# PAC-Bayesian Generalization Bound for Multi-class Learning

**Loubna BENABBOU**
Department of Industrial Engineering
Ecole Mohammadia d'Ingènieurs
Mohammed V University in Rabat, Morocco
Benabbou@emi.ac.ma

**Pascal LANG**
Operations and Decision Systems Department
Faculty of Business Administration
Laval University, Quèbec, Canada
Pascal.Lang@fsa.ulaval.ca

## Abstract

We generalize the PAC-Bayes theorem in the setting of multi-class learning. We set off by defining empirical and generalization risks in the multi-class setting with valued asymmetric loss function reflecting unequal gravity of misclassification and taking into account the option to not classify an example. We provide a simplified proof of the PAC-Bayes theorem in this multi-class setting. The generalization risk of a Gibbs Classifier is upper-bounded by its empirical risk and parameters of the multi-class classifier. We also formulate a convex mathematical program to estimate multi-class PAC-Bayes bound.

## 1 Introduction

PAC stands for Probably Approximately Correct; it was first used by Valiant in the PAC learnability (1984). The first PAC-Bayes bound was given in (McAllester (1998)) other tight versions can be find in (McAllester, (1999; 2003); Seeger, (2002;2003); Langford and Shawe-Taylor, 2003; Langford, 2005; Laviolette and Marchand, 2005; Maurer, 2006; Catoni, 2007; Lacasse et al., 2007; Germain et al., 2009; Laviolette et al., 2011, McAllester, 2013; Germain et al., 2016). To the best of our knowledge, bounds are generally developed for binary classifiers with (0-1) loss functions. The emerging of the multi-class learning problems in practice has stimulated the development of various bounds. Some attempts have been made to generalize PAC-Bayes theorem in multi-class case with a (0-1) loss function (Seeger, 2003; McAllester, 2013). The case of valued loss function has been treated generally in the binary classification (Seeger 2003; Germain et al. 2006). Morvant et al. (2012) have proposed a PAC-Bayes bounds based on confusion matrices of a multi-class classifier as an error measure. A PAC-Bayesian margin bound for generalization loss in structured classification has been suggested by Bartlett et al. (2004). In this paper we generalize PAC-Bayes theorem in more real-world multi-class setting with three characteristics: (*i*) Number of classes is more than 2 (*ii*) the possibility to not classify an example and (*iii*) loss function is valued and asymmetric reflecting unequal gravity of misclassification.

We are concerned with a multi-class problem in which each example $\tilde{z} = (\tilde{x}, \tilde{y}) \in \mathrm{Z}$ is constituted from an input-output pair $(x, y)$ where $x \in \mathrm{X}$ and $y \in \mathrm{Y}$; such that $|Z| = n$ and $|\mathrm{Y}| > 2$; a finite set of *observed* classes. We adopt the PAC setting where each example $\tilde{z}$ from $Z$ is drawn independent and identically distributed (i.i.d) according to a fixed, but unknown distribution $P_{\tilde{z}}$. According to the characteristic (ii) the multi-class classifier can choose to not classify an example. Many reasons may justify adding to $Y$ new predicted classes like "Unclassified" "hesitation between classes $y_1$ and $y_2$", etc. In practice, it's more prudent to not classify an example than to give him a wrong class (See Bartlett and Wegkamp (2008) for an interesting classification model with rejection option). Let C a set of *predicted* classes such that $\mathrm{Y} \subseteq \mathrm{C}$. In our multi-class setting, the classification task consists in assigning to each input object $x$ a predicted class $c$. The accuracy of classifier $h : \mathrm{X} \to \mathrm{C}$ is measured

31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.

through a *loss function*. In the binary case, this is usually a zero-one loss function. In our multi-class context, however, different types of errors may deserve different error costs according to their relative gravity.

We thus posit a more general *valued, cardinal, normalized,* loss function $\mathrm{Q} : \mathrm{C} \times \mathrm{Y} \rightarrow [0, 1]$, with $\mathrm{Q}(y, y) = 0 \forall y \in \mathrm{Y}$ and $Max_{c,y}\mathrm{Q}(c, y) = 1$. The real risk of the classifier $h$ is defined as $R_h = E_{\tilde{z}} Q(h(\tilde{x}), y)$ and empirical risk $\tilde{R}_h = \frac{1}{n}\sum_{j=1}^{n} Q(h(x_j), y_j)$. Now consider the ordered set of distinct $s$ values that may be incurred of the valued loss function Q(error costs) noted $0 = q_1 < q_2 < ... < q_s = 1$ such that: $\mathrm{Q}_{c,y} = q_i \forall (c, y) \in \mathrm{C} \times \mathrm{Y}, 1 \leq i \leq s$. In the sequel, "multi-class" is taken to mean that the error cost takes on at least one fractional value, reflecting intensity or relative gravity of errors – so that $s > 2$. Consider the random vector $\tilde{K} = (\tilde{K}_1, ..., \tilde{K}_s)$ , where $\tilde{K}_i$ is the number of examples falling into error cost category $i$, $1 \leq i \leq s$. Let $K = \{k \in Z_+^s \mid \sum_{i=1}^{s} k_i = n\}$ denote the range of $\tilde{K}$, the empirical risk is defined by:

$$\tilde{R}_h = \frac{1}{n} q^T \tilde{K}(p(h)) = \sum_{i=1}^{s} q_i k_i \tag{1}$$

We remark that $\tilde{K} = (\tilde{K}_1, ..., \tilde{K}_s)$ has a multinomial distribution with probabilities $p_i(h), 1 \leq i \leq s$. As a consequence, the true risk can equivalently be expressed as:

$$R_h = q^T p(h) = \sum_{i=1}^{s} q_i p_i(h) \tag{2}$$

According to the previous remark, the probability that the classifier $h$ makes exactly $k_i$ errors on the loss category $i$ for $0 \leq s \leq 1$ is:

$$Pr_{\tilde{z}}\{\tilde{K}(p(h)) = k \mid p(h)\} = \frac{n!}{\prod_{i=1}^{s} k_i!} \prod_{i=1}^{s} p_i(h)^{k_i} \tag{3}$$

In the following section we generalize PAC-Bayes theorem in this multi-class setting with a simplified proof. A convex mathematical program is formulated in order to estimate the tightest PAC-Bayes multi-class bound.

## 2 PAC-Bayes Generalization bound for Multi-class learning

The powerful PAC-Bayes theorem provides a tight upper bound on the risk of a stochastic classifier called the Gibbs classifier $G$. Given an input example $x$, the label $G_{\mathfrak{Q}}(x)$ assigned to $x$ by the stochastic Gibbs classifier $G_{\mathfrak{Q}}$ is defined by the following process. We first choose randomly a deterministic classifier $h$ from a family of classifiers $H$ according to the posterior distribution $\mathfrak{Q}$ and then use $h$ to assign the label to $x$. The risk of $G_{\mathfrak{Q}}$ is defined as the expected risk of classifier drawn according to $\mathfrak{Q}$ :

$$R(G_{\mathfrak{Q}}) = E_{h \sim \mathfrak{Q}} R(h)$$

And the empirical risk:

$$\tilde{R}(G_{\mathfrak{Q}}) = E_{h \sim \mathfrak{Q}} \tilde{R}(h)$$

Remember that in our multi-class setting, for each multi-class classifier $h$ the real risk is $R_h = q^T p(h)$ and the empirical risk is $\tilde{R}_h = \frac{1}{n} q^T \tilde{K}(p(h))$ with $\tilde{K}_i(p(h))$ is the number of examples falling, with probability $p_i(h)$, into loss category $i$, and $q_i$ the unit error cost of category $i$, $1 \leq i \leq s$. Let define for each loss category $i$, $1 \leq i \leq s$:

$$\bar{p}_i(\mathfrak{Q}) = E_{h \sim \mathfrak{Q}} p_i(h) \text{ and}$$

$$\tilde{\kappa}_i(\mathfrak{Q}) = \tfrac{1}{n} E_{h \sim \mathfrak{Q}} \tilde{K}_i(p(h)) \text{ and}$$

$\tilde{\kappa}_i(\mathfrak{Q})$ is a random variable, with $E_{\tilde{z}} \tilde{\kappa}_i(\mathfrak{Q}) = \bar{p}(\mathfrak{Q})$

In the multi-class setting, the Gibbs real risk can be written:

$$R(\mathfrak{Q}) = R(G_{\mathfrak{Q}}) = q^T \bar{p}(\mathfrak{Q}) \quad (4)$$

And the Gibbs empirical risk:

$$\tilde{R}(\mathfrak{Q}) = \tilde{R}(G_{\mathfrak{Q}}) = q^T \tilde{\kappa}(\mathfrak{Q}) \quad (5)$$

The Kullback-Leibler divergence (KL) measures the relative entropy between probability measures (see Thomas and Cover , 1991). In the multi-class setting, we have a *meta-Bernoulli* instead of *Bernoulli alea*. Hence we can define the KL divergence between two meta-Bernoulli aleas with parameters $(a, s)$ and $(b, s)$ as:

$$kl(b \| a) = \sum_{i=1}^{s} b_i ln \frac{b_i}{a_i} \quad (6)$$

We define for each classifier $h$ from a set of multi-classifiers $H$, and for each $k \in K = \{k \in Z_+^s \mid \sum_{i=1}^{s} k_i = n\}$:

$$B(k,h) = Pr_{\tilde{z}}\{\tilde{K}(p(h)) = k \mid p(h)\} = \frac{n!}{\prod_{i=1}^{s} k_i!} \prod_{i=1}^{s} p_i(h)^{k_i}, (k \in K, h \in H) \quad (7)$$

the probability that the classifier $h$ makes exactly $k_i$ errors on the loss category $i$ for $1 \leq i \leq s$.

**Lemma 1:** *For each prior distribution* $\mathfrak{B}$ *on* $H$, *for each* $\delta \in [0, 1]$ we have:

$$Pr_{\tilde{z}}(E_{h\sim\mathfrak{B}} \tfrac{1}{B(k,h)} \leq \tfrac{1}{\delta}(\tfrac{(n+s-1)!}{(s-1)!n!})) \geq 1 - \delta$$

***Proof*:**Remember that $K = \{k \in Z_+^s \mid \sum_{i=1}^{s} k_i = n\}$, we have:

$$\begin{aligned} E_{\tilde{z}} \tfrac{1}{B(k,h)} &= \textstyle\sum_{k\in K} Pr_{\tilde{z}}(\tilde{K} = k) \times E_{\tilde{z}|\tilde{K}=k} \tfrac{1}{B(k,h)} \\ &= \textstyle\sum_{k\in K} Pr_{\tilde{z}}(\tilde{K} = k) \times \tfrac{1}{Pr_{\tilde{z}}(\tilde{K}=k)} \\ &= |K| \\ &= \tfrac{(n+s-1)!}{(s-1)!n!} \end{aligned}$$

So, for each distribution $\mathfrak{B}$:

$$E_{\tilde{z}} E_{h\sim\mathfrak{B}} \tfrac{1}{B(k,h)} = E_{h\sim\mathfrak{B}} E_{\tilde{z}} \tfrac{1}{B(k,h)} = \tfrac{(n+s-1)!}{(s-1)!n!}$$

We get Lemma 1 by applying Markov inequality.

**Lemma 2:** *For each posterior distribution* $\mathfrak{Q}$, *for each* $k \in K$:

$$E_{h\sim\mathfrak{Q}}(\tfrac{1}{n} ln(\tfrac{1}{B(k,h)})) \geq kl(\tilde{\kappa}(\mathfrak{Q}) \| \bar{p}(\mathfrak{Q}))$$

***Proof*:**By definition:

$$B(k,h) = \tfrac{n!}{\prod_{i=1}^{s} k_i!} \textstyle\prod_{i=1}^{s} p_i(h)^{k_i}$$

By using Stirling approximation we get:

$$ln(B(k,h)) = -nkl(\kappa(h) \| p(h)) + o(n)$$

For two Meta-Bernoulli aleas KL divergence is defined as:

$$kl(\kappa(h) \| p(h)) = \textstyle\sum_{i=1}^{s} \ \kappa_i(h) ln \frac{\kappa_i(h)}{p_i}$$

Hence:

$$\tfrac{1}{n} ln(\tfrac{1}{B(k,h)}) \geq kl(\kappa(h) \| p(h))$$

By applying Jensen inequality we get:

$$E_{h\sim\mathfrak{Q}}\frac{1}{n}ln(\frac{1}{B(k,h)}) \geq kl(\tilde{\kappa}(\mathfrak{Q})\|\bar{p}(\mathfrak{Q}))$$

By applying these two lemmas we get the following PAC-Bayes theorem in the multi-class setting:

**Theorem 1:** *For any set of multi-classes classifiers $H$, for any prior distribution $\mathfrak{B}$, for each $\delta \in [0,1]$*

$$Pr_{\tilde{z}}(\forall\mathfrak{Q} : kl(\tilde{\kappa}(\mathfrak{Q})\|\bar{p}(\mathfrak{Q})) \leq \frac{KL(\mathfrak{Q}\|\mathfrak{B}) + ln(\frac{1}{\delta}(\frac{(n+s-1)!}{(s-1)!n!}))}{n}) \geq 1-\delta \tag{8}$$

***Proof:***

For each prior distribution $\mathfrak{B}$ we have:

$$\begin{aligned} ln[E_{h\sim\mathfrak{B}}(\frac{1}{B(k,h)})] &= ln[E_{h\sim\mathfrak{Q}}\frac{\mathfrak{B}(h)}{\mathfrak{Q}(h)}(\frac{1}{B(k,h)})]\forall\mathfrak{Q} \\ &\geq E_{h\sim\mathfrak{Q}}ln[\frac{\mathfrak{B}(h)}{\mathfrak{Q}(h)}(\frac{1}{B(k,h)})] \text{ by Jensen inequality} \\ &= -KL(\mathfrak{Q}\|\mathfrak{B}) + E_{h\sim\mathfrak{Q}}ln[\frac{1}{B(k,h)}] \end{aligned}$$

By applying Lemma 1 we have:

$$Pr_{\tilde{z}}(\forall\mathfrak{Q}, E_{h\sim\mathfrak{Q}}ln[\frac{1}{B(k,h)}] \leq KL(\mathfrak{Q}\|\mathfrak{B}) + ln(\frac{1}{\delta}(\frac{(n+s-1)!}{(s-1)!n!}))) \geq 1-\delta \tag{9}$$

By applying Lemma 2 we get multi-class PAC-Bayes theorem.

Multi-class PAC-Bayes theorem tells us that the KL divergence between empirical loss and generalization loss of the posterior distribution $\mathfrak{Q}$ is bounded by the KL divergence between $\mathfrak{Q}$ and the prior distribution $\mathfrak{B}$. We try now to determine a bound on the generalization risk of a Gibbs Classifier. For a fixed $\delta$ and $\mathfrak{B}$ let define the function :

$$r(\kappa,\mathfrak{Q}) = Sup_{x\in\Re^s}\{q^Tx|n.kl(\kappa\|x) \leq KL(\mathfrak{Q}\|\mathfrak{B}) + ln(\frac{1}{\delta}(\frac{(n+s-1)!}{(s-1)!n!}))\}$$

It's clear that $\tilde{r}(\mathfrak{Q}) = r(\tilde{\kappa}(\mathfrak{Q}),\mathfrak{Q})$ is a random variable driven from $\tilde{z} = (\tilde{x},\tilde{y}) \in Z$.

**Theorem 2:**

$$Pr_{\tilde{z}}(\forall\mathfrak{Q} : R(\mathfrak{Q}) \leq \tilde{r}(\mathfrak{Q})) \geq 1-\delta$$

***Proof:*** We define:

$$\varepsilon(\mathfrak{Q}) = KL(\mathfrak{Q}\|\mathfrak{B}) + ln(\frac{1}{\delta}(\frac{(n+s-1)!}{(s-1)!n!}))$$

$$A(\mathfrak{Q}) = \{\kappa \in \mathcal{U}_s | n.kl(\kappa\|\bar{p}(\mathfrak{Q})) \leq \varepsilon(\mathfrak{Q})\} \text{ with } U_s = \{x \in \Re_s^+, e^Tx = 1$$

$$B(\mathfrak{Q}) = \{\kappa \in \mathcal{U}_s | R(\mathfrak{Q}) \leq r(\kappa,\mathfrak{Q})\}$$

By definition of $r(\kappa,\mathfrak{Q})$ :

$$\forall(\mathfrak{Q}) : \{(\kappa,\bar{p}(\mathfrak{Q}))|n.kl(\kappa\|\bar{p}(\mathfrak{Q})) \leq \varepsilon(\mathfrak{Q})\} \subseteq \{(\kappa,\bar{p}(\mathfrak{Q}))|\kappa \in \mathcal{U}_s|R(\mathfrak{Q}) \leq r(\kappa,\mathfrak{Q})\}$$

The projection and the intersection preserve the inclusion, then :

$$\forall\mathfrak{Q} : A(\mathfrak{Q}) \subseteq B(\mathfrak{Q}), \text{ and}$$

$$1-\delta \leq Pr_{\tilde{z}}(\tilde{\kappa}(\mathfrak{Q}) \in A(\mathfrak{Q})) \leq Pr_{\tilde{z}}(\tilde{\kappa}(\mathfrak{Q}) \in B(\mathfrak{Q}))$$

## 3 PAC-Bayes Multi-class Bound Estimation

Theorem 2 gives a determinist bound (expected real risk) for each posterior distribution $\mathfrak{Q}$. Remember that $H$ is a finite set of multi-class classifiers, in the following we will prove that for a fixed $\mathfrak{Q}$, estimate PAC-Bayes bound is equivalent to solve a convex mathematical program.

The probability distributions $\mathfrak{B}$ and $\mathfrak{Q}$ can be represented by vectors: $\pi \in \Re_+^{|H|}$ and $\rho \in \Re_+^{|H|}$. We try to estimate PAC-Bayes Multi-class bound for a fixed $\rho$.

Let $\kappa(\rho) = n^{-1}\sum_{h\in H} K(h)$ the observed realization of $\tilde{\kappa}(\rho)$. We define:

$$f(\rho) = \frac{1}{n}\sum_{h\in H}\rho_h ln\frac{\rho_h}{\pi_h} + \frac{1}{n}ln(\frac{1}{\delta}(\frac{(n+s-1)!}{(s-1)!n!})) - \sum_{i=1}^{s}\kappa_i(\rho)ln\sum_{i=1}^{s}\kappa_i(\rho)$$
$$g(p,\rho) = -\sum_{i=1}^{s}\kappa_i(\rho)lnp_i(h)$$

We determine a Gibbs Risk bound $\bar{B}(\rho)$ by looking for probabilities which maximize this risk :

$$\bar{B}(\rho) = Max_p q^T p \tag{10}$$
$$\text{S.t. } g(p,\rho) \geq f(\rho)$$
$$\sum_{i=1}^{s} p_i = 1$$
$$p \geq 0$$

The first constraint of the mathematical program is equivalent to:

$$n.kl(\kappa\|p) \leq KL(\rho\|\pi) + ln(\frac{1}{\delta}(\frac{(n+s-1)!}{(s-1)!n!}))$$

Then $\bar{B}(\rho) = r(\kappa(\rho),\rho)$. The optimal solution $p^*(\rho)$ of the mathematical programm gives for each $\rho$ a pessimist estimation of $\bar{p}(\rho)$ with an uniform confidence level $1-\delta$.

The function $g(.,\rho)$ is strictly positive, differentiable with a negative gradient and strictly convex. So the first constraint has the form (convex function <= constant), furthermore, second and third constraints are linear so the feasible region is convex. The objective function to maximize is concave so the mathematical program (10) is convex.

## References


[1] A. Lacasse, F. Laviolette, M. Marchand, P. Germain, and N. Usunier. PAC-Bayes bounds for the risk of the majority vote and the variance of the Gibbs classifier. In NIPS, pages 769-776, 2006.

[2] A. Maurer. A note on the pac-bayesian theorem. arXiv preprint cs/0411099, 2004.

[3] D. A. McAllester. Some PAC-Bayesian theorems. In Proceedings of COLT, 1998.

[4] D. A. McAllester. Some PAC-Bayesian theorems. Machine Learning, 37(3)355-363, 1999.

[5] D. A. McAllester. PAC-Bayesian stochastic model selection. Machine Learning, 51(1)5-21, 2003.

[6] D. McAllester. A PAC-Bayesian tutorial with a dropout bound. CoRR, abs/1307.2118, 2013.

[7] E. Morvant , S. Koco, L. Ralaivola. PAC-Bayesian Generalization Bound on Confusion Matrix for Multi-Class Classification. In Proceedings of International Conference on Machine Learning, 2012.

[8] F. Laviolette and M. Marchand. PAC-Bayes risk bounds for sample-compressed Gibbs classifiers. In ICML, pages 481-488, 2005.

[9] J. Langford and J. Shawe-Taylor. PAC-Bayes and margins. In Proceedings of Advances in Neural Information Processing Systems, pages 439–446, 2002.

[10] J. Langford, Tutorial on practical prediction theory for classification. Journal of Machine Learning Research, 6, 273-306, 2005.

[11] L. G. Valiant. A theory of the learnable. Commun. ACM, 27(11):1134–1142, 1984.

[12] M. Seeger. PAC-Bayesian generalization bounds for Gaussian processes. Journal of Machine Learning Research, 3:233-269, 2002.

[13] M. Seeger. Bayesian Gaussian Process Models: PAC-Bayesian Generalisation Error Bounds and Sparse Approximations. PhD thesis, University of Edinburgh, 2003.

[14] O. Catoni. PAC-Bayesian supervised classification: the thermodynamics of statistical learning. Institute of Mathematical Statistics Lecture Notes—Monograph Series, 56. Institute of Mathematical Statistics, 2007.

[15] P. Bartlett, C. Michael, D. Mcallester, B. Taskar. Large Margin Methods for Structured Classification: Exponentiated gradient algorithms and PAC-Bayesian generalization bounds, (2004).

[16] P. Germain, A. Lacasse, F. Laviolette, and M. Marchand. PAC-Bayesian learning of linear classifiers. In Proceedings of the 26th International Conference on Machine Learning, pages 353–360, 2009.

[17] P. Germain, A. Lacoste, F. Laviolette, M. Marchand, and S. Shanian. A PAC-Bayes sample-compression approach to kernel methods. In ICML, pages 297-304, 2011.

[18] P. Germain, A. Lacoste, F. Laviolette, M. Marchand and J.F Roy. Risk Bounds for the Majority Vote: From a PAC-Bayesian Analysis to a Learning Algorithm. Journal of Machine Learning Research 16 (2015) 787-860.

[19] T. Cover and J. Thomas. Elements of Information Theory. Series in Telecommunications. John Wiley and Sons, 1st edition, 1991.