PAC-Bayesian Bounds based on the Rényi Divergence

Luc Bégin, Pascal Germain, François Laviolette, Jean-Francis Roy

PAC-Bayesian Bounds based on the Rényi Divergence

Luc Bégin[1 †] luc.begin@umoncton.ca
Pascal Germain[2 ‡] pascal.germain@inria.fr
François Laviolette[3] **Jean-Francis Roy**[3] {francois.laviolette, jean-francis.roy}@ift.ulaval.ca

[1] Campus d'Edmundston, Université de Moncton, Nouveau-Brunswick, Canada
[2] INRIA - Sierra Project-Team, École Normale Supérieure, Paris, France
[3] Département d'informatique et de génie logiciel, Université Laval, Québec, Canada

Abstract

We propose a simplified proof process for PAC-Bayesian generalization bounds, that allows to divide the proof in four successive inequalities, easing the "customization" of PAC-Bayesian theorems. We also propose a family of PAC-Bayesian bounds based on the Rényi divergence between the *prior* and *posterior* distributions, whereas most PAC-Bayesian bounds are based on the Kullback-Leibler divergence. Finally, we present an empirical evaluation of the tightness of each inequality of the simplified proof, for both the classical PAC-Bayesian bounds and those based on the Rényi divergence.

1 INTRODUCTION

Many learning algorithms output prediction functions that can be seen as a *weighted majority vote* of simpler functions (named the *voters* in this paper). Boosting [Schapire and Singer, 1999] and Random Forests [Breiman, 2001] are classical examples of ensemble methods that output a weight vector over a set of voters (such as decision trees). The dual form of many *kernel methods* can also be seen as majority votes, where each voter is the output of a *kernel* function. The PAC-Bayesian theory [McAllester, 1999] aims to provide *Probably Approximately Correct* (PAC) guarantees to learning algorithms that output a weighted majority vote. This approach considers a *prior* distribution P over the voters—that characterizes prior beliefs before observing any data—, and a *posterior* distribution Q—that takes into account the information provided by the training data. Distribution Q characterizes the output of the learning algorithm executed on the training data.

Classical PAC-Bayesian generalization bounds indirectly bound the *risk* of the (deterministic) majority vote classifier by bounding the risk of the (stochastic) *Gibbs classifier*. Given a family of voters $\mathcal{H}$ and a prior distribution P on $\mathcal{H}$, the general PAC-Bayesian theorem of Germain et al. [2009, 2015] bounds the *real* risk of the Gibbs classifier simultaneously for all posterior distributions Q using two main ingredients: a convex function $\Delta : [0,1]^2 \to \mathbb{R}$ that links the real and empirical risks of the Gibbs classifier, and a complexity term that depends on the *Kullback-Leibler* (KL) divergence between Q and P. Likewise, most PAC-Bayesian bounds on the risk of the Gibbs classifier depend on the KL divergence [*e.g.*, McAllester, 1999, Langford and Shawe-Taylor, 2002, Seeger, 2003].[1]

In this paper, we first provide a new proof of the general theorem of Germain et al. [2009, 2015], that streamlines the steps to four inequalities: Jensen's inequality, the *change of measure* inequality, Markov's inequality, and a supremum inequality. This proof helps to highlight each step that introduces looseness into the bound. Our new proof also eases forthcoming "customizations" of the proof to obtain novel bounds.

We later focus our study on the use of a new *change of measure inequality*, based on the *Rényi divergence*, alongside our proposed proving methodology. This quantity, that generalizes the KL divergence (see the extensive study of van Erven and Harremoës [2014]), gives rise to a family of PAC-Bayesian bounds that depend on the Rényi divergence instead of the usual KL

†All authors contributed equally to this work.

‡Most of this work was carried out while P. Germain was affiliated with Université Laval, Québec, Canada.

[1] Notable exceptions are bounds that consider restricted families of posterior distributions and have no divergence at all [Catoni, 2007, Parrado-Hernández et al., 2012, Lever et al., 2013, Germain et al., 2015].

Appearing in Proceedings of the 19th International Conference on Artificial Intelligence and Statistics (AISTATS) 2016, Cadiz, Spain. JMLR: W&CP volume 51. Copyright 2016 by the authors.

divergence. Furthermore, as a particular case of this new result, we state a bound based on the *Chi-squared divergence*, which is very similar to the one of Honorio and Jaakkola [2014].

We finally make use of the simplified proof to provide the first empirical analysis that evaluates each of the bound's inequalities, opening the way for a better understanding of the parts that induce a tightness loss.

The paper is organized as follows. Section 2 introduces the classical PAC-Bayesian result and presents the new "customizable" proof approach. Section 3 introduces the new family of bounds based on the Rényi divergence. Section 4 provides the empirical evaluation of the proof steps for both bounds based on the KL and Rényi divergences, and we conclude in Section 5.

2 A FRESH LOOK AT PAC-BAYESIAN PROOFS

In this section we first present basic definitions and notation, we recall the classical PAC-Bayesian theorem and present our new streamlined proof.

2.1 The Setting

Let us consider an arbitrary input space $\mathcal{X}$ and a binary output space $\mathcal{Y}=\{-1,1\}$. The examples $(x,y)\in\mathcal{X}\times\mathcal{Y}$ are input-output pairs; x is a description, and y is a label. We study the inductive learning setting where each example (x,y) is drawn *i.i.d.* from an unknown probability distribution D on $\mathcal{X}\times\mathcal{Y}$. Given a *training set* $S=\{(x_i,y_i)\}_{i=1}^m\sim D^m$, a machine learning algorithm builds a *classifier* $h:\mathcal{X}\to\mathcal{Y}$ that is later used to classify new examples drawn from D. The *risk* of a classifier h on a distribution D is the probability that h misclassifies an example,

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

and the *empirical risk* of h on a discrete set S is

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

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

In the PAC-Bayesian framework, we consider a hypothesis space $\mathcal{H}$ of classifiers, a *prior* distribution P on $\mathcal{H}$, and a *posterior* distribution Q on $\mathcal{H}$. The *prior* is specified before exploiting the information contained in S, while the *posterior* is obtained by running a learning algorithm on S. The PAC-Bayesian theory usually studies the stochastic *Gibbs classifier* G_Q. Given a distribution Q on $\mathcal{H}$, G_Q classifies an example x by drawing at random a classifier h according to Q, and returns $h(x)$. The risk of G_Q is then defined as follows.

Definition 1. For any probability distribution Q on a set of voters, the *Gibbs risk* $R_D(G_Q)$ is the expected risk of the Gibbs classifier G_Q relative to D. Hence,

$$R_D(G_Q) \;=\; \mathop{\mathbf{E}}_{(x,y)\sim D}\mathop{\mathbf{E}}_{h\sim Q} I\Big[h(x)\neq y\Big]\,.$$

Usual PAC-Bayesian bounds give guarantees on the generalization risk $R_D(G_Q)$.[2] Typically, these bounds rely on the empirical risk $R_S(G_Q)$,

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

and the Kullback-Leibler divergence between the prior and posterior distributions, as defined below.

Definition 2 (Kullback-Leibler divergence)**.** The Kullback-Leibler divergence between distributions Q and P is given by

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

Note that throughout this paper, we will always suppose that the support of Q is included in the support of P, that is, if $P(h)=0$, we also have $Q(h)=0$.

2.2 Change of Measure Inequality

A key step of most PAC-Bayesian proofs is summarized by the following *change of measure inequality* [Seldin and Tishby, 2010, McAllester, 2013, Germain et al., 2015]. Note that the same result is derived from Fenchel's inequality [Banerjee, 2006] and Donsker-Varadhan's variational formula for relative entropy [Seldin et al., 2012, Tolstikhin and Seldin, 2013].

Lemma 3 (Kullback-Leibler change of measure)**.** *For any set $\mathcal{H}$, for any distributions P and Q on $\mathcal{H}$, and for any measurable function $\phi:\mathcal{H}\to\mathbb{R}$, we have*

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

Proof idea. The result is obtained by exploiting the definition of the KL divergence (Definition 2), and then by using Jensen's inequality on the concave function $\ln(\cdot)$. □

[2] A part of PAC-Bayesian literature studies how to convert the Gibbs risk into the more commonly used Bayes risk (*i.e.*, the risk of the deterministic majority vote classifier). While twice the Gibbs risk upper-bounds the Bayes risk [*e.g.*, Herbrich and Graepel, 2000], tighter bounds are obtained by specializing to linear classifiers [Langford and Shawe-Taylor, 2002], or by exploiting the "voters' disagreement" [Germain et al., 2015]. Based on these works, tighter Gibbs bounds lead to tighter Bayes ones.

2.3 Customizable Proof

The statement of the following PAC-Bayesian theorem originally comes from Germain et al. [2009, 2015]. Note that, even if the proof presented below incorporate ideas from many other works [*e.g.*, McAllester, 1999, Langford and Shawe-Taylor, 2002, Seeger, 2003], the approach is new. In particular, this allows to divide the proof in four successive inequalities, as presented schematically by Figure 1 (left-hand side). As we will see in Section 3, this approach eases the "customization" of the proof: one can replace a particular step to tailor the theorem to his need. Also, the proof highlights all approximations leading to the risk bound, as studied empirically in Section 4.

Theorem 4 relies on the choice of a convex function $\Delta : [0,1]\times[0,1] \to \mathbb{R}$, that measures the "distance" between the observed empirical Gibbs risk $R_S(G_Q)$ and the *true* Gibbs risk $R_D(G_Q)$ on distribution D. By upper-bounding the value of this Δ-*function*, Theorem 4 provides an interval in which lies $R_D(G_Q)$ with high probability. The extremities of this interval give both a lower bound and an upper bound of $R_D(G_Q)$.

Theorem 4. *For any distribution D on $\mathcal{X}\times\mathcal{Y}$, for any set $\mathcal{H}$ of voters $\mathcal{X} \to \{-1,1\}$, for any prior distribution P on $\mathcal{H}$, for any $\delta\in(0,1]$, for any $m'>0$, and for any convex function $\Delta : [0,1]\times[0,1] \to \mathbb{R}$, with probability at least $1-\delta$ over the choice of $S\sim D^m$, we have*

$$\forall Q \text{ on } \mathcal{H}:\quad \Delta\Big(R_S(G_Q), R_D(G_Q)\Big) \le \frac{1}{m'}\left[\mathrm{KL}(Q\|P) + \ln\frac{\mathcal{I}^{\mathrm{K}}_{\Delta}(m,m')}{\delta}\right],$$

with $$\mathcal{I}^{\mathrm{K}}_{\Delta}(m,m') \overset{\text{def}}{=} \sup_{r\in[0,1]}\left[\sum_{k=0}^{m}\mathrm{Bin}_k^m\big(r\big)e^{m'\Delta(\frac{k}{m},r)}\right], \quad (1)$$

and $\mathrm{Bin}_k^m\big(r\big)$ is the binomial probability mass function:

$$\mathrm{Bin}_k^m\big(r\big) \overset{\text{def}}{=} \tbinom{m}{k}\big(r\big)^k\big(1-r\big)^{m-k}.$$

Proof. To upper-bound $\Delta\Big(R_S(G_Q), R_D(G_Q)\Big)$, we apply Jensen's inequality on convex function Δ, and Donsker-Varadhan's change of measure (Lemma 3) with $\phi(f){=}m'\Delta\Big(R_S(h), R_D(h)\Big)$. Hence, $\forall Q$ on $\mathcal{H}$:

$$\begin{aligned} & m'\Delta\Big(R_S(G_Q), R_D(G_Q)\Big)\\ &= m'\Delta\Big(\mathop{\mathbf{E}}_{h\sim Q}R_S(h), \mathop{\mathbf{E}}_{h\sim Q}R_D(h)\Big)\\ &\le \mathop{\mathbf{E}}_{h\sim Q} m'\Delta\Big(R_S(h), R_D(h)\Big)\\ &\le \mathrm{KL}(Q\|P) + \ln\Bigg(\underbrace{\mathop{\mathbf{E}}_{h\sim P} e^{m'\Delta\big(R_S(h),R_D(h)\big)}}_{X_P(S)}\Bigg). \end{aligned}$$

Now, consider the random variable

$$X_P(S) = \mathop{\mathbf{E}}_{h\sim P} e^{m'\Delta\big(R_S(h),R_D(h)\big)},$$

and apply Markov's inequality to obtain

$$\Pr_{S\sim D^m}\left(X_P(S) \le \frac{1}{\delta}\mathop{\mathbf{E}}_{S'\sim D^m} X_P(S')\right) \ge 1-\delta\,.$$

This, in turn, implies that with probability at least $1-\delta$ over the choice of $S\sim D^m$, we have $\forall\, Q$ on $\mathcal{H}$:

$$m'\Delta\big(R_S(G_Q), R_D(G_Q)\big) \le \mathrm{KL}(Q\|P) + \ln\frac{\mathop{\mathbf{E}}_{S'\sim D^m} X_P(S')}{\delta}. \quad (2)$$

We now upper-bound $\mathbf{E}\,X_P(S')$, first by swapping the expectations over D^m and over P, and then using the fact that the number of errors $mR_{S'}(h)$ follows a binomial distribution[3] with parameters m and $R_D(h)$:

$$\begin{aligned} &\mathop{\mathbf{E}}_{S'\sim D^m} X_P(S') && (3)\\ &= \mathop{\mathbf{E}}_{S'\sim D^m}\mathop{\mathbf{E}}_{h\sim P} e^{m'\Delta(R_{S'}(h),R_D(h))}\\ &= \mathop{\mathbf{E}}_{h\sim P}\mathop{\mathbf{E}}_{S'\sim D^m} e^{m'\Delta(R_{S'}(h),R_D(h))}\\ &= \mathop{\mathbf{E}}_{h\sim P}\sum_{k=0}^{m}\Pr_{S'\sim D^m}\big(R_{S'}(h)=\tfrac{k}{m}\big)\,e^{m'\Delta(\frac{k}{m},R_D(h))}\\ &= \mathop{\mathbf{E}}_{h\sim P}\sum_{k=0}^{m}\mathrm{Bin}_k^m\Big(R_D(h)\Big)e^{m'\Delta(\frac{k}{m},R_D(h))} && (4)\\ &\le \sup_{r\in[0,1]}\left[\sum_{k=0}^{m}\mathrm{Bin}_k^m\big(r\big)\,e^{m'\Delta(\frac{k}{m},r)}\right]\\ &= \mathcal{I}^{\mathrm{K}}_{\Delta}(m,m')\,. \end{aligned}$$

The final result is obtained by replacing $\mathbf{E}\,X_P(S')$ by its upper bound $\mathcal{I}^{\mathrm{K}}_{\Delta}(m,m')$ inside Equation (2). $\square$

Note that usual PAC-Bayesian theorems use $m'=m$. In this particular case, we use the shorthand notation $\mathcal{I}^{\mathrm{K}}_{\Delta}(m) \overset{\text{def}}{=} \mathcal{I}^{\mathrm{K}}_{\Delta}(m,m)$.

2.4 Some Choices of Δ-Functions

As discussed in Germain et al. [2009, 2015], Theorem 4 is a generic tool to derive various inductive PAC-Bayesian bounds, as Δ may be any convex function. However, one needs to calculate (or upper-bound) the value of $\mathcal{I}^{\mathrm{K}}_{\Delta}(m,m')$ to express a computable bound. A common choice is $\Delta=\Delta_{\mathrm{KL}}$, the Kullback-Leibler (KL) divergence between two Bernoulli distributions of probability of success p and q, defined by

$$\Delta_{\mathrm{KL}}(q,p) \overset{\text{def}}{=} q\ln\tfrac{q}{p} + (1-q)\ln\tfrac{1-q}{1-p}\,. \quad (5)$$

[3] Maurer [2004] allows to generalize the PAC-Bayesian theorem to real-valued voters $\mathcal{X}\to[-1,1]$. In this case, one can replace the equality between Lines (3) to (4) with an inequality ($\le$) and the statement of Theorem 4 holds.

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

$$\mathcal{I}^{\rm K}_{\Delta_{\rm KL}}(m) \;=\; \sum_{k=0}^{m} \binom{m}{k} \left(\tfrac{k}{m}\right)^k \left(1-\tfrac{k}{m}\right)^{m-k}. \qquad (6)$$

Hence, it is straightforward to compute the exact value of $\mathcal{I}^{\rm K}_{\Delta_{\rm KL}}(m)$. However, this computation can be time-consuming when m is large. To avoid the computation of the sum of Equation (6), it is also possible to upper bound the value of $\mathcal{I}^{\rm K}_{\Delta_{\rm KL}}(m)$ with a simpler expression. Indeed, Maurer [2004] shows the following:

$$\sqrt{m} \;\le\; \mathcal{I}^{\rm K}_{\Delta_{\rm KL}}(m) \;\le\; 2\sqrt{m}\,. \qquad (7)$$

This leads to the following PAC-Bayesian bound, attributed to Seeger [2002] (in the former result, $m{+}1$ appeared instead of $2\sqrt{m}$).

Corollary 5 (Seeger [2002])**.** *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 $\mathcal{H}$:

$$\Delta_{\rm KL}\big(R_S(G_Q), R_D(G_Q)\big) \le \tfrac{1}{m}\Big[{\rm KL}(Q\|P) + \ln \tfrac{2\sqrt{m}}{\delta}\Big].$$

Another common PAC-Bayesian result of McAllester [2003] is obtained by using the following Δ-function:

$$\Delta_{V^2}(q,p) \;\stackrel{\rm def}{=}\; 2(q-p)^2\,. \qquad (8)$$

Using the fact that $\Delta_{\rm KL}(q,p) \ge \Delta_{V^2}(q,p)$ (which is known as Pinsker's inequality), the result of Equation (7) gives $\mathcal{I}^{\rm K}_{\Delta_{V^2}}(m) \le 2\sqrt{m}$. This allows us to state the following explicit PAC-Bayesian bound.

Corollary 6 (McAllester [2003])**.** *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 $\mathcal{H}$:

$$R_D(G_Q) \;\le\; R_S(G_Q) + \sqrt{\tfrac{1}{2m}\Big[{\rm KL}(Q\|P) + \ln \tfrac{2\sqrt{m}}{\delta}\Big]}.$$

Other choices of Δ-functions lead to different bounds that can be found in the literature. For instance, using $\Delta_c(q,p) = \ln \frac{e^{-c\,q}}{1-p\,(1-e^{-c})}$ for any constant $c > 0$ leads to the bound of Catoni [2007]. We can also recover bounds that are similar to the ones of Pentina and Lampert [2015] and Alquier et al. [2015] by considering a linear function $\Delta_{\rm lin}(q,p) = p - q$. In the transductive learning setting [Vapnik, 1998], where one has access to a subset of m labeled examples drawn from a set of N examples to classify, using $\Delta_{{\rm KL},\beta}(q,p) = \Delta_{\rm KL}(q,p) + \frac{1-\beta}{\beta}\Delta_{\rm KL}(\frac{p-\beta q}{1-\beta}, p)$ with $\beta = \frac{m}{N}$ leads to the PAC-Bayesian bounds of Derbeko et al. [2004] and Bégin et al. [2014]. The latter also experiment with other Δ-functions in the transductive setting, such as the variation distance $\Delta_V(q,p) = 2|p-q|$ and the triangular discrimination $\Delta_\triangle(q,p) = \frac{(q-p)^2}{q+p} + \frac{(q-p)^2}{2-q-p}$.

In the next section, we customize the proof of Theorem 4 by introducing a change of measure inequality based on the Rényi divergence.

3 FROM THE KL-DIVERGENCE TO THE RÉNYI DIVERGENCE

We first introduce the Rényi divergence [Rényi, 1961], on which we will base a new change of measure inequality and a new family of PAC-Bayesian bounds.

Definition 7 (Rényi divergence)**.** For any $\alpha > 1$, the Rényi divergence between distributions Q and P is given by

$$D_\alpha(Q\|P) \;\stackrel{\rm def}{=}\; \frac{1}{\alpha-1}\ln\left[\mathop{\mathbf{E}}_{h\sim P}\left(\frac{Q(h)}{P(h)}\right)^{\alpha}\right],$$

where $D_\alpha(Q\|P) = {\rm KL}(Q\|P)$ when α tends to 1.

It is noteworthy that the value of $D_\alpha(Q\|P)$ is always greater to or equal than ${\rm KL}(Q\|P)$. Moreover, given a uniform prior $U_\mathcal{H}$ over $\mathcal{H}$ and a posterior $U_{\mathcal{H}'}$ which is uniform over a subset $\mathcal{H}' \subseteq \mathcal{H}$, the KL divergence and the Rényi divergence are equal for any α value. In particular, when $\mathcal{H}$ is a discrete set, we have $U_\mathcal{H}(h){=}\frac{1}{|\mathcal{H}|}$ for all $h \in \mathcal{H}$, and $U_{\mathcal{H}'}(h){=}\frac{1}{|\mathcal{H}'|}$ for all $h \in \mathcal{H}'$ or $U_{\mathcal{H}'}(h){=}0$ otherwise. Therefore, $\forall \alpha \in (1,\infty)$:

$$D_\alpha\big(U_{\mathcal{H}'}\|U_\mathcal{H}\big) \;=\; {\rm KL}\big(U_{\mathcal{H}'}\|U_\mathcal{H}\big) \;=\; -\ln\left(\frac{|\mathcal{H}'|}{|\mathcal{H}|}\right).$$

This corresponds to the case where distribution $U_{\mathcal{H}'}$ describes a democratic majority vote classifier, like those output by *Bagging* and *Random Forests* learning algorithms.

3.1 Change of Measure Inequality

We now present a change of measure inequality that, instead of being based on the Kullback-Leibler divergence like this is the case in the usual Lemma 3, is based on the Rényi divergence of Definition 7.

Theorem 8 (Rényi change of measure)**.** *For any set $\mathcal{H}$, for any distributions P and Q on $\mathcal{H}$, for any $\alpha > 1$, and for any measurable function $\phi : \mathcal{H} \to \mathbb{R}$, we have*

$$\frac{\alpha}{\alpha{-}1}\ln \mathop{\mathbf{E}}_{h\sim Q}\phi(h) \;\le\; D_\alpha(Q\|P) + \ln\left(\mathop{\mathbf{E}}_{h\sim P}\phi(h)^{\frac{\alpha}{\alpha-1}}\right).$$

	KL-divergence	Rényi divergence with $\alpha'=\frac{\alpha}{\alpha-1}$
	$\Delta\Big(\mathbf{E}_{h\sim Q} R_S(h), \mathbf{E}_{h\sim Q} R_D(h)\Big)$	$\ln\Delta\Big(\mathbf{E}_{h\sim Q} R_S(h), \mathbf{E}_{h\sim Q} R_D(h)\Big)$
Jensen's inequality	$\leq \mathbf{E}_{h\sim Q}\, \Delta\Big(R_S(h), R_D(h)\Big)$	$\leq \ln\Big(\mathbf{E}_{h\sim Q}\, \Delta\Big(R_S(h), R_D(h)\Big)\Big)$
Change of measure	$\leq \frac{1}{m'}\Big[\mathrm{KL}(Q\\|P) + \ln\Big(\mathbf{E}_{h\sim P}\, e^{m'\Delta\left(R_S(h), R_D(h)\right)}\Big)\Big]$	$\leq \frac{1}{\alpha'}\Big[D_\alpha(Q\\|P) + \ln\Big(\mathbf{E}_{h\sim P}\, \Delta\big(R_S(h), R_D(h)\big)^{\alpha'}\Big)\Big]$
Markov's inequality	$\underset{1-\delta}{\leq} \frac{1}{m'}\Big[\mathrm{KL}(Q\\|P) + \ln\Big(\frac{1}{\delta}\mathbf{E}_{S'\sim D^m}\mathbf{E}_{h\sim P}\, e^{m'\Delta(R_{S'}(h), R_D(h))}\Big)\Big]$	$\underset{1-\delta}{\leq} \frac{1}{\alpha'}\Big[D_\alpha(Q\\|P) + \ln\Big(\frac{1}{\delta}\mathbf{E}_{S'\sim D^m}\mathbf{E}_{h\sim P}\, \Delta\big(R_{S'}(h), R_D(h)\big)^{\alpha'}\Big)\Big]$
Expectations swap	$= \frac{1}{m'}\Big[\mathrm{KL}(Q\\|P) + \ln\Big(\frac{1}{\delta}\mathbf{E}_{h\sim P}\mathbf{E}_{S'\sim D^m}\, e^{m'\Delta(R_{S'}(h), R_D(h))}\Big)\Big]$	$= \frac{1}{\alpha'}\Big[D_\alpha(Q\\|P) + \ln\Big(\frac{1}{\delta}\mathbf{E}_{h\sim P}\mathbf{E}_{S'\sim D^m}\, \Delta\big(R_{S'}(h), R_D(h)\big)^{\alpha'}\Big)\Big]$
Binomial law	$= \frac{1}{m'}\Big[\mathrm{KL}(Q\\|P) + \ln\Big(\frac{1}{\delta}\mathbf{E}_{h\sim P}\sum_{k=0}^{m}\mathrm{Bin}_k^m\big(R_D(h)\big) e^{m'\Delta\left(\frac{k}{m}, R_D(h)\right)}\Big)\Big]$	$= \frac{1}{\alpha'}\Big[D_\alpha(Q\\|P) + \ln\Big(\frac{1}{\delta}\mathbf{E}_{h\sim P}\sum_{k=0}^{m}\mathrm{Bin}_k^m\big(R_D(h)\big)\,\Delta\big(\frac{k}{m}, R_D(h)\big)^{\alpha'}\Big)\Big]$
Supremum over risk	$\leq \frac{1}{m'}\Big[\mathrm{KL}(Q\\|P) + \ln\Big(\frac{1}{\delta}\sup_{r\in[0,1]}\Big\{\sum_{k=0}^{m}\mathrm{Bin}_k^m(r)\, e^{m'\Delta\left(\frac{k}{m}, r\right)}\Big\}\Big)\Big]$	$\leq \frac{1}{\alpha'}\Big[D_\alpha(Q\\|P) + \ln\Big(\frac{1}{\delta}\sup_{r\in[0,1]}\Big\{\sum_{k=0}^{m}\mathrm{Bin}_k^m(r)\,\Delta\big(\frac{k}{m}, r\big)^{\alpha'}\Big\}\Big)\Big]$

Figure 1: Proof sketch comparing the classical PAC-Bayesian bound of Theorem 4 (on the left) with the new bound based on the Rényi divergence of Theorem 9 (on the right), using the proof process introduced in Section 2.3. The symbol $\underset{1-\delta}{\leq}$ denotes that the inequality holds with probability at least $1-\delta$.

Proof. We first change the expectation over Q for an expectation over P, and then apply Hölder's inequality with $r=\alpha$ and $s=\frac{\alpha}{\alpha-1}$. More precisely, we have

$$
\begin{aligned}
&\frac{\alpha}{\alpha-1}\ln \mathbf{E}_{h\sim Q}\phi(h)\\
&\leq \frac{\alpha}{\alpha-1}\ln \mathbf{E}_{h\sim P}\left[\frac{Q(h)}{P(h)}\,\phi(h)\right]\\
&\leq \frac{\alpha}{\alpha-1}\ln\left(\left[\mathbf{E}_{h\sim P}\left(\frac{Q(h)}{P(h)}\right)^{\alpha}\right]^{\frac{1}{\alpha}}\left[\mathbf{E}_{h\sim P}\phi(h)^{\frac{\alpha}{\alpha-1}}\right]^{\frac{\alpha-1}{\alpha}}\right)\\
&= \frac{1}{\alpha-1}\ln\left[\mathbf{E}_{h\sim P}\left(\frac{Q(h)}{P(h)}\right)^{\alpha}\right] + \ln\left[\mathbf{E}_{h\sim P}\phi(h)^{\frac{\alpha}{\alpha-1}}\right]\\
&= D_\alpha(Q\|P) + \ln\left[\mathbf{E}_{h\sim P}\phi(h)^{\frac{\alpha}{\alpha-1}}\right].
\end{aligned}
$$

Note that Hölder's inequality holds when $\frac{1}{r}+\frac{1}{s}=1$, which is the case for these choices of r and s. □

Theorem 8, with $\phi(h)$ replaced by $e^{(\alpha-1)\phi(h)}$, has been presented in Atar and Merhav [2015, Equation (8)] as the *risk-sensitive functional comparison bounds*[4] (see also Atar et al. [2015, Corollary 2.4]). The proof presented in this paper is much simpler. Note also that function ϕ in Atar and Merhav [2015] is required to be bounded, and this limitation is not necessary here. However, Theorem 8 is not interesting in situations where ϕ is not bounded, as the right-hand side of the inequality is infinite.

[4] Atar and Merhav [2015] use a different definition of the Rényi divergence that differs by a factor of α.

Observe that applying Jensen's inequality on the concave function $\ln(\cdot)$ of the left-hand side inequality of Theorem 8 (with $\phi(h)$ replaced by $e^{\frac{\alpha-1}{\alpha}\phi(h)}$) gives rise to the following looser change of measure inequality that is also based on the Rényi divergence:

$$
\mathbf{E}_{h\sim Q}\phi(h) \;\leq\; D_\alpha(Q\|P) + \ln\left(\mathbf{E}_{h\sim P} e^{\phi(h)}\right). \qquad (9)
$$

This inequality has the same form as Lemma 3, with the $\mathrm{KL}(Q\|P)$ divergence replaced with $D_\alpha(Q\|P)$. New PAC-Bayesian bounds could be derived using this inequality, but would however be looser than traditional ones as the Rényi divergence has a higher value than the KL divergence for all $\alpha>1$. For this reason, in this paper we will always rely on Theorem 9 below, instead of using the bound that one can derive from Equation (9).

3.2 Bounds Based on the Rényi Divergence

We now present the main result of this paper. Note that the proof of Theorem 9, below, follows the "customizable" approach introduced in Section 2.3. This highlights that our new PAC-Bayesian proof is based on the same inequalities that the usual ones (see Theorem 4), except that we substitute the *Kullback-Leibler* change of measure (Lemma 3) with the *Rényi* change of measure (Theorem 8). Figure 1 presents sketches of the proofs that allow to compare the two approaches.

Theorem 9. *For any distribution D on $\mathcal{X}\times\mathcal{Y}$, for any set $\mathcal{H}$ of voters $\mathcal{X}\to\{-1,1\}$, for any prior distribution P on $\mathcal{H}$, for any $\delta\in(0,1]$, for any $\alpha>1$, and for any convex function $\Delta:[0,1]\times[0,1]\to\mathbb{R}$, with probability at least $1-\delta$ over the choice of $S\sim D^m$, we have*

$$\forall Q \text{ on } \mathcal{H}: \quad \ln \Delta\Big(R_S(G_Q), R_D(G_Q)\Big) \leq \frac{1}{\alpha'}\left[D_\alpha(Q\|P) + \ln \frac{\mathcal{I}^{\mathrm{R}}_\Delta(m,\alpha')}{\delta}\right],$$

where $\alpha'=\frac{\alpha}{\alpha-1}$, and

$$\mathcal{I}^{\mathrm{R}}_\Delta(m,\alpha') \overset{\text{def}}{=} \sup_{r\in[0,1]}\left[\sum_{k=0}^m \mathrm{Bin}_k^m\big(r\big)\,\Delta\big(\tfrac{k}{m},\,r\big)^{\alpha'}\right]. \quad (10)$$

Proof. We apply Jensen's inequality on the convex function $\Delta(\cdot,\cdot)$, and Rényi change of measure (Theorem 8) with $\phi(h)=\Delta\Big(R_S(h), R_D(h)\Big)$. Hence,$\forall Q$ on $\mathcal{H}$:

$$\begin{aligned}
&\alpha' \ln \Delta\Big(R_S(G_Q), R_D(G_Q)\Big)\\
&= \alpha' \ln \Delta\Big(\mathop{\mathbf{E}}_{h\sim Q} R_S(h), \mathop{\mathbf{E}}_{h\sim Q} R_D(h)\Big)\\
&\leq \alpha' \ln \mathop{\mathbf{E}}_{h\sim Q} \Delta\Big(R_S(h), R_D(h)\Big)\\
&\leq D_\alpha(Q\|P) + \ln\Bigg(\underbrace{\mathop{\mathbf{E}}_{h\sim P} \Delta\big(R_S(h), R_D(h)\big)^{\alpha'}}_{X_P(S)}\Bigg).
\end{aligned}$$

Let $X_P(S) = \mathbf{E}_{h\sim P}\,\Delta\big(R_S(h), R_D(h)\big)^{\alpha'}$. By Markov's inequality, we have, with probability at least $1-\delta$ over the choice of $S\sim D^m$, $\forall\, Q$ on $\mathcal{H}$:

$$\alpha' \ln \Delta\big(R_S(G_Q), R_D(G_Q)\big) \leq D_\alpha(Q\|P) + \ln \frac{\mathop{\mathbf{E}}_{S'\sim D^m} X_P(S')}{\delta}. \quad (11)$$

We now upper-bound $\mathbf{E}\,X_P(S')$ by applying the same steps that in the proof of Theorem 4 (from Line (3)).

$$\begin{aligned}
\mathop{\mathbf{E}}_{S'\sim D^m} X_P(S') &= \mathop{\mathbf{E}}_{S'\sim D^m}\mathop{\mathbf{E}}_{h\sim P} \Delta(R_{S'}(h), R_D(h))^{\alpha'}\\
&\leq \sup_{r\in[0,1]}\left[\sum_{k=0}^m \mathrm{Bin}_k^m\big(r\big)\,\Delta\big(\tfrac{k}{m},\,r\big)^{\alpha'}\right]\\
&= \mathcal{I}^{\mathrm{R}}_\Delta(m,\alpha').
\end{aligned}$$

The final statement is obtained by replacing $\mathbf{E}\,X_P(S')$ by its upper bound $\mathcal{I}^{\mathrm{R}}_\Delta(m,\alpha')$ in Equation (11). □

When comparing the bounds of Theorems 4 and 9, we see that both can be parameterized, using m' for the bounds based on the KL divergence, and using α for those relying on the Rényi divergence. In the latter, the value of α also impacts the divergence value. We also notice that the Δ-function appears as an exponent in Theorem 4, and as the base of an exponent in Theorem 9. As the values might be much smaller in the latter, this opens the way to exploring alternatives for the remaining steps of the proof. We discuss an alternative in concluding remarks (Section 5).

Theorem 9 is stated as an upper bound on the log of the chosen Δ-function to ease the comparison with Theorem 4, as its right-hand side has a similar form. To bound the Δ-function directly, one can simply apply an exponential function on both sides of Theorem 9 inequality. Then, by simple arithmetic, we obtain

$$\Delta\Big(R_S(G_Q), R_D(G_Q)\Big) \leq \left[\mathop{\mathbf{E}}_{h\sim P}\Big(\tfrac{Q(h)}{P(h)}\Big)^{\alpha}\right]^{\frac{1}{\alpha}}\left[\tfrac{\mathcal{I}^{\mathrm{R}}_\Delta(m,\alpha')}{\delta}\right]^{\frac{1}{\alpha'}}.$$

By choosing $\alpha=2$ (and therefore $\alpha'=2$) in the latter equation, we obtain an interesting special case of Theorem 9 that relies on the chi-squared divergence $\chi^2(Q\|P) \overset{\text{def}}{=} \mathop{\mathbf{E}}_{h\sim P}\Big[\Big(\frac{Q(h)}{P(h)}\Big)^2 - 1\Big]$. With this observation, and the linear function $\Delta_{\mathrm{lin}}(q,p) = p-q$, we obtain Corollary 10 below, which turns out to be similar to Honorio and Jaakkola [2014, Lemma 7]. This previous result cannot be directly compared to ours, as it applies to a parameterized family of linear classifiers in a different setting than the one we study. Nevertheless, Corollary 10 does have a smaller complexity term, due to the factor $\frac{1}{4}$ inside the square root.

Corollary 10. *For any distribution D on $\mathcal{X}\times\mathcal{Y}$, for any set $\mathcal{H}$ of voters $\mathcal{X}\to\{-1,1\}$, for any prior distribution P on $\mathcal{H}$, and for any $\delta\in(0,1]$, with probability at least $1-\delta$ over the choice of $S\sim D^m$, we have*

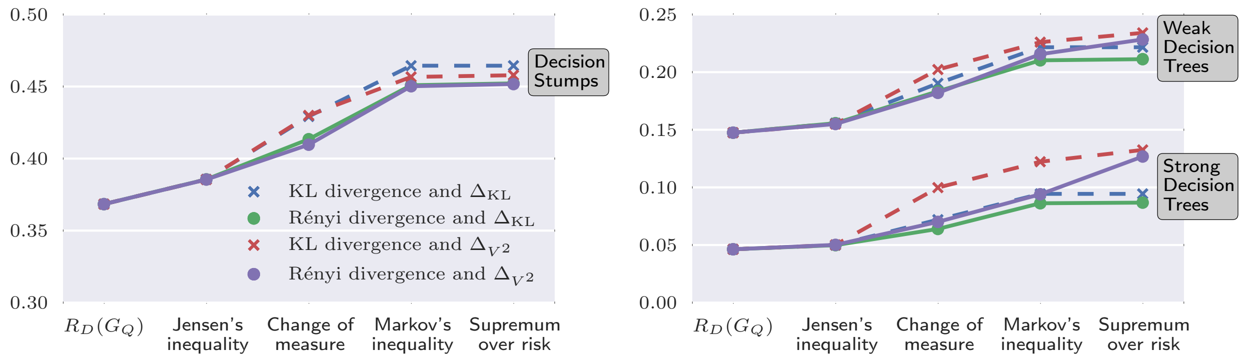
$$\forall Q \text{ on } \mathcal{H}: \; R_D(G_Q) \leq R_S(G_Q) + \sqrt{\frac{\chi^2(Q\|P)+1}{4m\,\delta}}.$$

Proof. We apply Theorem 9 with $\alpha=2$ and $\Delta=\Delta_{\mathrm{lin}}$. In this case, the value of Equation (10) turns out to be the variance of a binomial random variable (with m trials and success $r=\frac{1}{2}$) divided by m^2:

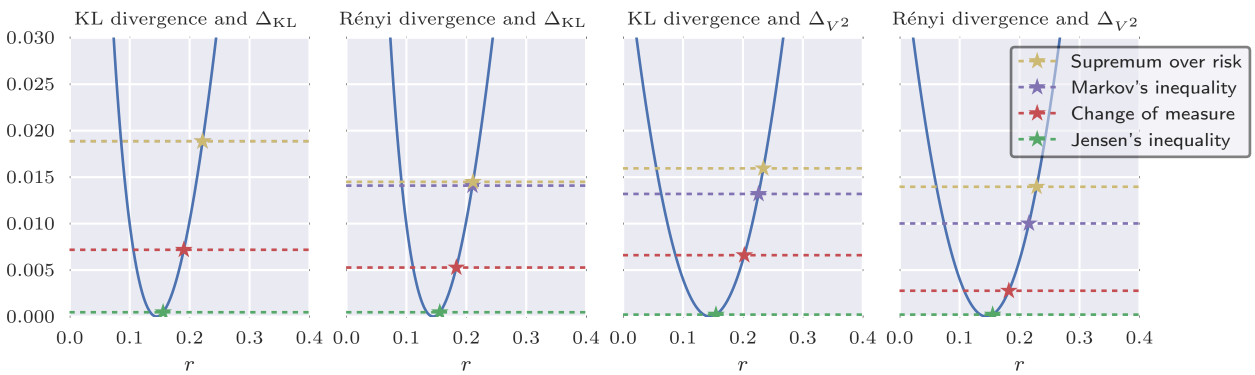
$$\begin{aligned}
\mathcal{I}^{\mathrm{R}}_{\Delta_{\mathrm{lin}}}(m,2) &= \frac{1}{m^2}\sup_{r\in[0,1]}\left[\sum_{k=0}^m \mathrm{Bin}_k^m\big(r\big)\,(m\,r-k)^2\right]\\
&= \frac{1}{m^2}\sup_{r\in[0,1]}\Big[m\,r(1-r)\Big] \;=\; \frac{1}{4m}.
\end{aligned}$$ □

4 EMPIRICAL STUDY

The following experiments compare the accuracy of the new PAC-Bayesian bounds based on the Rényi divergence, to the usual ones based on the KL divergence. Moreover, we aim to study the effect of each inequality used to state the bound (see Figure 1). To do so, as we need to know every quantity intervening at each step of the proof including the data-generating distribution D,



(a) Values for each inequality computed with for three kinds of voters: *decision stumps*, *weak decision trees* and *strong decision trees*. The dashed lines correspond to the traditional bounds with the Kullback-Leibler divergence. The full lines correspond to the bounds considering the Rényi divergence. The value at last step gives the final bound. The majority vote risk on these experiments is 0.01 using decision stumps, 0.001 using weak decision trees and 0.002 using strong decision trees (see Footnote 2 for more details about the links between the Gibbs and the majority vote risks).



(b) Alternate representation of the quantities obtained using the weak decision trees. The blue curve corresponds to the function $\Delta(R_D(G_Q), r)$. Each dashed horizontal line corresponds to the value given by the right-hand side of the bound after each inequality. On each of these lines, the location of the star gives the value of the inequality (on the x axis). Note that on the leftmost figure, the supremum inequality is an equality (as the KL-based bound with Δ_{KL} offers an analytic value for the supremum), and thus the horizontal line appears directly over the line related to Markov's inequality.

Figure 2: Values for each inequality of the proof process of Theorems 4 and 9, applied with the KL divergence between two Bernoulli distributions Δ_{KL} of Equation (5), and the quadratic distance Δ_{V^2} of Equation (8).

we consider the following synthetic distribution. Each example generated by D is a random draw among the 8124 examples of the *mushroom* dataset coming from the UCI Machine Learning Repository [Lichman, 2013]. That is, the training set $S \sim D^m$ contains m examples drawn *with replacement* and *uniform probability* from the full dataset. From training set S, we learn a majority vote using AdaBoost [Schapire and Singer, 1999]. We conduct three experiments with different kinds of voters:

– *Decision Stumps.* For each of the 22 attributes of mushroom dataset, we build 10 decision stumps with equally distributed thresholds between the minimum and the maximum values of the attribute. For each so obtained voter, we also consider its inverse. Thus, we obtain a total of 440 weak voters.

– *Weak Decision Trees.* We generate 500 decision trees using the *scikit-learn* library [Pedregosa et al., 2011]. Each tree is learned using 100 examples randomly selected among the full mushroom dataset.[5] We set parameters *depth* $= 3$, and *max_features* $= 2$.

– *Strong Decision Trees.* We generate 500 decision trees using the same procedure described above, but with parameters *depth* $= 6$, and *max_features* $= 5$.

In all three experiments, we set the prior to be a uniform distribution over the above described vot-

[5] Note that the bounds are only valid when the voters must not rely on training examples. As our goal is to study the behavior of the bounds using voters of different capabilities, the decision trees simulate the situation where one has strong prior knowledge on the data distribution.

ers. We use two Δ-functions: the Kullback-Leibler divergence between two Bernoulli distributions Δ_{KL} and the quadratic distance Δ_{V^2}. Recall that these Δ-function allow to recover Corollaries 5 and 6 respectively when using the KL change of measure and $m' = m$. In our experiments, we observed that choosing $m' = m$ for KL-based bounds and $\alpha = 1.1$ for Rényi-based bounds provides near-optimal bound values, regardless the values of other quantities intervening in the bound expression. We present the results obtained for these choices. We do not show results using the linear distance Δ_{lin} and $\alpha = 2$ giving Corollary 10, as the resulting bounds were significantly looser.

The four steps displayed in Figure 2 correspond to the four inequalities of the PAC-Bayesian proof (see the proof sketch of Figure 1). For example, the values displayed at *Jensen's inequality* step, for an experiment with the KL divergence and the Δ-function Δ_{KL}, is computed by finding the value $r \geq R_S(G_Q)$ such that $\Delta_{\mathrm{KL}}\big(R_S(G_Q), r\big) = \sum_{h\in\mathcal{H}} Q(h)\Delta_{\mathrm{KL}}\big(R_S(h), R_D(h)\big)$.

Similarly, the value of the *Change of measure* step is computed by finding r such that $m\,\Delta_{\mathrm{KL}}\big(R_S(G_Q), r\big) = \mathrm{KL}(Q\|P) + \ln \sum_{h\in\mathcal{H}} P(h)\exp\big(m\Delta_{\mathrm{KL}}(R_S(h), R_D(h))\big)$.

The two remaining steps are computed using the same method. Note that the final inequality is a supremum over continuous value r, and therefore must be approximated when the choice of Δ-function does not provide a closed-form expression. As our experiments show that the argument of the supremum is smooth and only have one or two local maximums, a simple root finding method such as the classic *Brent method* [Brent, 1973] can be used to obtain a precise approximation.

Using the weak decision trees and inequality values of Figure 2a, Figure 2b puts in relation the value of each Δ-function (in function of the empirical Gibbs risk) with the right-hand side value of each inequality of the proof process. This figure offers a different view of the same experiment, and helps understanding the impact of the choice if a Δ-function.

We observe that, for a given majority vote and a given Δ-function, the final bounds obtained with the Rényi approach are *slightly* tighter than the traditional Kullback-Leibler approach.[6] With weak voters, we observe that the *change of measure* proof step is significantly tighter with the Rényi bounds than with the KL ones (Theorem 8 versus Lemma 3). However, this edge is lost in further steps, mainly when applying *Markov's inequality*. Note that *Markov's inequality* is not problematic with our strongest voters. In this case, the *supremum over risk* step degrades the accuracy of the Rényi bound used with the quadratic function Δ_{V^2}.

[6]Note that this observation does not rely on our specific choice of m' value and α values. Indeed, we observed that the Rényi bound with the *best* α value is always tighter than the KL bound with the best m'.

5 CONCLUSION & FUTURE WORK

We exposed a "customizable" PAC-Bayesian proving methodology relying on four inequalities steps. We showed that when replacing the usual *Kullback-Leibler change of measure* step by a new *Rényi change of measure* (Theorem 8), we obtain a PAC-Bayesian theorem (Theorem 9) that allows to expresses a new family of generalization bounds. We empirically studied these bounds by comparing them to usual ones. The Rényi based bounds are *slightly* tighter, but it turns out that other steps of the proving process counteract the gain obtained by the new change of measure.

Nevertheless, we think that our proving scheme can motivate interventions on other inequality steps to improve the bound value. In particular, we have seen that Markov's inequality step is loose in the context of weak voters. We plan to replace the Markov inequality by the Chebyshev inequality, that would take into account the variance of the studied random variable.[7] We also plan to explore the relations of our proving scheme with the *Occam's Hammer* bound of Blanchard and Fleuret [2007].

Finally, the new bounds provided in this work are not explicit (except for Corollary 10 that leads to deceptive empirical bound values). Therefore, they may be less attractive for practitioners than the classical PAC-Bayesian bound of McAllester [2003] (Corollary 6). To state an explicit bound, one first needs to find a Δ-function such that the function $\mathcal{I}^{\mathrm{R}}_{\Delta}(m, \alpha')$ of Equation (10) is upper-bounded by a closed-form expression. New explicit bounds may be a source of inspiration for designing learning algorithms. So far, most algorithms derived from PAC-Bayesian bounds are KL-regularized [*e.g.* Germain et al., 2009, Parrado-Hernández et al., 2012, Pentina and Lampert, 2015, Alquier et al., 2015]. Our new result might lead to a different kind of regularization.

Acknowledgements

This work has been supported by National Science and Engineering Research Council (NSERC) Discovery grant 262067.

[7]This variance is huge in classical bounds, as the random variable relies on the exponential of the Δ-function (*i.e.*, $e^{m'\Delta(\cdot,\cdot)}$). Thus, the Chebyshev inequality is of little use for bounds based on the KL change of measure, but might lead to an improvement in our new Rényi bounds, as the Δ-function appears at the base of the exponent (*i.e.*, $\Delta(\cdot,\cdot)^{\alpha'}$). See the definition of $X_P(S)$ in Theorems 4 and 9 proofs to compare KL and Rényi *Markov's inequality* step.

References

Pierre Alquier, James Ridgway, and Nicolas Chopin. On the properties of variational approximations of Gibbs posteriors. *ArXiv e-prints*, 1506.04091, 2015.

Rami Atar and Neri Merhav. Information-theoretic applications of the logarithmic probability comparison bound. *IEEE International Symposium on Information Theory (ISIT)*, 2015.

Rami Atar, Kenny Chowdhary, and Paul Dupuis. Robust bounds on risk-sensitive functionals via renyi divergence. *SIAM/ASA Journal on Uncertainty Quantification*, 3(1):18–33, 2015.

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

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

Gilles Blanchard and François Fleuret. Occam's hammer. In *COLT*, volume 4539 of *Lecture Notes in Computer Science*, pages 112–126. Springer, 2007.

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

Richard P Brent. *Algorithms for Minimization Without Derivatives*. Courier Corporation, 1973.

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

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.

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

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*, 16:787–860, 2015. URL `http://jmlr.org/papers/v16/germain15a.html`.

Ralf Herbrich and Thore Graepel. A PAC-Bayesian margin bound for linear classifiers: Why svms work. In *NIPS*, pages 224–230, 2000.

Jean Honorio and Tommi Jaakkola. Tight bounds for the expected risk of linear classifiers and PAC-Bayes finite-sample guarantees. In *AISTATS*, pages 384–392, 2014.

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

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

M. Lichman. UCI machine learning repository, 2013. URL `http://archive.ics.uci.edu/ml`.

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, 2003.

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

Emilio Parrado-Hernández, Amiran Ambroladze, John Shawe-Taylor, and Shiliang Sun. PAC-bayes bounds with data dependent priors. *Journal of Machine Learning Research*, 13:3507–3531, 2012.

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

Anastasia Pentina and Christoph H. Lampert. Lifelong learning with non-i.i.d. tasks. In *NIPS*, 2015.

Alfréd Rényi. On measures of entropy and information. In *Fourth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 547–561, 1961.

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.

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

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

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

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

Tim van Erven and Peter Harremoës. Rényi divergence and Kullback-Leibler divergence. *IEEE Trans. Information Theory*, 60(7):3797–3820, 2014.

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