PAC-Bayesian Generalization Bound on Confusion Matrix for Multi-Class Classification
Emilie Morvant, Sokol Koço, Liva Ralaivola

To cite this version:

HAL Id: hal-00674847
https://hal.archives-ouvertes.fr/hal-00674847v4
Submitted on 30 Jun 2012 (v4), last revised 22 Oct 2013 (v6)

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
PAC-Bayesian Generalization Bound on Confusion Matrix for Multi-Class Classification

Emilie Morvant
Sokol Koço
Liva Ralaivola
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 $\mathbb{P}$ over $\mathcal{F}$, a learning sample $S$ and a posterior distribution $\mathbb{Q}$ over $\mathcal{F}$. Distribution $\mathbb{P}$ 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 $\mathbb{P}$. The posterior distribution $\mathbb{Q}$ 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 $\mathbb{Q}$ (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...
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 $S = \{(x_i, y_i)\}_{i=1}^m$, where each example is drawn i.i.d. from a fixed (but unknown) probability distribution $\mathcal{D}$ over $X \times Y$. $\mathcal{D}_m$ 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 $\text{KL}(\mathfrak{Q}||\mathfrak{P})$ between $\mathfrak{Q}$ and $\mathfrak{P}$ is

$$\text{KL}(\mathfrak{Q}||\mathfrak{P}) = \mathbb{E}_{f \sim \mathfrak{Q}} \log \frac{\mathfrak{Q}(f)}{\mathfrak{P}(f)}.$$  

The sign function $\text{sign}(x) = 1$ if $x \geq 0$ and $-1$ otherwise; The indicator function $\mathbb{1}(x)$ 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). $\mathbf{C}^\top$ is the transpose of the matrix $\mathbf{C}$, $\mathbf{I}_Q$ 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

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

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

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

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

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

3. The Usual PAC-Bayes Theorem

Here, we recall the main PAC-Bayesian bound in the binary classification case as presented in (McAllester, 2003; Seeger, 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:

$$R(f) \overset{\text{def}}{=} \mathbb{E}_{(x, y) \sim \mathcal{D}} \mathbb{1}(f(x) \neq y),$$

$$R_S(f) \overset{\text{def}}{=} \frac{1}{m} \sum_{i=1}^m \mathbb{1}(f(x_i) \neq y_i).$$

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

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

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:

$$R(B_{\mathfrak{Q}}) \overset{\text{def}}{=} \mathbb{P}_{(x, y) \sim \mathcal{D}} (B_{\mathfrak{Q}}(x) \neq y). \quad (5)$$

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 $x \in X$ by first drawing $f$ according to $\mathfrak{Q}$ and then returning $f(x)$. The true risk $R(G_{\mathfrak{Q}})$ and the empirical error $R_S(G_{\mathfrak{Q}})$ of $G_{\mathfrak{Q}}$ are therefore:

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

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 \( \mathcal{D} \), any \( \mathcal{F} \), any \( \mathcal{Q} \) of support \( \mathcal{F} \), any \( \delta \in (0, 1] \), we have,

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

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

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 \( S = \{(x_i, y_i)\}_{i=1}^{m} \) where each example is drawn i.i.d according to \( \mathcal{D} \), 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 \( S = \{(x_i, y_i)\}_{i=1}^{m} \sim \mathcal{D}_m \), the empirical confusion matrix \( \mathbf{D}^f_S = (\hat{d}_{pq})_{1 \leq p, q \leq Q} \) of \( f \) is defined as follows:

\[
\forall (p, q), \hat{d}_{pq} = \frac{1}{m_{y_q}} \mathbb{I}(f(x_i) = q) \mathbb{I}(y_i = p).
\]

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

\[
\forall (p, q), d_{pq} = \mathbb{P}_{x \mid y = p}(f(x) = q) = \mathbb{P}_{(x, y) \sim \mathcal{D}}(f(x) = q | y = p).
\]

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 \( \mathbf{C}^f = (c_{pq})_{1 \leq p, q \leq Q} \) and \( \mathbf{C}^f = (c_{pq})_{1 \leq p, q \leq Q} \) such that for all \( (p, q) \):

\[
\hat{c}_{pq} = \begin{cases} 0 & \text{if } q = p \\ \hat{d}_{pq} & \text{otherwise}, \end{cases} \quad \text{and} \quad c_{pq} = \begin{cases} 0 & \text{if } q = p \\ \hat{d}_{pq} - \mathbb{P}_{(x, y) \sim \mathcal{D}}(f(x) = q | y = p) & \text{otherwise}. \end{cases}
\]

Note that if \( f \) correctly classifies every example of a given sample \( S \), then the empirical confusion matrix \( \mathbf{C}^f_S \) is equal to \( \mathbf{0} \). Similarly, if \( f \) is a perfect classifier over the distribution \( \mathcal{D} \), 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 \( G_{\mathcal{D}} \) 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 \( G_{\mathcal{D}} \) respectively by:

\[
\mathbf{C}^{G_{\mathcal{D}}} = \mathbb{E}_{f \sim \mathcal{D}} \mathbb{E}_{S \sim \mathcal{D}_m} \mathbf{C}^f_S; \quad \mathbf{C}^{G_{\mathcal{D}}} = \mathbb{E}_{f \sim \mathcal{D}} \mathbf{C}^f_S.
\]

Given \( f \sim \mathcal{Q} \) and a sample \( S \sim \mathcal{D}_m \), our objective is to bound the difference between \( \mathbf{C}^{G_{\mathcal{D}}} \) and \( \mathbf{C}^{G_{\mathcal{D}}} \), the true and empirical errors of the Gibbs classifier. Remark that the error rate \( P(f(x) \neq y) \) of a classifier \( f \) might be directly computed as the 1-norm of \( \| \mathbf{C}^{f} p \|_1 \), where \( p \) 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 \( P(f(x) \neq y) \leq \sqrt{Q} \| \mathbf{C}^{f} \|_o \). 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 \( \mathbf{C}^{G_{\mathcal{D}}} \) and \( \mathbf{C}^{G_{\mathcal{D}}} \).

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

$$P_{S \sim \mathcal{D}_m} \left\{ \forall \Omega \text{ on } \mathcal{F}, \|C^G_{\Omega} - C^G_{\Omega} \| \leq \sqrt{\frac{8Q}{m_- - 8Q} \left[ KL(\Omega \| \mathcal{P}) + \ln \left( \frac{m_-}{4\delta} \right) \right]} \right\} \geq 1 - \delta,$$

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

**Proof.** Deferred to Section 5. \(\Box\)

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 $\|C^G_{\Omega}\|$. **Corollary 1.** We consider the hypothesis of the Theorem 2. We have:

$$P_{S \sim \mathcal{D}_m} \left\{ \forall \Omega \text{ on } \mathcal{F}, \|C^G_{\Omega}\| \leq \|C^G_{\mathcal{S}}\| + \sqrt{\frac{8Q}{m_- - 8Q} \left[ KL(\Omega \| \mathcal{P}) + \ln \left( \frac{m_-}{4\delta} \right) \right]} \right\} \geq 1 - \delta.$$

**Proof.** By application of the reverse triangle inequality $\|A\| - \|B\| \leq \|A - B\|$ to Theorem 2. \(\Box\)

Both Theorem 2 and Corollary 1 yield a bound on the empirical risk of the Gibbs classifier over the posterior distribution $\Omega$, 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 $B_{\Omega}$ is well known as majority vote classifier under a given posterior distribution $\Omega$. In the multiclass setting, $B_{\Omega}$ is such that for any example it returns the majority class under the measure $\Omega$ and we define it as:

$$B_{\Omega}(x) = \argmax_{c \in Y} \left[ E_{f \sim \Omega}(f(x) = c) \right].$$

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

$$R(G_{\Omega}, p, q) = E_{x \sim D_{\Omega}} E_{f \sim \Omega}(f(x) = q),$$

$$R(B_{\Omega}, p, q) = E_{x \sim D_{\Omega}} I(\argmax_{c \in Y} g(c, q)).$$

where $g(c, q) = \left[ E_{f \sim \Omega}(f(x) = c) \right] = q$

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

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

$$\forall (q, p), R(B_{\Omega}, p, q) \leq Q R(G_{\Omega}, p, q).$$

**Proof.** Given in (Morvant et al., 2012). \(\Box\)

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

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

$$\|C_{B_{\Omega}}\| \leq Q \|C^G_{\Omega}\|.$$

**Proof.** Given in (Morvant et al., 2012). \(\Box\)

### 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; Seeger, 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 \( \{M_i\} \) of independent, random, square matrices of order \( Q \), and let \( \{a_i\} \) be a sequence of fixed scalars. Assume that each random matrix satisfies \( \mathbb{E}_i M_i = 0 \) and \( \|M_i\| \leq a_i \), almost surely. Then, with \( \sigma^2 \triangleq \sum_i a_i^2 \), we have,

\[
\forall \epsilon \geq 0, \quad P \left\{ \left\| \sum_i M_i \right\| \geq \epsilon \right\} \leq 2Q \exp \left( -\frac{\epsilon^2}{8\sigma^2} \right). \tag{14}
\]

**Proof.** We want to verify the hypothesis given in Theorem 3 in order to apply it. Let \( \{M_i\} \) be a finite sequence of independent, random, square matrices of order \( Q \) such that \( \mathbb{E}_i M_i = 0 \) and \( \{a_i\} \) be a sequence of fixed scalars such that \( \|M_i\| \leq a_i \). We consider the sequence \( \{S(M_i)\} \) of random self-adjoint matrices with dimension \( 2Q \). By the definition of the dilation, we obtain \( \mathbb{E}_i S(M_i) = 0 \).

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

\[
\| \sum_i M_i \| = \lambda_{\max} \left( \sum_i M_i \right) = \lambda_{\max} \left( \sum_i S(M_i) \right).
\]

On the other hand, we have:

\[
\|M_i\| = \|S(M_i)\| = \lambda_{\max}(S(M_i)) \leq a_i.
\]

To assure the hypothesis \( S(M_i)^2 \preceq A_i^2 \), we need to find a suitable sequence of fixed self-adjoint matrices \( \{A_i\} \) of dimension \( 2Q \) (where \( \preceq \) refers to the semidefinite order on self-adjoint matrices). Indeed, it suffices to construct a diagonal matrix defined as \( \lambda_{\max}(S(M_i)) \text{Id}_{2Q} \) for ensuring \( S(M_i)^2 \preceq \lambda_{\max}(S(M_i)) \text{Id}_{2Q} \).

More precisely, since for every \( i \) we have \( \lambda_{\max}(S(M_i)) \leq a_i \), we fix \( A_i \) as a diagonal matrix with \( a_i \) on the diagonal, i.e., \( A_i \triangleq a_i \text{Id}_{2Q} \), with \( \| \sum_i A_i^2 \| = \sum_i a_i^2 = \sigma^2 \). Finally, we can invoke Theorem 3 to obtain the concentration inequality (14). \( \Box \)

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 \( (x_i, y_i) \in S \), we define its empirical confusion matrix by \( C_i^f = (\tilde{c}_{pq}(i))_{1 \leq p, q \leq Q} \) as follows:

\[
\forall p, q, \tilde{c}_{pq}(i) \triangleq \begin{cases} 
0 & \text{if } q = p \\
\frac{1}{m_{y_i}} \mathbb{I}(f(x) = q) \mathbb{I}(y_i = p) & \text{otherwise}.
\end{cases}
\]

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

We further introduce the random square matrices \( C_i^{f'} \):

\[
C_i^{f'} = C_i^f - \mathbb{E}_{S \sim \mathcal{D}_m} C_i^f,
\]

which verifies \( \mathbb{E}_i C_i^{f'} = 0 \).

We have yet to find a suitable \( a_i \) for a given \( C_i^{f'} \). Let \( \lambda_{\max, i} \) be the maximum singular value of \( C_i^{f'} \). 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:

\[
P \left\{ \left\| \sum_i C_i^{f'} \right\| \geq \epsilon \right\} \leq 2Q \exp \left( -\frac{\epsilon^2}{8\sigma^2} \right). \tag{16}
\]

This inequality (16) allows us to demonstrate our Theorem 2 by following the process of (McAllester, 2003; Seeger, 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 \( C_S^f \) and \( C_i^{f'} = C_i^f - \mathbb{E}_{S \sim \mathcal{D}_m} C_i^f \) defined as in (15). Then the following bound holds for any \( \delta \in (0, 1) \):

\[
P_{S \sim \mathcal{D}_m}\left\{ \mathbb{E}_f \sup_{\|C\| \leq 1} \left[ \exp \left( \frac{1 - 8\sigma^2}{8\sigma^2} \|C\| \right) \right] \leq \frac{2Q}{\sigma^2 \delta} \right\} \geq 1 - \delta
\]

**Proof.** For readability reasons, we note \( C_S^{f'} = \sum_{i=1}^m C_i^{f'} \). If \( Z \) is a real valued random variable so that \( P(Z \geq z) \leq k \exp(-(n,g(z)) \) with \( g(z) \) non-negative, non-decreasing and \( k \) a constant, then \( P(\exp((n-1)g(Z)) \geq \nu) \leq \min(1, k\nu^{-1/(n-1)}) \). 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:

\[
P \left\{ \exp \left( \frac{1 - 8\sigma^2}{8\sigma^2} \|C_S^{f'}\| \right) \geq \nu \right\} \leq \min\left(1, 2Q \nu^{-1/(1-8\sigma^2)}\right).
\]

Note that \( \exp \left( \frac{1 - 8\sigma^2}{8\sigma^2} \|C_S^{f'}\| \right) \) is always non-negative.
Hence it allows us to compute its expectation as:

\[
\begin{align*}
\mathbb{E}\left[ \exp\left( \frac{1 - 8\sigma^2}{8\sigma^2} \| C_S^f \| \right) \right] \\
= \int_0^\infty \mathbb{P}\left\{ \exp\left( \frac{1 - 8\sigma^2}{8\sigma^2} \| C_S^f \| \right) \geq \nu \right\} d\nu \\
\leq 2Q + \int_1^\infty 2Q \nu^{1 - 1/(1 - 8\sigma^2)} d\nu \\
= 2Q - 2Q \frac{1 - 8\sigma^2}{8\sigma^2} \left[ \nu^{1 - 1/(1 - 8\sigma^2)} \right]_1^\infty \\
= \frac{2Q}{8\sigma^2}.
\end{align*}
\]

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

\[
\mathbb{E}_{S \sim \mathcal{D}} \left[ \exp\left( \frac{1 - 8\sigma^2}{8\sigma^2} \| C_S^f \| \right) \right] \leq \frac{2Q}{8\sigma^2} \tag{17}
\]

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

\[
\mathbb{E}_{S \sim \mathcal{D}} \mathbb{E}_{f \sim \mathcal{P}} \left[ \exp\left( \frac{1 - 8\sigma^2}{8\sigma^2} \| C_S^f \| \right) \right] \leq \frac{2Q}{8\sigma^2} \tag{18}
\]

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 \cite{DonskerVaradhan75}). Given the Kullback-Leibler divergence\(^2\) \( KL(\Omega||\mathcal{P}) \) between two distributions \( \mathcal{P} \) and \( \Omega \) and let \( g(\cdot) \) be a function, we have:

\[
\mathbb{E}_{b \sim \Omega} \left[ g(b) \right] \leq KL(\Omega||\mathcal{P}) + \ln \mathbb{E}_{b \sim \mathcal{P}} \left[ \exp(g(b)) \right].
\]

**Proof.** See (McAllester, 2003). \( \square \)

Recall that \( C_S^f = \sum_{i=1}^m C_{i}^f \). With \( g(b) = \frac{1 - 8\sigma^2}{8\sigma^2} b^2 \) and \( b = \| C_S^f \| \), Lemma 2 implies:

\[
\mathbb{E}_{f \sim \Omega} \left[ \frac{1 - 8\sigma^2}{8\sigma^2} \| C_S^f \|^2 \right] \leq KL(\Omega||\mathcal{P}) + \ln \mathbb{E}_{f \sim \mathcal{P}} \left[ \exp\left( \frac{1 - 8\sigma^2}{8\sigma^2} \| C_S^f \|^2 \right) \right]. \tag{19}
\]

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:

\[
\mathbb{E}_{f \sim \Omega} \left[ \frac{1 - 8\sigma^2}{8\sigma^2} \| C_S^f \|^2 \right] \leq KL(\Omega||\mathcal{P}) + \ln \frac{2Q}{8\sigma^2}. \tag{20}
\]

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 \( \Omega \) on \( \mathcal{F} \), we have:

\[
\mathbb{E}_{f \sim \Omega} \left[ \frac{1 - 8\sigma^2}{8\sigma^2} \| C_S^f \|^2 \right] \leq \frac{8\sigma^2}{1 - 8\sigma^2} \left( KL(\Omega||\mathcal{P}) + \ln \frac{2Q}{8\sigma^2} \right). \tag{21}
\]

Since \( C_S^f = \sum_{i=1}^m C_{i}^f - ES_{\sim \mathcal{D}} C_{i}^f \), 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 \( \sigma^2 = \sum_{i=1}^m a_i \). For that purpose, in Section 5.1 we showed that for each \( i \in \{1, \ldots, m\} \), we can choose \( a_i = \frac{1}{m_y} \), where \( y_i \) is the class of the \( i \)-th example and \( m_y \) is the number of examples of class \( y \). Thus we have:

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

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 \( \sigma^2 \), we propose to upper-bound \( \sigma^2 \):

\[
\sigma^2 = \sum_{y=1}^Q \frac{1}{m_{y}} \leq \frac{Q}{\min_{y=1, \ldots, Q} m_{y}}. \tag{22}
\]

Let \( m_{\cdot \cdot} = \min_{y=1, \ldots, Q} m_{y} \), then using Equation (22), we obtain the following bound from Equation (21):

\[
\mathbb{E}_{f \sim \Omega} \left[ \| C_S^f \|^2 \right] \leq \frac{8Q}{m_{\cdot \cdot} - 8Q} \left( KL(\Omega||\mathcal{P}) + \ln \frac{m_{\cdot \cdot}}{4d} \right).
\]

Then:

\[
\mathbb{E}_{f \sim \Omega} \left[ \| C_S^f \|^2 \right] \leq \sqrt{\frac{8Q}{m_{\cdot \cdot} - 8Q} \left( KL(\Omega||\mathcal{P}) + \ln \frac{m_{\cdot \cdot}}{4d} \right)}. \tag{23}
\]

It remains to replace \( C_S^f = \sum_{i=1}^m C_{i}^f - ES_{\sim \mathcal{D}} C_{i}^f \). Recall that \( C_{G\cdot} = ES_{\sim \mathcal{D}} C_{S}^f \) and \( C_{G\cdot}^f = ES_{\sim \mathcal{D}} C_{f}^f \), where

\( ^1\)see Theorem 4 in Appendix.

\( ^2\)The KL-divergence is defined in Equation (1).

\( ^3\)see Theorem 5 in Appendix.
PAC-Bayesian Generalization Bounds on the Confusion Matrix

\[ E_{f \sim \Omega} C^f_S, \text{ we obtain:} \]

\[
E_{f \sim \Omega} \| C^f_S \| = E_{f \sim \Omega} \left[ \| \sum_{i=1}^{m} [C^f_i - E_{S \sim \mathcal{D}_m} C^f_i] \| \right] \\
= E_{f \sim \Omega} \left[ \| \sum_{i=1}^{m} C^f_i - \sum_{i=1}^{m} E_{S \sim \mathcal{D}_m} C^f_i \| \right] \\
\geq \| E_{f \sim \Omega} \left[ C^f_S - E_{S \sim \mathcal{D}_m} C^f_S \right] \| \\
= \| E_{f \sim \Omega} C^f_S - E_{f \sim \Omega} E_{S \sim \mathcal{D}_m} C^f_S \| \\
= \| C^f_G - C^f_Q \|. \quad (24)
\]

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 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 \( m_\rightarrow \) 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 \( O(1/\sqrt{m_\rightarrow}) \), 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 \( \{M_i\} \) of independent, random, self-adjoint matrices with dimension \( Q \), and let \( \{A_i\} \) be a sequence of fixed self-adjoint matrices. Assume that each random matrix satisfies \( EM_i = 0 \) and \( M_i^2 \preceq A_i^2 \) almost surely. Then, for all \( \epsilon \geq 0 \),

\[
P \left\{ \lambda_{\max} \left( \sum_i M_i \right) \geq \epsilon \right\} \leq Q \exp \left( -\frac{\epsilon^2}{8\sigma^2} \right),
\]

where \( \sigma^2 \overset{\text{def}}{=} \| \sum_i A_i^2 \| \) and \( \preceq \) 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:

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

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

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

**References**


