the bound of $\mathcal{C}_Q^D$ when one bounds simultaneously the values of $R_D(G_Q)$ and d_Q^D. This degradation problem is avoided in the transductive setting, since we have access to the exact value of the expected disagreement on the full sample d_Q^Z. Indeed, as we see in Equation (17), the labels are not required to compute the value of the expected disagreement. Hence, an upper bound on $R_Z(G_Q)$ can be directly converted to an upper bound on $\mathcal{C}_Q^Z$, and therefore on $R_Z(B_Q)$. By using the expected disagreement on the full sample, these bounds extract more information about the learning problem. Theorem 8 then becomes a powerful tool to derive transductive risk bounds on majority vote classifiers.[4]

3 EMPIRICAL STUDY

In this section, we conduct two different experiments.[5] Section 3.1 explores how the choice of a $\mathcal{D}$-function in Theorem 5 impacts the value of the bound. Section 3.2 shows bound values on multiple real-world datasets.

3.1 Exploiting different $\mathcal{D}$-functions

An interesting aspect of Theorem 5 is the possibility to use any $\mathcal{D}$-function to compute a transductive PAC-Bayesian bound. The choice of a $\mathcal{D}$-function leads to more or less accurate bounds according to the other bounds parameters: the training risk $R_S(G_Q)$, the divergence between the posterior and the prior $\mathrm{KL}(Q\|P)$, the training set size m, the full sample size N, and (less importantly) the confidence value δ.

In this section, we conduct experiments with five $\mathcal{D}$-functions. The first two have been presented earlier. These are the KL-divergence between two Bernoulli distributions of Equation (3), and the $\mathcal{D}^\star_\beta$-function of Equation (8). We also experimented many other candidates of $\mathcal{D}$-functions. We choose to present three of those, that are simple and have an interesting behavior. These are the *variation distance* $\mathcal{D}_V$, the *quadratic distance* $\mathcal{D}_{V^2}$, and the *triangular discrimination* $\mathcal{D}_\triangle$, defined as

$$\mathcal{D}_V(q,p) \stackrel{\text{def}}{=} 2\,|q-p|\,, \quad \mathcal{D}_{V^2}(q,p) \stackrel{\text{def}}{=} 2\,(q-p)^2\,,$$
$$\mathcal{D}_\triangle(q,p) \stackrel{\text{def}}{=} \tfrac{(q-p)^2}{q+p} + \tfrac{(q-p)^2}{2-q-p}\,.$$

Note that the above are three well-known divergences

[4] Laviolette et al. [2011] presented a similar bound, but in an asymptotic form, making it unusable in practice.

[5] The code to reproduce these experiments is available at: http://graal.ift.ulaval.ca/aistats2014

in the literature (*e.g.*, Topsøe [2000]). Note also that the quadratic distance leads to the PAC-Bayesian bound of McAllester [2003a] by the Pinsker's inequality, *i.e.*, $\mathcal{D}_{\mathrm{KL}}(q,p) \geq \mathcal{D}_{V^2}(q,p)$.

Figure 1 compares the above-mentioned $\mathcal{D}$-functions ($\mathcal{D}_{\mathrm{KL}}$, $\mathcal{D}^{\star}_{m/N}$, $\mathcal{D}_{V^2}$, $\mathcal{D}_{V}$, and $\mathcal{D}_{\triangle}$), with a fixed training set risk of $R_S(G_Q) = 0.2$, a fixed divergence of $\mathrm{KL}(Q\|P)=5$ and a fixed confidence value of $\delta = 0.05$, when varying the full sample size N and the ratio $\frac{m}{N}$ of training set size over full sample size. More precisely, we consider nine possible pairs (N,m) where $N \in \{200, 500, 500\}$ and $m \in \{\frac{1}{10}N, \frac{1}{2}N, \frac{9}{10}N\}$. For each set of parameters, an upper bound of the full sample risk $R_Z(G_Q)$ is computed according to Theorem 5, by finding the value of $r \geq R_S(G_Q)$ such that

$$\mathcal{D}\big(R_S(G_Q), r\big) = \tfrac{1}{m}\Big[\mathrm{KL}(Q\|P) + \ln\tfrac{\mathcal{T}_{\mathcal{D}}(m,N)}{\delta}\Big], \quad (18)$$

where the exact value of $\mathcal{T}_{\mathcal{D}}(m,N)$ is computed according to Equation (6) for all $\mathcal{D}$-functions. For instance, $\mathcal{T}_{\mathcal{D}_{\mathrm{KL}}}(m,N)$ and $\mathcal{T}_{\mathcal{D}^{\star}_{m/N}}(m,N)$ are respectively given by Equations (7) and (10). Figure 1 highlights that, once the $\mathcal{D}$-function is chosen, the risk bound relies on a trade-off between the growing rate of $\mathcal{D}\big(R_S(G_Q), r\big)$ and the value of the right-hand side of Equation (18), influenced by the amplitude of $\mathcal{T}_{\mathcal{D}}(m,N)$.

Quite surprisingly, the KL-divergence $\mathcal{D}_{\mathrm{KL}}$, giving the best known PAC-Bayesian bounds on the inductive setting, is often less accurate than other $\mathcal{D}$-functions in the transductive setting. The bounds given by $\mathcal{D}_{\mathrm{KL}}$, $\mathcal{D}_{V^2}$ and $\mathcal{D}_{\triangle}$ are similar on Figure 1. The variation distance $\mathcal{D}_{V}$ gives the lowest bounds on small datasets, but loses its edge as N grows. The function $\mathcal{D}^{\star}_{m/N}$ shows good accuracies when m is the half of N (*i.e.*, the sizes of training set and unlabeled set are the same), and is especially tight when m is close to N (*i.e.*, there is a small amount of unlabeled examples). This phenomenon is related to the fact that $\mathcal{D}^{\star}_{m/N}$ is the only function that adjusts itself to the ratio $\frac{m}{N}$. Thereby, its value is always inside the interval of realizable risks, deduced from Equation (1) (provided that $0 \leq R_U(G_Q) \leq 1$):

$$\tfrac{m}{N} R_S(G_Q) \;\leq\; R_Z(G_Q) \;\leq\; \tfrac{m}{N} R_S(G_Q) + \tfrac{N-m}{N}. \quad (19)$$

In the supplementary material, we present analogous experimentations for other training set risks $R_S(G_Q)$.

3.2 Bound Values on Natural Data

We now compare bound values on reasonably large binary classification datasets, coming from the UCI Machine Learning Repository [Blake and Merz, 1998]. For each dataset of N examples, we draw at random (without replacement) a training set of m examples, for ratios m/N of 0.1 and 0.5. To obtain a posterior distribution Q on a set of classifiers, we run the AdaBoost algorithm [Schapire and Singer, 1999], using *decision stumps* as weak classifiers, for 200 rounds.

In Table 1, for each dataset and m/N ratio, we compute the risk of the Gibbs classifier on the full sample Z and the training set S. We compute explicit bounds of Corollary 7-(b) and Equation (15) (stated as "Derbeko"). We also compute bounds from Theorem 5, using $\mathcal{D} = \mathcal{D}_{\mathrm{KL}}$ and $\mathcal{D} = \mathcal{D}^{\star}_{m/N}$, as they are the most interesting choices in this setting (see the discussion of Section 3.1). Finally, we compute the risks of the majority vote classifier on S and Z, a bound of $R_Z(B_Q)$ using twice the bound from Theorem 5 with $\mathcal{D} = \mathcal{D}^{\star}_{m/N}$, and that same bound converted to a bound on $\mathcal{C}^Z_Q$ using Theorem 8. Recall that any bound on the Gibbs classifier's risk can be converted to a bound on $\mathcal{C}^Z_Q$. All bounds were calculated with $\delta = 0.05$.

Table 1 confirms that the explicit bound of Corollary 7 outperforms the bound of Derbeko et al. [2004]. It also corroborates observations from Figure 1, as the bound from Theorem 5 with $\mathcal{D} = \mathcal{D}_{\mathrm{KL}}$ performs well when m/N is low, and the version with $\mathcal{D} = \mathcal{D}^{\star}_{m/N}$ performs well when m/N is high. Finally, we see that when a non-trivial bound on $\mathcal{C}^Z_Q$ can be calculated (*i.e.*, when twice the Gibbs classifier's risk bound is lower than 1), the resulting bound on the risk of the majority vote classifier is most of the time much tighter than twice the bound on the Gibbs classifier's risk.

4 CONCLUSION

We have presented a transductive general PAC-Bayesian theorem that allows one to use any convex function $\mathcal{D}$ of the risk on the training set and the risk on the full sample. Each choice of $\mathcal{D}$-function leads to a new transductive bound, and all such bounds can be calculated, but possibly at a computational price if the full sample size N is big. For the bound obtained with the $\mathcal{D}^{\star}_{\beta}$-function of Equation (8), we derived a tight closed form that can easily be calculated for any N.

In our analysis of the behavior of different proposed bounds, we noticed that the KL-divergence $\mathcal{D}_{\mathrm{KL}}$ does not always give rise to the tightest bound. This is in opposition with the inductive case. We also proposed a way to use the expected disagreement over the full sample to obtain much tighter bounds. These bounds really take into account the unlabeled examples.

Acknowledgements

Work supported by NSERC discovery grant 262067.

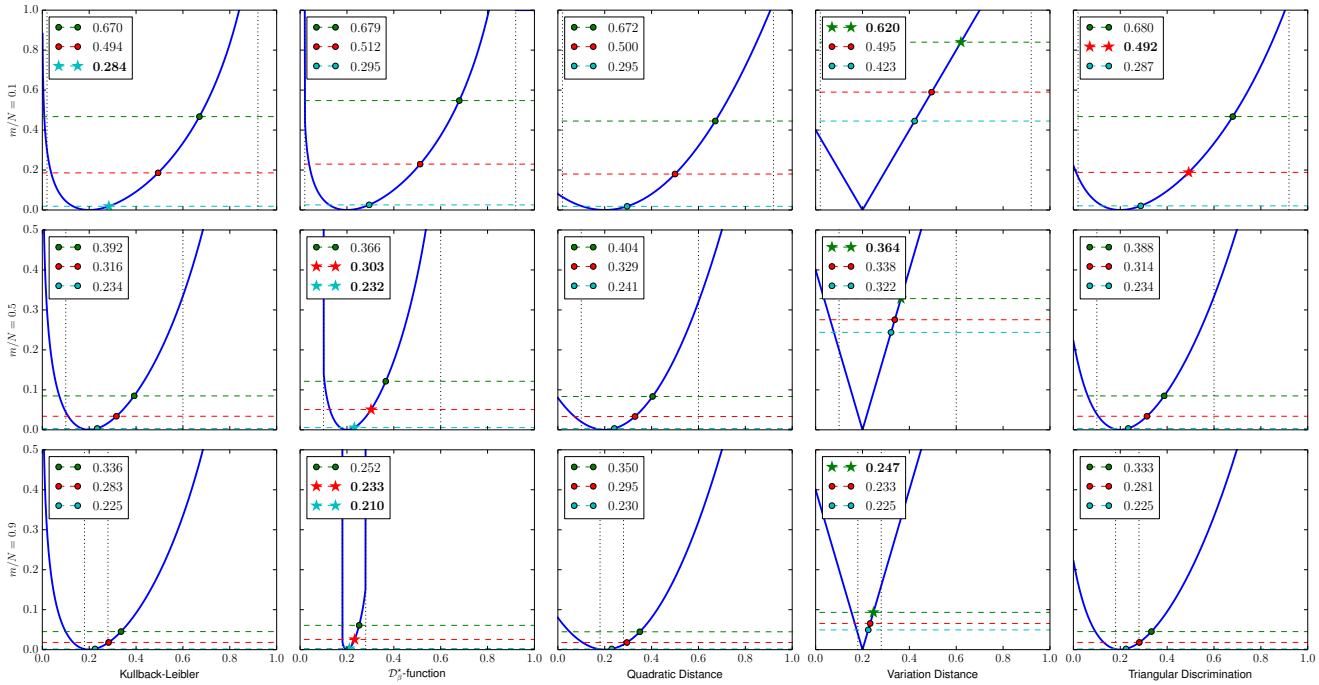

Figure 1: Study of the behavior of bounds obtained by Theorem 5. All graphics consider $R_S(G_Q) = 0.2$, $\mathrm{KL}(Q\|P) = 5$ and $\delta = 0.05$. The three graphics of each column share a particular $\mathcal{D}$-function, and the five graphics of each line share a particular ratio $\frac{m}{N}$. One graphic is interpreted as follows. The two vertical dotted lines show the minimum and the maximum value of $R_Z(G_Q)$ (see Equation (19)). The blue curve corresponds to the function $\mathcal{D}(0.2, r)$. Each dashed horizontal line corresponds to the value given by $\frac{1}{m}\big[\mathrm{KL}(Q\|P) + \ln\frac{\mathcal{T}_\mathcal{D}(m,N)}{\delta}\big]$ for three values of N: $N = 200$ (green line), $N = 500$ (red line), and $N = 5000$ (cyan line). On each of these lines, the location of the dot points out the bound value (*i.e.*, the r solving Equation (18)). Finally, this bound value is reported on the graphic legend. A star replaces the dot if the bound is the lowest obtained for all $\mathcal{D}$-functions.

Dataset information			Gibbs Classifier						Majority Vote Classifier			
			Observed Risk		Bounds of $R_Z(G_Q)$				Observed Risk		Bounds of $R_Z(B_Q)$	
Dataset	N	m/N	$R_S(G_Q)$	$R_Z(G_Q)$	Cor 7-(b)	Derbeko	Thm 5-$\mathcal{D}_{\mathrm{KL}}$	Thm 5-$\mathcal{D}^\star_{m/N}$	$R_S(B_Q)$	$R_Z(B_Q)$	$\mathcal{D}^\star_{m/N}$	$\mathcal{C}$-$\mathcal{D}^\star_{m/N}$
car	1728	0.1	0.193	0.194	**0.555**	0.793	**0.527**	0.546	0.105	0.159	1.092	-
car	1728	0.5	0.179	0.181	**0.418**	0.496	0.418	**0.415**	0.115	0.125	0.830	**0.819**
letter_AB	1555	0.1	0.146	0.149	**0.469**	0.718	**0.437**	0.457	0.000	0.017	**0.914**	0.961
letter_AB	1555	0.5	0.171	0.171	**0.402**	0.485	0.401	**0.399**	0.000	0.001	0.797	**0.626**
mushroom	8124	0.1	0.202	0.202	**0.486**	0.609	**0.471**	0.482	0.000	0.000	**0.964**	0.966
mushroom	8124	0.5	0.205	0.205	**0.439**	0.479	0.438	0.438	0.000	0.000	0.875	**0.546**
nursery	12959	0.1	0.169	0.168	**0.404**	0.504	**0.389**	0.399	0.009	0.016	0.798	**0.692**
nursery	12959	0.5	0.167	0.168	**0.357**	0.391	0.356	0.356	0.010	0.012	0.711	**0.379**
optdigits	3823	0.1	0.208	0.213	**0.533**	0.703	**0.513**	0.527	0.000	0.077	1.055	-
optdigits	3823	0.5	0.210	0.211	**0.460**	0.516	0.460	**0.458**	0.026	0.042	0.917	**0.793**
pageblock	5473	0.1	0.199	0.201	**0.495**	0.642	**0.476**	0.490	0.048	0.063	**0.979**	0.992
pageblock	5473	0.5	0.208	0.208	**0.448**	0.497	0.448	**0.447**	0.057	0.059	0.894	**0.697**
pendigits	7494	0.1	0.209	0.210	**0.499**	0.629	**0.481**	0.495	0.023	0.051	**0.989**	0.997
pendigits	7494	0.5	0.215	0.215	**0.457**	0.500	**0.455**	0.456	0.041	0.045	0.912	**0.706**
segment	2310	0.1	0.206	0.207	**0.558**	0.769	**0.533**	0.550	0.000	0.059	1.101	-
segment	2310	0.5	0.206	0.206	**0.462**	0.532	0.462	**0.460**	0.014	0.016	0.920	**0.834**
spambase	4601	0.1	0.222	0.227	**0.553**	0.708	**0.535**	0.548	0.115	0.161	1.096	-
spambase	4601	0.5	0.225	0.226	**0.488**	0.539	0.489	**0.486**	0.137	0.143	0.973	**0.961**

Table 1: Comparison of multiple bounds on the risks of the Gibbs classifier and the majority vote classifier.

References

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

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

Monroe D Donsker and SR Srinivasa Varadhan. Asymptotic evaluation of certain markov process expectations for large time. *Communications on Pure and Applied Mathematics*, 28(1):1–47, 1975.

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

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

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

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

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

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

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

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

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

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

Flemming Topsøe. Some inequalities for information divergence and related measures of discrimination. *IEEE Transactions on Information Theory*, 46(4): 1602–1609, 2000.

Vladimir Vapnik. *Statistical learning theory*. Wiley, 1998. ISBN 978-0-471-03003-4.