PAC-Bayesian Generalization Bound on Confusion Matrix for Multi-Class Classification

Emilie Morvant EMILIE.MORVANT@LIF.UNIV-MRS.FR
Sokol Koço SOKOL.KOCO@LIF.UNIV-MRS.FR
Liva Ralaivola LIVA.RALAIVOLA@LIF.UNIV-MRS.FR
Aix-Marseille Univ., QARMA, LIF, CNRS, UMR 7279, F-13013, Marseille, France

Abstract

In this paper, we propose a PAC-Bayes bound for the generalization risk of the Gibbs classifier in the multi-class classification framework. The novelty of our work is the critical use of the *confusion matrix* of a classifier as an error measure; this puts our contribution in the line of work aiming at dealing with performance measure that are richer than mere scalar criterion such as the misclassification rate. Thanks to very recent and beautiful results on matrix concentration inequalities, we derive two bounds showing that the true confusion risk of the Gibbs classifier is upper-bounded by its empirical risk plus a term depending on the number of training examples in each class. To the best of our knowledge, this is the first PAC-Bayes bounds based on confusion matrices.

1. Introduction

The PAC-Bayesian framework, first introduced in (McAllester, 1999a), is an important field of research in learning theory. It borrows ideas from the philosophy of Bayesian inference and mixes them with techniques used in statistical approaches of learning. Given a family of classifiers $\mathcal{F}$, the ingredients of a PAC-Bayesian bound are a *prior distribution* $\mathfrak{P}$ over $\mathcal{F}$, a learning sample S and a *posterior distribution* $\mathfrak{Q}$ over $\mathcal{F}$. Distribution $\mathfrak{P}$ conveys some prior belief on what are the best classifiers from $\mathcal{F}$ (prior any access to S); the classifiers expected to be the most performant for the classification task at hand therefore have the largest weights under $\mathfrak{P}$. The posterior distribution $\mathfrak{Q}$ is learned/adjusted using the information provided by the training set S. The essence of PAC-Bayesian results is to bound the risk of the *stochastic* Gibbs classifier associated with $\mathfrak{Q}$ (Catoni, 2004).

When specialized to appropriate function spaces $\mathcal{F}$ and relevant families of prior and posterior distributions, PAC-Bayes bounds can be used to characterize the error of different existing classification methods, such as support vector machines (Langford & Shawe-Taylor, 2002; Langford, 2005). PAC-Bayes bounds can also be used to derive new supervised learning algorithms. For example, Lacasse et al. (2007) have introduced an elegant bound on the risk of the majority vote, which holds for any space $\mathcal{F}$. This bound is used to derive an algorithm, `MinCq` (Laviolette et al., 2011), which achieves empirical results on par with state-of-the-art methods. Some other important results are given in (Catoni, 2007; Seegerr, 2002; McAllester, 1999b; Langford et al., 2001).

In this paper, we address the multi-class classification problem. Given the ability of PAC-Bayesian bounds to explain performances of learning methods, related contributions to our work are multi-class formulations for the SVMs, such as those of Weston & Watkins (1998); Lee et al. (2004) and Crammer & Singer (2002). As majority vote methods, we may relate our work to boosting multi-class extensions of AdaBoost (Freund & Schapire, 1996), such as in the framework of Mukherjee & Schapire (2011), and to the algorithms `AdaBoost.MH/AdaBoost.MR` (Schapire & Singer, 1999) and `SAMME` (Zhu et al., 2009).

The originality of our work is that we consider the *confusion matrix* of the Gibbs classifier as an error measure. We believe that in the multi-class framework, it is more relevant to consider the confusion matrix as the error measure than the mere misclassification error, which corresponds to the probability for some classifier h to err for its prediction on $\mathbf{x}$. The information as to

Appearing in *Proceedings of the 29th International Conference on Machine Learning*, Edinburgh, Scotland, UK, 2012.

what is the probability for an instance of class p to be classified into class q (with $p \neq q$) by some predictor is indeed crucial in some applications (think of the difference between false-negative and false-positive predictions in a diagnosis automated system). To the best of our knowledge, we are the first to propose a generalization bound on the confusion matrix in the PAC-Bayesian framework. The result we propose heavily relies on a matrix concentration inequality for sums of random matrices of Tropp (2011).

The rest of this paper is organized as follows. Sec. 2 introduces the setting of multi-class learning and some of the basic notation used throughout the paper. Sec. 3 briefly recalls the folk PAC-Bayes bound as introduced in (McAllester, 2003). In Sec. 4, we present the main contribution of this paper, our PAC-Bayes bound on the confusion matrix, followed by its proof in Sec. 5. We discuss some future works in Sec. 6.

2. Setting and Notation

This section presents the general setting that we consider and the different tools that we will make use of.

2.1. General Problem Setting

We consider classification tasks over some *input space* X. The *output space* is $Y = \{1, \ldots, Q\}$, where Q is the number of classes. The learning sample is denoted by $S = \{(\mathbf{x}_i, y_i)\}_{i=1}^m$ where each example is drawn *i.i.d.* from a fixed (but unknown) probability distribution $\mathfrak{D}$ over $X \times Y$; $\mathfrak{D}_m$ denotes the distribution of an m-sample. We consider the family $\mathcal{F} \subseteq Y^X$ of classifiers, and $\mathfrak{P}$ and $\mathfrak{Q}$ will respectively refer to *prior* and *posterior* distributions over $\mathcal{F}$. Given the prior distribution $\mathfrak{P}$ and the training set S, learning aims at finding a posterior $\mathfrak{Q}$ leading to good generalization.

The Kullback-Leibler divergence $\mathrm{KL}(\mathfrak{Q}||\mathfrak{P})$ between $\mathfrak{Q}$ and $\mathfrak{P}$ is

$$\mathrm{KL}(\mathfrak{Q}||\mathfrak{P}) = \mathbb{E}_{f\sim\mathfrak{Q}} \log \frac{\mathfrak{Q}(f)}{\mathfrak{P}(f))}; \tag{1}$$

$\mathrm{sign}(x) = 1$ if $x \geq 0$ and -1 otherwise; The indicator function $\mathbb{I}(x)$ is equal to 1 if x is true and 0 otherwise.

2.2. Conventions and Basics on Matrices

Throughout the paper we consider only real-valued square matrices $\mathbf{C}$ of order Q (the number of classes). ${}^t\mathbf{C}$ is the transpose of the matrix $\mathbf{C}$, $\mathbf{Id}_Q$ denotes the identity matrix of size Q and $\mathbf{0}$ is the zero matrix.

The results given in this paper are based on a concentration inequality of Tropp (2011) for sums of random self-adjoint matrices, which extends to general real-valued matrices thanks to the dilation technique Paulsen (2002): the dilation $\mathcal{S}(\mathbf{C})$ of matrix $\mathbf{C}$ is

$$\mathcal{S}(\mathbf{C}) \overset{def}{=} \begin{pmatrix} \mathbf{0} & \mathbf{C} \\ {}^t\mathbf{C} & \mathbf{0} \end{pmatrix}. \tag{2}$$

Notation $\|\cdot\|$ refers to the *operator or spectral norm.* It returns the maximum singular value of its argument:

$$\|\mathbf{C}\| = \max\{\lambda_{\max}(\mathbf{C}), -\lambda_{\min}(\mathbf{C})\}, \tag{3}$$

where $\lambda_{\max}$ and $\lambda_{\min}$ are respectively the algebraic maximum and minimum singular value of $\mathbf{C}$. Note that dilation preserves spectral information, so

$$\lambda_{\max}\big(\mathcal{S}(\mathbf{C})\big) = \|\mathcal{S}(\mathbf{C})\| = \|\mathbf{C}\|. \tag{4}$$

3. The Usual PAC-Bayes Theorem

Here, we recall the main PAC-Bayesian bound in the binary classification case as presented in (McAllester, 2003; Seegerr, 2002; Langford, 2005), where the set of labels is $Y = \{-1, 1\}$ The true risk $R(f)$ and the empirical error $R_S(f)$ of f are defined as:

$$R(f) \overset{def}{=} \mathbb{E}_{(\mathbf{x},y)\sim\mathfrak{D}} \mathbb{I}(f(\mathbf{x} \neq y)),$$

$$R_S(f) \overset{def}{=} \frac{1}{m}\sum_{i=1}^{m} \mathbb{I}(f(\mathbf{x}_i \neq y_i)).$$

The learner's aim is to choose a posterior distribution $\mathfrak{Q}$ on $\mathcal{F}$ such that the risk of the $\mathfrak{Q}$-weighted majority vote (also called the Bayes classifier) $B_\mathfrak{Q}$ is as small as possible. $B_\mathfrak{Q}$ makes a prediction according to

$$B_\mathfrak{Q}(\mathbf{x}) = \mathrm{sign}\left[\mathbb{E}_{f\sim\mathfrak{Q}} f(\mathbf{x})\right].$$

The true risk $R(B_\mathfrak{Q})$ and the empirical error $R_S(B_\mathfrak{Q})$ of the Bayes classifier are defined as the probability that it commits an error on an example:

$$R(B_\mathfrak{Q}) \overset{def}{=} \mathbb{P}_{(\mathbf{x},y)\sim\mathfrak{D}}\left(B_\mathfrak{Q}(\mathbf{x}) \neq y\right). \tag{5}$$

However, the PAC-Bayes approach does not directly bound the risk of $B_\mathfrak{Q}$. Instead, it bounds the risk of the stochastic Gibbs classifier $G_\mathfrak{Q}$ which predicts the label of $\mathbf{x} \in X$ by first drawing f according to $\mathfrak{Q}$ and then returning $f(\mathbf{x})$. The true risk $R(G_\mathfrak{Q})$ and the empirical error $R_S(G_\mathfrak{Q})$ of $G_\mathfrak{Q}$ are therefore:

$$R(G_\mathfrak{Q}) = \mathbb{E}_{f\sim\mathfrak{Q}} R(f)\ ;\ R_S(G_\mathfrak{Q}) = \mathbb{E}_{f\sim\mathfrak{Q}} R_S(f). \tag{6}$$

Note that in this setting, we have $R(B_\mathfrak{Q}) \leq 2R(G_\mathfrak{Q})$.

We present the PAC-Bayes theorem which gives a bound on the error of the stochastic Gibbs classifier.

Theorem 1. *For any $\mathfrak{D}$, any $\mathcal{F}$, any $\mathfrak{P}$ of support $\mathcal{F}$, any $\delta \in (0,1]$, we have,*

$$\mathbb{P}_{S\sim\mathfrak{D}_m}\Bigg(\forall \mathfrak{Q} \text{ on } \mathcal{F},\ kl\big(R_S(G_\mathfrak{Q}), R(G_\mathfrak{Q})\big) \leq \frac{1}{m}\left[KL(\mathfrak{Q}\|\mathfrak{P}) + \ln\frac{\xi(m)}{\delta}\right]\Bigg) \geq 1-\delta,$$

where $kl(a,b) \overset{def}{=} a\ln\frac{a}{b} + (1-a)\ln\frac{1-a}{1-b}$,
and $\xi \overset{def}{=} \sum_{i=0}^{m}\binom{m}{i}(i/m)^i(1-i/m)^{m-i}$.

We now provide a novel PAC-Bayes bound in the context of multi-class classification by considering the confusion matrix as an error measure.

4. Multiclass PAC-Bayes Bound

4.1. Definitions and Setting

As said earlier, we focus on multi-class classification. The output space is $Y=\{1,\ldots,Q\}$, with $Q>2$. We only consider learning algorithms acting on learning sample $S=\{(\mathbf{x}_i,y_i)\}_{i=1}^m$ where each example is drawn *i.i.d* according to $\mathfrak{D}$, such that $|S|\geq Q$ and $m_{y_j}\geq 1$ for every class $y_j\in Y$, where m_{y_j} is the number of examples of real class y_j. In the context of multi-class classification, an error measure can be a performance tool called *confusion matrix*. We consider the classical definition of the confusion matrix based on conditional probabilities: it is inherent (and desirable) to minimize the effects of unbalanced classes. Concretely, for a given classifier $f\in\mathcal{F}$ and a sample $S=\{(\mathbf{x}_i,y_i)\}_{i=1}^m \sim \mathfrak{D}_m$, the *empirical confusion matrix* $\mathbf{D}_S^f=(\hat{d}_{pq})_{1\leq p,q\leq Q}$ of f is defined as follows:

$$\forall(p,q),\ \hat{d}_{pq} \overset{def}{=} \sum_{i=1}^{m}\frac{1}{m_{y_i}}\mathbb{I}(f(\mathbf{x}_i)=q)\mathbb{I}(y_i=p).$$

The *true confusion matrix* $\mathbf{D}^f = (d_{pq})_{1\leq p,q\leq Q}$ of f over $\mathfrak{D}$ corresponds to:

$$\begin{aligned}\forall(p,q),\ d_{pq} &\overset{def}{=} \mathbb{E}_{\mathbf{x}|y=p}\mathbb{I}\big(f(\mathbf{x})=q\big)\\ &= \mathbb{P}_{(\mathbf{x},y)\sim\mathfrak{D}}(f(\mathbf{x})=q|y=p).\end{aligned}$$

If f correctly classifies every example of the sample S, then all the elements of the confusion matrix are 0, except for the diagonal ones which correspond to the correctly classified examples. Hence the more there are non-zero elements in a confusion matrix outside the diagonal, the more the classifier is prone to err. Recall that in a learning process the objective is to learn a classifier $f\in\mathcal{F}$ with a low true error (*i.e.* with good generalization guarantees), we are thus only interested in the errors of f. Our objective is then to find f leading to a confusion matrix with the more zero elements outside the diagonal. Since the diagonal gives the conditional probabilities of 'correct' predictions, we propose to consider a different kind of confusion matrix by discarding the diagonal values. Then the only non-zero elements of the new confusion matrix correspond to the examples that are misclassified by f. For all $f\in\mathcal{F}$ we define the empirical and true confusion matrices of f by respectively $\mathbf{C}_S^f=(\hat{c}_{pq})_{1\leq p,q\leq Q}$ and $\mathbf{C}^f=(c_{pq})_{1\leq p,q\leq Q}$ such that for all (p,q):

$$\hat{c}_{pq} \overset{def}{=} \begin{cases}0 & \text{if } q=p\\ \hat{d}_{pq} & \text{otherwise,}\end{cases} \tag{7}$$

$$c_{pq} \overset{def}{=} \begin{cases}0 & \text{if } q=p\\ d_{pq}=\mathbb{P}_{(\mathbf{x},y)\sim\mathfrak{D}}(f(\mathbf{x})=q|y=p) & \text{otherwise.}\end{cases} \tag{8}$$

Note that if f correctly classifies every example of a given sample S, then the empirical confusion matrix $\mathbf{C}_S^f$ is equal to $\mathbf{0}$. Similarly, if f is a perfect classifier over the distribution $\mathfrak{D}$, then the true confusion matrix is equal to $\mathbf{0}$. Therefore a relevant task is to minimize the size of the confusion matrix, thus having a confusion matrix as close to $\mathbf{0}$ as possible.

4.2. Main Result: Confusion PAC-Bayes Bound for the Gibbs Classifier

Our main result is a PAC-Bayes generalization bound over the Gibbs classifier $G_\mathfrak{Q}$ in this particular context, where the empirical and true error measures are respectively given by the confusion matrices from (7) and (8). In this case, we can define the true and the empirical confusion matrices of $G_\mathfrak{Q}$ respectively by:

$$\mathbf{C}^{G_\mathfrak{Q}} = \mathbb{E}_{f\sim\mathfrak{Q}}\mathbb{E}_{S\sim\mathfrak{D}_m}\mathbf{C}_S^f\ ;\ \mathbf{C}_S^{G_\mathfrak{Q}} = \mathbb{E}_{f\sim\mathfrak{Q}}\mathbf{C}_S^f.$$

Given $f\sim\mathfrak{Q}$ and a sample $S\sim\mathfrak{D}_m$, our objective is to bound the difference between $\mathbf{C}^{G_\mathfrak{Q}}$ and $\mathbf{C}_S^{G_\mathfrak{Q}}$, the true and empirical errors of the Gibbs classifier. Remark that the error rate $P(f(\mathbf{x})\neq y)$ of a classifier f might be directly computed as the 1-norm of ${}^t\mathbf{C}^f\mathbf{p}$, where $\mathbf{p}$ is the vector of prior probabilities. However, in our case, concentration inequalities are only available for the operator norm. Since we have $\|\mathbf{u}\|_1\leq\sqrt{Q}\|\mathbf{u}\|_2$ for any Q-dimensional vector $\mathbf{u}$, we have that $P(f(\mathbf{x})\neq y)\leq\sqrt{Q}\|\mathbf{C}^f\|_{op}$. Thus trying to minimize the operator norm of $\mathbf{C}^f$ is a relevant strategy to control the risk. Here is our main result, a bound on the operator norm of the difference between $\mathbf{C}^{G_\mathfrak{Q}}$ and $\mathbf{C}_S^{G_\mathfrak{Q}}$.

Theorem 2. *Let $X\subseteq\mathbb{R}^d$ be the input space, $Y=\{1,\ldots,Q\}$ the output space, $\mathfrak{D}$ a distribution over $X\times Y$ (with $\mathfrak{D}_m$ the distribution of a m-sample) and $\mathcal{F}$ a* *family of classifiers from X to Y. Then for every prior distribution $\mathfrak{P}$ over $\mathcal{F}$ and any $\delta \in (0,1]$, we have:*

$$\mathbb{P}_{S\sim\mathfrak{D}_m}\Bigg\{\forall\mathfrak{Q} \text{ on } \mathcal{F}, \|\mathbf{C}_S^{G_\mathfrak{Q}} - \mathbf{C}^{G_\mathfrak{Q}}\| \le \sqrt{\frac{8Q}{m_- - 8Q}\left[KL(\mathfrak{Q}\|\mathfrak{P}) + \ln\left(\frac{m_-}{4\delta}\right)\right]}\Bigg\} \ge 1-\delta,$$

where $m_- = \min_{y=1,\dots,Q} m_y$ is the minimal number of examples from S which belong to the same class.

Proof. Deferred to Section 5. □

Note that, for all $y \in Y$, we need the following hypothesis: $m_y > 8$, which is not too strong a limitation.

Finally, we rewrite Theorem 2 in order to provide a bound on the size $\|\mathbf{C}^{G_\mathfrak{Q}}\|$.

Corollary 1. *We consider the hypothesis of the Theorem 2. We have:*

$$\mathbb{P}_{S\sim\mathfrak{D}_m}\Bigg\{\forall\mathfrak{Q} \text{ on } \mathcal{F},\ \|\mathbf{C}^{G_\mathfrak{Q}}\| \le \|\mathbf{C}_S^{G_\mathfrak{Q}}\| + \sqrt{\frac{8Q}{m_- - 8Q}\left[KL(\mathfrak{Q}\|\mathfrak{P}) + \ln\left(\frac{m_-}{4\delta}\right)\right]}\Bigg\} \ge 1-\delta.$$

Proof. By application of the reverse triangle inequality $|\|\mathbf{A}\| - \|\mathbf{B}\|| \le \|\mathbf{A} - \mathbf{B}\|$ to Theorem 2. □

Both Theorem 2 and Corollary 1 yield a bound on the estimation (through the operator norm) of the true confusion matrix of the Gibbs classifier over the posterior distribution $\mathfrak{Q}$, though this is more explicit in the corollary. Let the number of classes Q be a constant, then the true risk is upper-bounded by the empirical risk of the Gibbs classifier and a term depending on the number of training examples, especially on the value m_- which corresponds to the minimal quantity of examples that belong to the same class. This means that the larger m_-, the closer the empirical confusion matrix of the Gibbs classifier is to its true matrix.

4.3. Upper Bound on the Risk of the Majority Vote Classifier

We recall that the Bayes classifier $B_\mathfrak{Q}$ is well known as majority vote classifier under a given posterior distribution $\mathfrak{Q}$. In the multiclass setting, $B_\mathfrak{Q}$ is such that for any example it returns the majority class under the measure $\mathfrak{Q}$ and we define it as:

$$B_\mathfrak{Q}(\mathbf{x}) = \operatorname{argmax}_{c\in Y}\left[\mathbb{E}_{f\in\mathfrak{Q}}\mathbb{I}(f(\mathbf{x}) = c)\right]. \quad (9)$$

We define the conditional Gibbs risk $R(G_\mathfrak{Q}, p, q)$ and Bayes risk $R(G_\mathfrak{Q}, p, q)$ as

$$R(G_\mathfrak{Q}, p, q) = \mathbb{E}_{\mathbf{x}\sim D_{|y=p}}\mathbb{E}_{f\sim\mathfrak{Q}}\mathbb{I}(f(\mathbf{x}) = q), \quad (10)$$
$$R(B_\mathfrak{Q}, p, q) = \mathbb{E}_{\mathbf{x}\sim D_{|y=p}}\mathbb{I}\left(\operatorname{argmax}_{c\in Y} g(c,q)\right). \quad (11)$$

where

$$g(c,q) = \left[\mathbb{E}_{f\in\mathfrak{Q}}\mathbb{I}(f(\mathbf{x}) = c) = q\right]$$

The former is the (p,q) entry of $\mathbf{C}^{G_\mathfrak{Q}}$ (if $p \neq q$) and the latter is the (p,q) entry of $\mathbf{C}^{B_\mathfrak{Q}}$.

Proposition 1. *Let $Q \ge 2$ be the number of classes. Then $R(B_\mathfrak{Q}, p, q)$ and $R(G_\mathfrak{Q}, p, q)$ are related by the following inequality :*

$$\forall(q,p), R(B_\mathfrak{Q}, p, q) \le QR(G_\mathfrak{Q}, p, q). \quad (12)$$

Proof. Given in (Morvant et al., 2012). □

This proposition implies the following result on the confusion matrices associated to $B_\mathfrak{Q}$ and $G_\mathfrak{Q}$.

Corollary 2. *Let $Q \ge 2$ be the number of class. Then $\mathbf{C}^{B_\mathfrak{Q}}$ and $\mathbf{C}^{G_\mathfrak{Q}}$ are related by the following inequality:*

$$\|\mathbf{C}^{B_\mathfrak{Q}}\| \le Q\|\mathbf{C}^{G_\mathfrak{Q}}\|. \quad (13)$$

Proof. Given in (Morvant et al., 2012). □

5. Proof of Theorem 2

This section gives the formal proof of Theorem 2. We first introduce a concentration inequality for a sum of random square matrices. This allows us to deduce the PAC-Bayes generalization bound for confusion matrices by following the same "three step process" as the one given in (McAllester, 2003; Seegerr, 2002; Langford, 2005) for the classic PAC-Bayesian bound.

5.1. Concentration Inequality for the Confusion Matrix

The main result of our work is based on the following corollary of a result on the concentration inequality for a sum of self-adjoint matrices given by Tropp (2011) (see Theorem 3 in Appendix) – this theorem generalizes Hoeffding's inequality to the case self-adjoint random matrices. The purpose of the following corollary is to restate Theorem 3 so that it carries over to matrices that are not self-adjoint. It is central to us to have such a result as the matrices we are dealing with, namely confusion matrices, are rarely symmetric.

Corollary 3. *Consider a finite sequence $\{\mathbf{M}_i\}$ of independent, random, square matrices of order Q, and let $\{a_i\}$ be a sequence of fixed scalars. Assume that* *each random matrix satisfies* $\mathbb{E}_i\mathbf{M}_i = \mathbf{0}$ *and* $\|\mathbf{M}_i\| \leq a_i$ *almost surely.. Then, with* $\sigma^2 \overset{def}{=} \sum_i a_i^2$*, we have,*

$$\forall \epsilon \geq 0,\ \mathbb{P}\left\{\|\sum_i \mathbf{M}_i\| \geq \epsilon\right\} \leq 2Q\exp\left(\frac{-\epsilon^2}{8\sigma^2}\right). \quad (14)$$

Proof. We want to verify the hypothesis given in Theorem 3 in order to apply it.

Let $\{\mathbf{M}_i\}$ be a finite sequence of independent, random, square matrices of order Q such that $\mathbb{E}_i\mathbf{M}_i = \mathbf{0}$ and let $\{a_i\}$ be a sequence of fixed scalars such that $\|\mathbf{M}_i\| \leq a_i$. We consider the sequence $\{\mathcal{S}(\mathbf{M}_i)\}$ of random self-adjoint matrices with dimension $2Q$. By the definition of the dilation, we obtain $\mathbb{E}_i\mathcal{S}(\mathbf{M}_i) = \mathbf{0}$.

From Equation (4), the dilation preserves the spectral information. Thus, on the one hand, we have:

$$\|\sum_i \mathbf{M}_i\| = \lambda_{\max}\left(\mathcal{S}\left(\sum_i \mathbf{M}_i\right)\right) = \lambda_{\max}\left(\sum_i \mathcal{S}(\mathbf{M}_i)\right).$$

On the other hand, we have:

$$\|\mathbf{M}_i\| = \|\mathcal{S}(\mathbf{M}_i)\| = \lambda_{\max}\big(\mathcal{S}(\mathbf{M}_i)\big) \leq a_i.$$

To assure the hypothesis $\mathcal{S}(\mathbf{M}_i)^2 \preccurlyeq \mathbf{A}_i^2$, we need to find a suitable sequence of fixed self-adjoint matrices $\{\mathbf{A}_i\}$ of dimension $2Q$ (where $\preccurlyeq$ refers to the semidefinite order on self-adjoint matrices). Indeed, it suffices to construct a diagonal matrix defined as $\lambda_{\max}\big(\mathcal{S}(\mathbf{M}_i)\big)\mathbf{Id}_{2Q}$ for ensuring $\mathcal{S}(\mathbf{M}_i)^2 \preccurlyeq \big(\lambda_{\max}\big(\mathcal{S}(\mathbf{M}_i)\big)\mathbf{Id}_{2Q}\big)^2$. More precisely, since for every i we have $\lambda_{\max}\big(\mathcal{S}(\mathbf{M}_i)\big) \leq a_i$, we fix $\mathbf{A}_i$ as a diagonal matrix with a_i on the diagonal, *i.e.* $\mathbf{A}_i \overset{def}{=} a_i\mathbf{Id}_{2Q}$, with $\|\sum_i \mathbf{A}_i^2\| = \sum_i a_i^2 = \sigma^2$. Finally, we can invoke Theorem 3 to obtain the concentration inequality (14). □

In order to make use of this corollary, we rewrite the confusion matrices as sums of example-based confusion matrices. That is, for each example $(\mathbf{x}_i, y_i) \in S$, we define its empirical confusion matrix by $\mathbf{C}_i^f = (\hat{c}_{pq}(i))_{1\leq p,q\leq Q}$ as follows:

$$\forall p, q, \hat{c}_{pq}(i) \overset{def}{=} \begin{cases} 0 & \text{if } q = p \\ \dfrac{1}{m_{y_i}}\mathbb{I}(f(\mathbf{x}) = q)\mathbb{I}(y_i = p) & \text{otherwise.} \end{cases}$$

where m_{y_i} is the number of examples of class $y_i \in Y$ belonging to S. Given an example $(\mathbf{x}_i, y_i) \in S$, the example-based confusion matrix contains at most one non zero-element when f misclassifies $(\mathbf{x}_i, y_i)$. In the same way, when f correctly classifies $(\mathbf{x}_i, y_i)$ then the example-based confusion matrix is equal to $\mathbf{0}$. Our error measure is then $\mathbf{C}_S^f = \sum_{i=1}^m \mathbf{C}_i^f$, that is we penalize only when f errs.

We further introduce the random square matrices ${\mathbf{C}'}_i^f$:

$${\mathbf{C}'}_i^f = \mathbf{C}_i^f - \mathbb{E}_{S\sim\mathfrak{D}_m}\mathbf{C}_i^f, \quad (15)$$

which verifies $\mathbb{E}_i{\mathbf{C}'}_i^f = 0$.

We have yet to find a suitable a_i for a given ${\mathbf{C}'}_i^f$. Let $\lambda_{\max_i}$ be the maximum singular value of ${\mathbf{C}'}_i^f$. It is easy to verify that $\lambda_{\max_i} \leq \frac{1}{m_{y_i}}$. Thus, for all i we fix a_i equal to $\frac{1}{m_{y_i}}$.

Finally, with the introduced notations, Corollary 3 leads to the following concentration inequality:

$$\mathbb{P}\left\{\|\sum_{i=1}^m {\mathbf{C}'}_i^f\| \geq \epsilon\right\} \leq 2Q\exp\left(\frac{-\epsilon^2}{8\sigma^2}\right). \quad (16)$$

This inequality (16) allows us to demonstrate our Theorem 2 by following the process of (McAllester, 2003; Seegerr, 2002; Langford, 2005).

5.2. "Three Step Proof" Of Our Bound

First, thanks to the concentration inequality (16), we prove the following lemma.

Lemma 1. *Let Q be the size of $\mathbf{C}_S^f$ and ${\mathbf{C}'}_i^f = \mathbf{C}_i^f - \mathbb{E}_{S\sim\mathfrak{D}_m}\mathbf{C}_i^f$ defined as in (15). Then the following bound holds for any $\delta \in (0, 1]$:*

$$\mathbb{P}_{S\sim\mathfrak{D}_m}\left\{\mathbb{E}_{f\sim\mathfrak{P}}\left[\exp\left(\frac{1-8\sigma^2}{8\sigma^2}\|\sum_{i=1}^m {\mathbf{C}'}_i^f\|^2\right)\right] \leq \frac{2Q}{8\sigma^2\delta}\right\} \geq 1-\delta$$

Proof. For readability reasons, we note ${\mathbf{C}'}_S^f = \sum_{i=1}^m {\mathbf{C}'}_i^f$. If Z is a real valued random variable so that $\mathbb{P}\left(Z \geq z\right) \leq k\exp(-n.g(z))$ with $g(z)$ non-negative, non-decreasing and k a constant, then $\mathbb{P}\left(\exp\left((n-1)g(Z)\right) \geq \nu\right) \leq \min(1, k\nu^{-n/(n-1)})$. We apply this to the concentration inequality (16). Choosing $g(z) = z^2$ (non-negative), $z = \epsilon$, $n = \frac{1}{8\sigma^2}$ and $k = 2Q$, we obtain the following result:

$$\mathbb{P}\left\{\exp\left(\frac{1-8\sigma^2}{8\sigma^2}\|{\mathbf{C}'}_S^f\|\right) \geq \nu\right\} \leq \min\left(1, 2Q\nu^{-1/(1-8\sigma^2)}\right).$$

Note that $\exp\left(\frac{1-8\sigma^2}{8\sigma^2}\|{\mathbf{C}'}_S^f\|\right)$ is always non-negative.

Hence it allows us to compute its expectation as:

$$
\begin{aligned}
&\mathbb{E}\left[\exp\left(\frac{1-8\sigma^2}{8\sigma^2}\|\mathbf{C'}^f_S\|\right)\right]\\
&=\int_0^\infty \mathbb{P}\left\{\exp\left(\frac{1-8\sigma^2}{(8\sigma^2)}\|\mathbf{C'}^f_S\|\right)\geq\nu\right\}d\nu\\
&\leq 2Q+\int_1^\infty 2Q\nu^{-1/(1-8\sigma^2)}d\nu\\
&=2Q-2Q\frac{1-8\sigma^2}{8\sigma^2}\left[\nu^{-8\sigma^2/(1-8\sigma^2)}\right]_1^\infty\\
&=\frac{2Q}{8\sigma^2}.
\end{aligned}
$$

For a given classifier $f\in\mathcal{F}$, we have:

$$\mathbb{E}_{S\sim\mathfrak{D}^m}\left[\exp\left(\frac{1-8\sigma^2}{8\sigma^2}\|\mathbf{C'}^f_S\|\right)\right]\leq\frac{2Q}{8\sigma^2} \tag{17}$$

Then, if $\mathfrak{P}$ is a probability distribution over $\mathcal{F}$, Equation (17) implies that:

$$\mathbb{E}_{S\sim\mathfrak{D}^m}\left[\mathbb{E}_{f\sim\mathfrak{P}}\exp\left(\frac{1-8\sigma^2}{8\sigma^2}\|\mathbf{C'}^f_S\|\right)\right]\leq\frac{2Q}{8\sigma^2} \tag{18}$$

Using Markov's inequality[1], we obtain the result of the lemma. □

The second step to prove Theorem 2 is to use the shift given in (McAllester, 2003). We recall this result in the following lemma.

Lemma 2 (Donsker-Varadhan inequality Donsker & Varadhan (1975))**.** *Given the Kullback-Leibler divergence*[2] $KL(\mathfrak{Q}\|\mathfrak{P})$ *between two distributions* $\mathfrak{P}$ *and* $\mathfrak{Q}$ *and let* $g(\cdot)$ *be a function, we have:*

$$\mathbb{E}_{b\sim\mathfrak{Q}}\Big[g(b)\Big]\leq KL(\mathfrak{Q}\|\mathfrak{P})+\ln\mathbb{E}_{b\sim\mathfrak{Q}}\Big[\exp(g(b))\Big].$$

Proof. See (McAllester, 2003). □

Recall that $\mathbf{C'}^f_S=\sum_{i=1}^m\mathbf{C'}^f_i$. With $g(b)=\frac{1-8\sigma^2}{8\sigma^2}b^2$ and $b=\|\mathbf{C'}^f_S\|$, Lemma 2 implies:

$$
\begin{aligned}
&\mathbb{E}_{f\sim\mathfrak{Q}}\left[\frac{1-8\sigma^2}{8\sigma^2}\|\mathbf{C'}^f_S\|^2\right]\\
&\leq KL(\mathfrak{Q}\|\mathfrak{P})+\ln\mathbb{E}_{f\sim\mathfrak{P}}\left[\exp\left(\frac{1-8\sigma^2}{8\sigma^2}\|\mathbf{C'}^f_S\|^2\right)\right].
\end{aligned}
\tag{19}
$$

The last step that completes the proof of Theorem 2 consists in applying the result we obtained in Lemma 1 to Equation (19). Then, we have:

$$\mathbb{E}_{f\sim\mathfrak{Q}}\left[\frac{1-8\sigma^2}{8\sigma^2}\|\mathbf{C'}^f_S\|^2\right]\leq KL(\mathfrak{Q}\|\mathfrak{P})+\ln\frac{2Q}{8\sigma^2\delta}. \tag{20}$$

Since $g(\cdot)$ is clearly convex, we apply Jensen's inequality[3] to (20). Then, with probability at least $1-\delta$ over S, and for every distribution $\mathfrak{Q}$ on $\mathcal{F}$, we have:

$$\Big(\mathbb{E}_{f\sim\mathfrak{Q}}\|\mathbf{C'}^f_S\|\Big)^2\leq\frac{8\sigma^2}{1-8\sigma^2}\Big(KL(\mathfrak{Q}\|\mathfrak{P})+\ln\frac{2Q}{8\sigma^2\delta}\Big). \tag{21}$$

Since $\mathbf{C'}^f_S=\sum_{i=1}^m\Big[\mathbf{C}^f_i-\mathbb{E}_{S\sim\mathfrak{D}_m}\mathbf{C}^f_i\Big]$, then the bound (21) is quite similar to the one given in Theorem 2.

We present in the next section, the calculations leading to our PAC-Bayesian generalization bound.

5.3. Simplification

We first compute the variance parameter $\sigma^2=\sum_{i=1}^m a_i^2$. For that purpose, in Section 5.1 we showed that for each $i\in\{1,\dots,m\}$, we can choose $a_i=\frac{1}{m_{y_i}}$, where y_i is the class of the i-th example and m_{y_i} is the number of examples of class y_i. Thus we have:

$$\sigma^2=\sum_{i=1}^m\frac{1}{m_{y_i}^2}=\sum_{y=1}^Q\sum_{i:y_i=y}\frac{1}{m_y^2}=\sum_{y=1}^Q\frac{1}{m_y}.$$

For sake of simplification of Equation (21) and since the term on the right side of this equation is an increasing function with respect to σ^2, we propose to upper-bound σ^2:

$$\sigma^2=\sum_{y=1}^Q\frac{1}{m_y}\leq\frac{Q}{\min_{y=1,\dots,Q}m_y}. \tag{22}$$

Let $m_-\overset{def}{=}\min_{y=1,\dots,Q}m_y$, then using Equation (22), we obtain the following bound from Equation (21):

$$\Big(\mathbb{E}_{f\sim\mathfrak{Q}}[\|\mathbf{C'}^f_S\|]\Big)^2\leq\frac{8Q}{m_--8Q}\Big(KL(\mathfrak{Q}\|\mathfrak{P})+\ln\frac{m_-}{4\delta}\Big).$$

Then:

$$\mathbb{E}_{f\sim\mathfrak{Q}}[\|\mathbf{C'}^f_S\|]\leq\sqrt{\frac{8Q}{m_--8Q}\Big(KL(\mathfrak{Q}\|\mathfrak{P})+\ln\frac{m_-}{4\delta}\Big)}. \tag{23}$$

It remains to replace $\mathbf{C'}^f_S=\sum_{i=1}^m\Big[\mathbf{C}^f_i-\mathbb{E}_{S\sim\mathfrak{D}_m}\mathbf{C}^f_i\Big]$. Recall that $\mathbf{C}^{G_\mathfrak{Q}}=\mathbb{E}_{f\sim\mathfrak{Q}}\mathbb{E}_{S\sim\mathfrak{D}_m}\mathbf{C}^f_S$ and $\mathbf{C}^{G_\mathfrak{Q}}_S=$

[1] see Theorem 4 in Appendix.

[2] The KL-divergence is defined in Equation (1).

[3] see Theorem 5 in Appendix.

$\mathbb{E}_{f\sim\mathfrak{Q}}\mathbf{C}_S^f$, we obtain:

$$\begin{aligned}
\mathbb{E}_{f\sim\mathfrak{Q}}[\|\mathbf{C'}_S^f\|] &= \mathbb{E}_{f\sim\mathfrak{Q}}\left[\|\sum_{i=1}^{m}\left[\mathbf{C}_i^f - \mathbb{E}_{S\sim\mathfrak{D}_m}\mathbf{C}_i^f\right]\|\right]\\
&= \mathbb{E}_{f\sim\mathfrak{Q}}\left[\|\sum_{i=1}^{m}\left[\mathbf{C}_i^f\right] - \sum_{i=1}^{m}\left[\mathbb{E}_{S\sim\mathfrak{D}_m}\mathbf{C}_i^f\right]\|\right]\\
&= \mathbb{E}_{f\sim\mathfrak{Q}}\left[\|\mathbf{C}_S^f - \mathbb{E}_{S\sim\mathfrak{D}_m}\mathbf{C}_S^f\|\right]\\
&\geq \|\mathbb{E}_{f\sim\mathfrak{Q}}\left[\mathbf{C}_S^f - \mathbb{E}_{S\sim\mathfrak{D}_m}\mathbf{C}_S^f\right]\|\\
&= \|\mathbb{E}_{f\sim\mathfrak{Q}}\mathbf{C}_S^f - \mathbb{E}_{f\sim\mathfrak{Q}}\mathbb{E}_{S\sim\mathfrak{D}_m}\mathbf{C}_S^f\|\\
&= \|\mathbf{C}_S^{G_\mathfrak{Q}} - \mathbf{C}^{G_\mathfrak{Q}}\|. \qquad (24)
\end{aligned}$$

By substituting the left part of the inequality (23) with the term (24), we find the bound of our Theorem 2.

6. Discussion and Future Work

This work gives rise to many interesting questions, among which the following ones.

Some future works will be focused on instantiating our bound given in Theorem 2 for specific multi-class frameworks, such as multi-class SVM and multi-class boosting. Taking advantage of our theorem while using the confusion matrices, may allow us to derive new generalization bounds for these methods.

Additionally, we are interested in seeing how effective learning methods may be derived from the risk bound we propose. For instance, in the binary PAC-Bayes setting, the algorithm `MinCq` proposed by Laviolette et al. (2011) minimizes a bound depending on the first two moments of the margin of the $\mathcal{Q}$-weighted majority vote. From our Theorem 2 and with a similar study, we would like to design a new multi-class algorithm and observe how sound such an algorithm could be. This would probably require the derivation of a Cantelli-Tchebycheff deviation inequality in the matrix case.

Besides, it might be very interesting to see how the noncommutative/matrix concentration inequalities provided by Tropp (2011) might be of some use for other kinds of learning problem such as multi-label classification, label ranking problems or structured prediction issues.

Finally, the question of extending the present work to the analysis of algorithms learning (possibly infinite-dimensional) operators as proposed by Abernethy et al. (2009) is also very exciting.

7. Conclusion

In this paper, we propose a new PAC-Bayesian generalization bound that applies in the multi-class classification setting. The originality of our contribution is that we consider the confusion matrix as an error measure. Coupled with the use of the operator norm on matrices, we are capable of providing generalization bound on the 'size' of confusion matrix (with the idea that the smaller the norm of the confusion matrix of the learned classifier, the better it is for the classification task at hand). The derivation of our result takes advantage of the concentration inequality proposed by Tropp (2011) for the sum of random self-adjoint matrices, that we directly adapt to square matrices which are not self-adjoint.

The main results are presented in Theorem 2 and Corollary 1. The bound in Theorem 2 is given on the difference between the true risk of the Gibbs classifier and its empirical error. While the one given in Corollary 1 upper-bounds the risk of the Gibbs classifier by its empirical error.

An interesting point is that our bound depends on the minimal quantity m_- of training examples belonging to the same class, for a given number of classes. If this value increases, *i.e.* if we have a lot of training examples, then the empirical confusion matrix of the Gibbs classifier tends to be close to its true confusion matrix. A point worth noting is that the bound varies as $O(1/\sqrt{m_-})$, which is a typical rate in bounds not using second-order information.

Last but not least, the present work has given rise to a few algorithmic and theoretical questions that we have discussed in the previous section.

Appendix

Theorem 3 (Concentration Inequality for Random Matrices Tropp (2011))**.** *Consider a finite sequence* $\{\mathbf{M}_i\}$ *of independant, random, self-adjoint matrices with dimension* Q*, and let* $\{\mathbf{A}_i\}$ *be a sequence of fixed self-adjoint matrices. Assume that each random matrix satisfies* $\mathbb{E}\mathbf{M}_i = \mathbf{0}$ *and* $\mathbf{M}_i^2 \preccurlyeq \mathbf{A}_i^2$ *almost surely. Then, for all* $\epsilon \geq 0$*,*

$$\mathbb{P}\left\{\lambda_{\max}\left(\sum_i \mathbf{M}_i\right) \geq \epsilon\right\} \leq Q\exp\left(-\frac{\epsilon^2}{8\sigma^2}\right),$$

where $\sigma^2 \overset{def}{=} \|\sum_i \mathbf{A}_i^2\|$ *and* $\preccurlyeq$ *refers to the semidefinite order on self-adjoint matrices.*

Theorem 4 (Markov's inequality)**.** *Let* Z *be a random* *variable and $z \geq 0$, then:*

$$\mathbb{P}\left(|Z| \geq z\right) \leq \frac{\mathbb{E}(|Z|)}{z}.$$

Theorem 5 (Jensen's inequality)**.** *Let X be an integrable real-valued random variable and $g(\cdot)$ be a convex function, then:*

$$f(\mathbb{E}[Z]) \leq \mathbb{E}[g(Z)].$$

References

Abernethy, Jacob, Bach, Francis, Evgeniou, Theodoros, and Vert, Jean-Philippe. A new approach to collaborative filtering: Operator estimation with spectral regularization. *Journal of Machine Learning Research*, 10:803–826, 2009.

Catoni, Olivier. 4. gibbs estimators. In *Statistical Learning Theory and Stochastic Optimization*, volume 1851, pp. 111–135. Springer, 2004.

Catoni, Olivier. PAC-bayesian supervised classification: The thermodynamics of statistical learning. *ArXiv e-prints*, 2007.

Crammer, Koby and Singer, Yoram. On the algorithmic implementation of multiclass kernel-based vector machines. *Journal of Machine Learning Research*, 2:265–292, 2002.

Donsker, David and Varadhan, S. S. Asymptotic evaluation of certain markov process expectations for large time. *Communications on Pure and Applied Mathematics*, 28, 1975.

Freund, Yoav and Schapire, Robert E. Experiments with a new boosting algorithm. In *In Proceedings of the International Conference on Machine Learning*, pp. 148–156, 1996.

Lacasse, Alexandre, Laviolette, François, Marchand, Mario, Germain, Pascal, and Usunier, Nicolas. PAC-bayes bounds for the risk of the majority vote and the variance of the Gibbs classifier. In *Adv. in Neural Processing Systems (NIPS)*, 2007.

Langford, John. Tutorial on practical prediction theory for classification. *Journal of Machine Learning Research*, 6:273–306, 2005.

Langford, John and Shawe-Taylor, John. PAC-bayes & margins. In *Advances in Neural Information Processing Systems 15*, pp. 439–446. MIT Press, 2002.

Langford, John, Seeger, Matthias, and Megiddo, Nimrod. An improved predictive accuracy bound for averaging classifiers. In *Proc. of the International Conference on Machine Learning*, pp. 290–297, 2001.

Laviolette, François, Marchand, Mario, and Roy, Jean-Francis. From PAC-Bayes Bounds to Quadratic Programs for Majority Votes. In *Proc. of the International Conference on Machine Learning*, June 2011.

Lee, Y., Lin, Y., and Wahba, G. Multicategory support vector machines, theory, and application to the classification of microarray data and satellite radiance data. *Journal of the American Statistical Association*, 99:67–81, 2004.

McAllester, David A. Some PAC-bayesian theorems. *Machine Learning*, 37:355–363, 1999a.

McAllester, David A. PAC-bayesian model averaging. In *Proceedings of the annual conference on Computational learning theory (COLT)*, pp. 164–170, 1999b.

McAllester, David A. Simplified PAC-bayesian margin bounds. In *Proc. of the annual conference on Computational learning theory (COLT)*, pp. 203–215, 2003.

Morvant, Emilie, Koço, Sokol, and Ralaivola, Liva. PAC-Bayesian Generalization Bound on Confusion Matrix for Multi-Class Classification. Arxiv: http://arxiv.org/abs/1202.6228, 2012.

Mukherjee, Indraneel and Schapire, Robert E. A theory of multiclass boosting. *CoRR*, abs/1108.2989, 2011.

Paulsen, V.I. *Completely bounded maps and operator algebras*. Cambridge studies in advanced mathematics. Cambridge University Press, 2002.

Schapire, Robert E. and Singer, Yoram. Improved boosting algorithms using confidence-rated predictions. In *Machine Learning*, pp. 80–91, 1999.

Seegerr, Matthias. PAC-bayesian generalization error bounds for gaussian process classification. *Journal of Machine Learning Research*, 3:233–269, 2002.

Tropp, Joel A. User-friendly tail bounds for sums of random matrices. *Foundations of Computational Mathematics*, pp. 1–46, 2011.

Weston, Jason and Watkins, Chris. Multi-class support vector machines, 1998.

Zhu, Ji, Zou, Hui, Rosset, Saharon, and Hastie, Trevor. Multi-class adaboost, 2009.