Tuning the Distribution Dependent Prior in the PAC-Bayes Framework based on Empirical Data

Luca Oneto[1], Sandro Ridella[2], and Davide Anguita[1]

1 - DIBRIS - University of Genova
Via Opera Pia 13, I-16145 Genova - Italy

2 - DITEN - University of Genova
Via Opera Pia 11A, I-16145 Genova - Italy

Abstract. In this paper we further develop the idea that the PAC-Bayes prior can be defined based on the data-generating distribution. In particular, following Catoni [1], we refine some recent generalisation bounds on the risk of the Gibbs Classifier, when the prior is defined in terms of the data generating distribution, and the posterior is defined in terms of the observed one. Moreover we show that the prior and the posterior distributions can be tuned based on the observed samples without worsening the convergence rate of the bounds and with a marginal impact on their constants.

1 Introduction

It is well known that combining the output of several classifiers results in much better performance than using any one of them alone. In fact many state-of-the-art algorithms search for a weighted combination of simpler classifiers [2]: Bagging [3], Boosting [4] and Bayesian approaches [5] and, in some sense, Kernel methods [6] and Neural Networks [7]. The major open problem in this scenario is how to weight the different classifiers in order to obtain good performance [8, 9, 1], and how this performance can be assessed [10, 6, 11, 12, 13, 14]. The PAC-Bayes approach [15, 11, 1, 16, 17, 2] is one of the sharpest analysis frameworks in this context, since it can provide a tight bound on the risk of the Gibbs Classifier (GC), and the Bayes Classifier (BC) [2]. The GC choses a classifiers in a set according to a posterior distribution each time a new sample has to be classified [17]. In particular, in the PAC-Bayes framework, a prior distribution over the classifiers must be defined before seeing the data, then, based on the available data, a posterior distribution is chosen, and the risk of the associate GC is estimated, based on the empirical risk and the divergence between the prior and posterior distributions [11]. The PAC-Bayes analysis bounds the risk of the GC [11, 16], while the $\mathcal{C}$-bound bounds the error of the BC, also called weighted majority vote classifier, based on the properties of the GC [2].

The major weakness in the conventional PAC-Bayes approach is that a posterior distribution that minimises the divergence between prior and posterior distributions must be chosen, since this divergence is part of the bound [17, 18]. In order to address this issue, Catoni [1] proposed a localised PAC-Bayes analysis, which exploys a Boltzmann prior distribution defined in terms of the unknown data distribution. Note that, since the prior depends on the data generating distribution, the PAC-Bayes analysis is still valid because the prior is defined before observing the data [1]. By tuning the prior to the distribution, Catoni was able to remove the divergence term from the bound, hence significantly reducing the complexity penalty [1]. More recently, this approach has been further developed in [17] but the prior still has a free parameter that needs to be fixed before observing the data and can affect the divergence penalty and consequently the tightness of the bound.

In this paper, in Section 3, we further refine the bound proposed in [17] and recalled in Section 2. Moreover, in Section 4 we show that the distribution dependent Boltzmann prior distribution developed by Catoni [1] can be tuned based on the observed samples in order to optimise the bounds over the risk of the GC. In particular we will show that this optimisation does not change the convergence rate of the bounds and has a marginal impact over the constants involved in the bounds.

2 State-of-the-art Results

Let us consider a set of n labeled samples $\mathcal{D}_n = \{(X_1, Y_1), \cdots, (X_n, Y_n)\} = \{Z_1, \cdots, Z_n\}$ drawn i.i.d. according to an unknown probability distribution μ over the cartesian product between the input space $\mathcal{X}$ and the output space $\mathcal{Y}{=}\{-1, 1\}$. Let us consider a function $f{\in}\mathcal{F}$, where $f{:}\mathcal{X}{\to}\overline{\mathcal{Y}}{=}[-1, 1]$. The error of f in approximating μ is measured with reference to some $[0, 1]$-bounded loss function $\ell{:}\mathcal{F}{\times}(\mathcal{X}{\times}\mathcal{Y}){\to}[0, 1]$. Then the risk of f, and its empirical estimator, can be defined as $L^\ell(f){=}\mathbb{E}_Z\{\ell(f, Z)\}$ and $\widehat{L}^\ell(f){=}{}^1\!/\!_n \sum_{i=1}^n \ell(f, Z_i)$. The GC draws a function $f{\in}\mathcal{F}$, according to a probability distribution Q over $\mathcal{F}$, each time a label for an input $X{\in}\mathcal{X}$ is required. For the GC, referred as G_Q, we can define its risk together with its empirical counterpart [17]: $L^\ell(G_Q){=}\mathbb{E}_{f\sim Q}\{L^\ell(f)\}$, and $\widehat{L}^\ell(G_Q){=}\mathbb{E}_{f\sim Q}\{\widehat{L}^\ell_{\text{emp}}(f)\}$. Given two probability distributions Q and P over $\mathcal{F}$, let us denote with $\mathtt{KL}[Q||P]$ the Kullback-Leibler Divergence (KLD) between P and Q, while $\mathtt{kl}[q||p]$ is the KLD for the Binomial distribution $\mathtt{kl}[q||p]{=}q\ln[{}^q\!/\!_p]+[1{-}q]\ln[{}^{1-q}\!/\!_{1-p}]$ where, thanks to the Pinsker's Inequality we can state that $|q{-}p|{\leq}\sqrt{{}^1\!/\!_2\mathtt{kl}[q||p]}$. Finally, let us recall the definition of a last fundamental quantity in the PAC-Bayes framework, which is a weighted sum of binomial coefficients $\xi_n{=}\sum_{k=0}^n \binom{n}{k}\left(\frac{k}{n}\right)^k\left(1{-}\frac{k}{n}\right)^{n-k}$ where $\sqrt{n}{\leq}\xi_n{\leq}2\sqrt{n}$ [16]. Based on these preliminary definitions we can recall the state of the art bound on the risk of the BC.

Theorem 1 ([16]). *For any probability distribution P over $\mathcal{F}$, chosen before seeing $\mathcal{D}_n$, $\forall Q$ we have* $\mathbb{P}\{\mathtt{kl}[\widehat{L}^\ell(G_Q)||L^\ell(G_Q)]{\geq}{}^{[\mathtt{KL}[Q||P]+\ln[\xi_n/\delta]]}\!/\!_n\}{\leq}\delta$.

The main problem of the PAC-Bayes Theory regards the choice of P and Q. Q should fit our observations, but, at the same time, Q should be close to P, in order the minimise the KLD. The milestone result of [1], later extended by [17], proposes to use a Boltzmann prior distribution P which depends on the data generating distribution μ. In particular, let us suppose that the density function associated to P is $p(f)=c_p e^{-\gamma L^\ell(f)}$, where $\gamma\in[0,\infty)$ and c_p is a normalisation term. Basically, this distribution gives more importance to functions that possess small risk. If we choose as posterior Q a distribution which gives more importance to functions with small empirical risk with the following density function $q(f)=c_q e^{-\gamma \widehat{L}^\ell(f)}$, where c_q is a normalisation term, it can be proved that this theorem, built on the result of Theorem 1, holds.

Theorem 2 ([17]). *Given the prior P and the posterior Q defined above, we can state that $\mathbb{P}\{\mathtt{KL}[Q||P]\geq\mathtt{KL}_1(\gamma,\delta,n)\doteq\gamma^2/n+\gamma\sqrt{2\ln[\xi_n/\delta]/n}\}\leq 2\delta$. Consequently, we have that $\mathbb{P}\{\mathtt{kl}[\widehat{L}^\ell(G_Q)||L^\ell(G_Q)]\geq[\mathtt{KL}_1(\gamma,\delta,n)+\ln[\xi_n/\delta]]/n\}\leq 3\delta$.*

3 Sharpening the Risk Bound

In this part of the paper we prove that the result of Theorem 2 can be further improved (Theorem 3). Furthermore, we show that, if the loss function exploited for assessing the risk of the BC is the same used to define the prior and posterior defined by Catoni [1] (generally they may be different), then Theorem 3 can be further improved (Theorem 4).

Theorem 3. *Under the same hypothesis of Theorem 2, we can state the following inequality $\mathbb{P}\{\mathtt{KL}[Q||P]\geq\mathtt{KL}_2(\gamma,\delta,n)\}\leq 2\delta$ where $\mathtt{KL}_2(\gamma,\delta,n)\leq\mathtt{KL}_1(\gamma,\delta,n)$ and $\mathtt{KL}_2(\gamma,\delta,n)\doteq\gamma\sqrt{\gamma\sqrt{2\ln[1/\delta]/n}/4n+\ln[\xi_n/\delta]/2n+\gamma^2/16n^2}+\gamma\sqrt{\ln[1/\delta]/2n}+\gamma^2/4n$. Moreover, we also have that $\mathbb{P}\{\mathtt{kl}[\widehat{L}^\ell(G_Q)||L^\ell(G_Q)]\geq[\mathtt{KL}_2(\gamma,\delta,n)+\ln[\xi_n/\delta]]/n\}\leq 3\delta$.*

Proof. Let us consider $\mathtt{KL}[Q||P]$, since $p(f)=c_p e^{-\gamma L^\ell(f)}$, then $c_p=p(f)e^{\gamma L^\ell(f)}$ and consequently:

$$\begin{aligned}&\mathtt{KL}[Q||P]=\mathbb{E}_{f\sim Q}\{\ln[q(f)/p(f)]\}=\mathbb{E}_{f\sim Q}\left\{\ln\left[c_q e^{-\gamma\widehat{L}^\ell(f)}/c_p e^{-\gamma L^\ell(f)}\right]\right\}=\ln[c_q/c_p]\\&+\mathbb{E}_{f\sim Q}\{\gamma[L^\ell(f)-\widehat{L}^\ell(f)]\}=\gamma[L^\ell(G_Q)-\widehat{L}^\ell(G_Q)]-\ln\textstyle\int_{\mathcal{F}}p(f)e^{\gamma[L^\ell(f)-\widehat{L}^\ell(f)]}df\\&\leq\gamma[L^\ell(G_Q)-\widehat{L}^\ell(G_Q)]-\gamma[L^\ell(G_P)-\widehat{L}^\ell(G_P)].\end{aligned}\tag{1}$$

By bounding the last two terms in square brackets through the Pinsker's inequality and Theorem 1 and by solving with respect to $\mathtt{KL}[Q||P]$, Theorem 2 can be derived. This is indeed sub-optimal. Since P is defined before seeing $\mathcal{D}_n$ we can exploit the Hoeffding's inequality [19] in order to bound the second term and, consequently, with probability $(1-2\delta)$, we have that:

$$\gamma[L^\ell(G_Q)-\widehat{L}^\ell(G_Q)]-\gamma[L^\ell(G_P)-\widehat{L}^\ell(G_P)]\leq\gamma\sqrt{\frac{\mathtt{KL}[Q||P]+\ln[\xi_n/\delta]}{2n}}+\gamma\sqrt{\frac{\ln[1/\delta]}{2n}}.\tag{2}$$

By bounding Eq. (1) through Eq. (2), solving with respect to $\mathtt{KL}[Q||P]$, and plugging the result in Theorem 1, the statement in this Theorem is proved. □

Theorem 4. *Under the same hypothesis of Theorem 2, if the losses used to define P, Q, $\widehat{L}^\ell(G_Q)$ and $L^\ell(G_Q)$ are the same, we can state the following inequality $\mathbb{P}\left\{\mathtt{kl}[\widehat{L}^\ell(G_Q)||L^\ell(G_Q)]\geq\left[\gamma|L^\ell(G_Q)-\widehat{L}^\ell(G_Q)|+\gamma\sqrt{\ln[1/\delta]/2n}+\ln[\xi_n/\delta]\right]/n\right\}\leq 2\delta$.*

Proof. The proof can be derived from Eq. (1). Since the losses used to define P, Q, $\widehat{L}^\ell(G_Q)$ and $L^\ell(G_Q)$ are the same we simply have to bound the last one of the two terms in square brackets of Eq. (1) through the Hoeffding's inequality [19], as in Theorem 3, and plug the result in Theorem 1. □

The results of Theorems 3 and 4 improve over the state-of the-art bound of Theorem 2.

4 Tuning the Prior on the Available Data

Unfortunately, even if an optimal choice of γ that minimises the above bounds exists, this parameter must be chosen before seeing the data in order to maintain their validity. In this section we deal with this problem by tuning γ based on the available data. The first step consists in proving the following theorem which bounds the risk of the GC when, given P^γ and Q^γ defined in Section 2 for different values of $\gamma \in \{\gamma_1, \cdots, \gamma_m\}$, one choses the γ_i with $i \in \{1, \cdots, m\}$ that minimises the bound on the GC risk.

Theorem 5. *Given the prior P^γ and Q^γ defined in Section 2 for different values of $\gamma \in \{\gamma_1, \cdots, \gamma_m\}$, we can state that, $\forall \gamma \in \{\gamma_1, \cdots, \gamma_m\}$:*

1. $\mathbb{P}\{\mathtt{kl}[\widehat{L}^\ell(G_Q)||L^\ell(G_Q)] \geq {}^{[\mathtt{KL}_1(\gamma,\delta,n) + \ln[\xi_n/\delta]]}/_n\} \leq 3m\delta$,
2. $\mathbb{P}\{\mathtt{kl}[\widehat{L}^\ell(G_Q)||L^\ell(G_Q)] \geq {}^{[\mathtt{KL}_2(\gamma,\delta,n) + \ln[\xi_n/\delta]]}/_n\} \leq 3m\delta$.

Moreover, if the losses used to define P^γ, Q^γ, $\widehat{L}^\ell(G_{Q^\gamma})$, and $L^\ell(G_{Q^\gamma})$ are the same, we can state that, $\forall \gamma \in \{\gamma_1, \cdots, \gamma_m\}$:

3. $\mathbb{P}\left\{\mathtt{kl}[\widehat{L}^\ell(G_Q)||L^\ell(G_Q)] \geq \left[\gamma|L^\ell(G_Q) - \widehat{L}^\ell(G_Q)| + \gamma\sqrt{\ln[1/\delta]/2n} + \ln[\xi_n/\delta]\right]/n\right\} \leq 2m\delta$.

Proof. In order to prove the statement the union bound must by applied over the different $\gamma \in \{\gamma_1, \cdots, \gamma_m\}$ to Theorems 2, 3 and 4. □

Given the result of Theorem 5, and since $\mathtt{KL}_2(\gamma,\delta,n) \leq \mathtt{KL}_1(\gamma,\delta,n)$ (Theorem 3), if γ is chosen among $m=(\xi_n)^\eta$ with $\eta \geq 0$ points equally spaced in logarithmic scale in $[\gamma_{\min}, \gamma_{\max}]$, we can state that with probability $(1-\delta)$:

$$\mathtt{kl}[\widehat{L}^\ell(G_Q)||L^\ell(G_Q)] \leq \frac{\mathtt{KL}_2(\gamma, {}^{\delta}/_{3\xi_n^\eta}, n) + (1+\eta)\ln[\xi_n] + \ln[{}^3/_\delta]}{n} \tag{3}$$

$$= \frac{\gamma\sqrt{\frac{\gamma\sqrt{2\eta\ln[\xi_n]+2\ln[3/\delta]}}{4n} + \frac{(1+\eta)\ln[\xi_n]+\ln[3/\delta]}{2n} + \frac{\gamma^2}{16n^2}} + \gamma\sqrt{\frac{\eta\ln[\xi_n]+\ln[3/\delta]}{2n}} + \frac{\gamma^2}{4n} + (1+\eta)\ln[\xi_n] + \ln[{}^3/_\delta]}{n}$$

$$\leq \frac{\mathtt{KL}_1(\gamma, {}^{\delta}/_{3\xi_n^\eta}, n) + (1+\eta)\ln[\xi_n] + \ln[{}^3/_\delta]}{n} = \frac{\frac{\gamma^2}{n} + \gamma\sqrt{\frac{2(1+\eta)\ln[\xi_n]+2\ln[3/\delta]}{n}} + (1+\eta)\ln[\xi_n] + \ln[{}^3/_\delta]}{n}$$

Moreover, if the losses used to define P^γ, Q^γ, $\widehat{L}^\ell(G_{Q^\gamma})$, and $L^\ell(G_{Q^\gamma})$ are the same, we can state that with probability $(1-\delta)$:

$$\mathtt{kl}[\widehat{L}^\ell(G_Q)||L^\ell(G_Q)] \leq \frac{\gamma|L^\ell(G_Q) - \widehat{L}^\ell(G_Q)| + \gamma\sqrt{\frac{\eta\ln[\xi_n]+\ln[2/\delta]}{2n}} + (1+\eta)\ln[\xi_n] + \ln[{}^2/_\delta]}{n} \tag{4}$$

Note that when $m=1$ (which means that $\eta=0$ and consequently Theorems 2, 3, and 4 can be applied) the bound has the same convergence rate as when γ is chosen among $m=(\xi_n)^\eta \in [n^{\eta/2}, 2n^{\eta/2}]$ possible values at the expenses of a slightly worse constant $\eta \geq 0$.

5 Discussion

In order to get more insights on the proposed bounds, we test them on an artificial problem. In particular, a dataset is created, consisting of n samples in a bidimensional input space: $n/2$ are equally spaced on a circle of radius 1 and center $[-c, -c]^T$ while the others $n/2$ are equally spaced on a circle of radius 1 and center $[c, c]^T$. We choose $c \in \{1/2, 1\}$, $\gamma_{\min}=10^{-2}$, and $\gamma_{\max}=10^3$. The Hard Loss $\ell(f, Z)=1-Y\text{sign}[f(X)]/2$ is exploited for the P^γ, Q^γ, $\widehat{L}^\ell(G_{Q^\gamma})$, and $L^\ell(G_{Q^\gamma})$. We choose, as hypothesis space $\mathcal{F}$, all the possible linear separators in the input space. In Figure 1 we have reported the upper bound of $L^\ell(G_Q)$ obtained by applying Theorems 2, 3, 4, and 5 (results 1, 2, and 3) in different situations. In Figures 1(a) and 1(d) we report the comparison of Theorems 2, 3 and 4 for different values of γ: note that Theorems 3 and 4 improve over the state of the art bound of Theorems 2, in particular Theorem 4 is the sharpest one (note that in this case we cannot choose the γ for which the bound is minimum). In Figures 1(b) and 1(e) we report the comparison between Theorem 4 and the bound obtained by adopting the strategy of Theorem 5 when we look at $m=(\xi_n)^\eta$ values of γ: note that the bound is worse as soon as we increase η but in this case we can choose the γ which minimises the bound and the loss, in terms of tightness of the risk bound, is smaller with respect to the loss of choosing a wrong γ. Finally in Figures 1(c) and 1(f) we report the upper bound of $L^\ell(G_Q)$, for different value of n, for $\eta = 1$ and $\gamma=\gamma^*$, where γ^* is the one that gives the minimum of Theorem 5 (results 1, 2, and 3) as described in Section 4 against the ideal case when γ^* is known a priori and Theorem 2, 3, and 4 can be used directly: obviously the bounds are looser but the loss is negligible if we take into account that γ has been tuned based on the observed samples.

The result that we have just presented, even if preliminary, gives interesting insights on the Theorems reported in this paper. In particular, in this paper we have shown that our results improve over the state-of-the-art PAC-Bayes GC risk bounds and that is possible to tune, in a fully empirical fashion, both the prior and posterior PAC-Bayes distributions without impacting on the rates of the bounds and with a marginal impact on their constants.

References

[1] O. Catoni. *PAC-Bayesian Supervised Classification.* Institute of Mathematical Statistics, 2007.

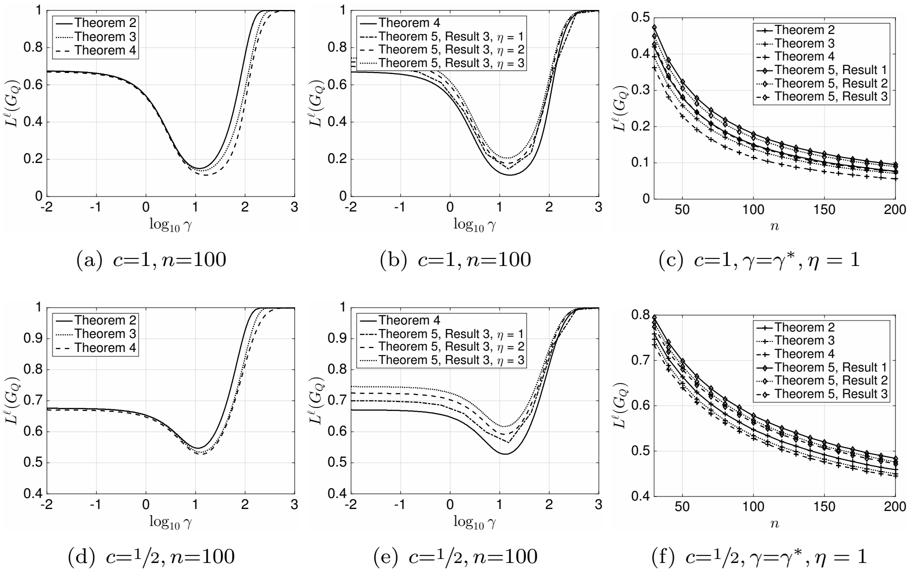
[2] P. Germain, A. Lacasse, A. Laviolette, M. Marchand, and Roy J. F. Risk bounds for the majority vote: From a pac-bayesian analysis to a learning algorithm. *JMLR*, 16(4):787–860, 2015.

[3] L. Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

[4] R. E. Schapire, Y. Freund, P. Bartlett, and W. S. Lee. Boosting the margin: A new explanation for the effectiveness of voting methods. *The Annals of Statistics*, 26(5):1651–1686, 1998.

[5] A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin. *Bayesian Data Analysis*, volume 2. Taylor & Francis, 2014.

[6] V. N. Vapnik. *Statistical Learning Theory.* Wiley New York, 1998.



(a) $c{=}1, n{=}100$

(b) $c{=}1, n{=}100$

(c) $c{=}1, \gamma{=}\gamma^*, \eta = 1$

(d) $c{=}1/2, n{=}100$

(e) $c{=}1/2, n{=}100$

(f) $c{=}1/2, \gamma{=}\gamma^*, \eta = 1$

Fig. 1: Performance of the different bounds over the artificial problem.

[7] B. M. Bishop. *Neural Networks for Pattern Recognition.* Oxford University Press, 1995.

[8] S. Nitzan and J. Paroush. Optimal decision rules in uncertain dichotomous choice situations. *International Economic Review*, 23(2):289–97, 1982.

[9] D. Berend and A. Kontorovitch. Consistency of weighted majority votes. In *NIPS*, 2014.

[10] S. Floyd and M. Warmuth. Sample compression, learnability, and the vapnik-chervonenkis dimension. *Machine Learning*, 21(3):269–304, 1995.

[11] D. A. McAllester. Some pac-bayesian theorems. In *Computational Learning Theory*, 1998.

[12] O. Bousquet and A. Elisseeff. Stability and generalization. *JMLR*, 2:499–526, 2002.

[13] P. L. Bartlett and S. Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *JMLR*, 3:463–482, 2003.

[14] P. L. Bartlett, O. Bousquet, and S. Mendelson. Local rademacher complexities. *Annals of Statistics*, pages 1497–1537, 2005.

[15] M. D. Donsker and S. R. S. Varadhan. Asymptotic evaluation of certain markov process expectations for large time, i. *Communications on Pure and Applied Mathematics*, 28(1):1–47, 1975.

[16] P. Germain, A. Lacasse, F. Laviolette, and M. Marchand. Pac-bayesian learning of linear classifiers. In *ICML*, 2009.

[17] G. Lever, F. Laviolette, and J. Shawe-Taylor. Tighter pac-bayes bounds through distribution-dependent priors. *Theoretical Computer Science*, 473:4–28, 2013.

[18] E. Parrado-Hernández, A. Ambroladze, J. Shawe-Taylor, and S. Sun. Pac-bayes bounds with data dependent priors. *JMLR*, 13(1):3507–3531, 2012.

[19] W. Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963.