



HAL
open science

Confusion Matrix Stability Bounds for Multiclass Classification

Pierre Machart, Liva Ralaivola

► **To cite this version:**

Pierre Machart, Liva Ralaivola. Confusion Matrix Stability Bounds for Multiclass Classification. 2012. hal-00674779v2

HAL Id: hal-00674779

<https://hal.science/hal-00674779v2>

Preprint submitted on 24 May 2012

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.

Confusion Matrix Stability Bounds for Multiclass Classification

Pierre Machart and Liva Ralaivola

QARMA, LIF UMR CNRS 7279

Aix-Marseille Université

39, rue F-Joliot Curie, F-13013 Marseille, France

{pierre.machart, liva.ralaivola}@lif.univ-mrs.fr

Abstract. We provide new theoretical results on the generalization properties of learning algorithms for multiclass classification problems. The originality of our work is that we propose to use the *confusion matrix* of a classifier as a measure of its quality; our contribution is in the line of work which attempts to set up and study the statistical properties of new evaluation measures such as, e.g. ROC curves. In the confusion-based learning framework we propose, we claim that a targetted objective is to minimize the size of the confusion matrix \mathcal{C} , measured through its *operator norm* $\|\mathcal{C}\|$. We derive generalization bounds on the (size of the) confusion matrix in an extended framework of uniform stability, adapted to the case of matrix valued loss. Pivotal to our study is a very recent matrix concentration inequality that generalizes McDiarmid’s inequality. As an illustration of the relevance of our theoretical results, we show how two SVM learning procedures can be proved to be confusion-friendly. To the best of our knowledge, the present paper is the first that focuses on the confusion matrix from a theoretical point of view.

1 Introduction

Multiclass classification is an important problem of machine learning. The issue of having at hand statistically relevant procedures to learn reliable predictors is of particular interest, given the need of such predictors in information retrieval, web mining, bioinformatics or neuroscience (one may for example think of document categorization, gene classification, fMRI image classification).

Yet, the literature on multiclass learning is not as voluminous as that of binary classification, while this multiclass prediction raises questions from the algorithmic, theoretical and practical points of view. One of the prominent questions is that of the measure to use in order to assess the quality of a multiclass predictor. Here, we develop our results with the idea that the *confusion matrix* is a performance measure that deserves to be studied as it provides a finer information on the properties of a classifier than the mere misclassification rate. We do want to emphasize that we provide theoretical results on the confusion matrix itself and that misclassification rate *is not* our primary concern — as we shall see, though, getting bounds on the confusion matrix entails, as a byproduct, bounds on the misclassification rate.

Building on matrix-based concentration inequalities [1–5], also referred to as non-commutative concentration inequalities, we establish a stability framework for confusion-based learning algorithm. In particular, we prove a generalization bound for *confusion stable* learning algorithms and show that there exist such algorithms in the literature. In a sense, our framework and our results extend those of [6], which are designed for scalar loss functions. To the best of our knowledge, this is the first work that establishes generalization bounds based on confusion matrices.

The paper is organized as follows. Section 2 describes the setting we are interested in and motivates the use of the confusion matrix as a performance measure. Section 3 introduces the new notion of *stability* that will prove essential to our study; the main theorem of this paper, together with its proof, are provided. Section 4 is devoted to the analysis of two *SVM* procedures in the light of our new framework. A discussion on the merits and possible extensions of our approach concludes the paper (Section 5).

2 Confusion Loss

2.1 Notation

As said earlier, we focus on the problem of multiclass classification. The input space is denoted by \mathcal{X} and the target space is

$$\mathcal{Y} = \{1, \dots, Q\}.$$

The training sequence

$$\mathbf{Z} = \{Z_i = (X_i, Y_i)\}_{i=1}^m$$

is made of m identically and independently random pairs $Z_i = (X_i, Y_i)$ distributed according to some unknown (but fixed) distribution D over $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$. The sequence of input data will be referred to as $\mathbf{X} = \{X_i\}_{i=1}^m$ and the sequence of corresponding labels $\mathbf{Y} = \{Y_i\}_{i=1}^m$, we may write $\mathbf{Z} = \{\mathbf{X}, \mathbf{Y}\}$. The realization of $Z_i = (X_i, Y_i)$ is $z_i = (x_i, y_i)$ and \mathbf{z} , \mathbf{x} and \mathbf{y} refer to the realizations of the corresponding sequences of random variables. For a sequence $\mathbf{y} = \{y_1, \dots, y_m\}$ of m labels, $m_q(\mathbf{y})$, or simply m_q when clear from context, denotes the number of labels from \mathbf{y} that are equal to q ; $\mathbf{s}(\mathbf{y})$ is the binary sequence $\{s_1(\mathbf{y}), \dots, s_Q(\mathbf{y})\}$ of size Q such that $s_q(\mathbf{y}) = 1$ if $q \in \mathbf{y}$ and $s_q(\mathbf{y}) = 0$ otherwise.

We will use $D_{X|y}$ for the conditional distribution of X given that $Y = y$; therefore, for a given sequence $\mathbf{y} = \{y_1, \dots, y_m\} \in \mathcal{Y}^m$, $D_{\mathbf{X}|\mathbf{y}} = \otimes_{i=1}^m D_{X|y_i}$ is the distribution of the random sample $\mathbf{X} = \{X_1, \dots, X_m\}$ over \mathcal{X}^m such that X_i is distributed according to $D_{X|y_i}$; for $q \in \mathcal{Y}$, and \mathbf{X} distributed according to $D_{\mathbf{X}|\mathbf{y}}$, $\mathbf{X}_q = \{X_{i_1}, \dots, X_{i_{m_q}}\}$ denotes the random sequence of variables such that X_{i_k} is distributed according to $D_{X|q}$. $\mathbb{E}[\cdot]$ and $\mathbb{E}_{X|y}[\cdot]$ denote the expectations with respect to D and $D_{X|y}$, respectively.

For a training sequence \mathbf{Z} , \mathbf{Z}^i denotes the sequence

$$\mathbf{Z}^i = \{Z_1, \dots, Z_{i-1}, Z'_i, Z_{i+1}, \dots, Z_m\}$$

where Z'_i is distributed as Z_i ; $\mathbf{Z}^{\setminus i}$ is the sequence

$$\mathbf{Z}^{\setminus i} = \{Z_1, \dots, Z_{i-1}, Z_{i+1}, \dots, Z_m\}.$$

These definitions directly carry over when conditioned on a sequence of labels \mathbf{y} (with, henceforth, $y'_i = y_i$).

We will consider a family \mathcal{H} of predictors such that

$$\mathcal{H} \subseteq \{h : h(x) \in \mathbb{R}^Q, \forall x \in \mathcal{X}\}.$$

For $h \in \mathcal{H}$, $h_q \in \mathbb{R}^{\mathcal{X}}$ denotes its q th coordinate. Also,

$$\ell = (\ell_q)_{1 \leq q \leq Q}$$

is a set of loss functions such that:

$$\ell_q : \mathcal{H} \times \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}_+.$$

Finally, for a given algorithm $\mathcal{A} : \cup_{m=1}^{\infty} \mathcal{Z}^m \rightarrow \mathcal{H}$, $\mathcal{A}_{\mathbf{Z}}$ will denote the hypothesis learned by \mathcal{A} when trained on \mathbf{Z} .

2.2 Confusion Matrix versus Misclassification Rate

We here provide a discussion as to why minding the *confusion matrix* or *confusion loss* (terms that we will use interchangeably) is crucial in multiclass classification. We also introduce the reason why we may see the confusion matrix as an operator and, therefore, motivate the recourse to the *operator norm* to measure the ‘size’ of the confusion matrix.

In many situations, e.g. class-imbalanced datasets, it is important not to measure the quality of a predictor h on its classification error $\mathbb{P}_{XY}(h(X) \neq Y)$ only, as this may lead to erroneous conclusions regarding the quality of h . Indeed, if, for instance, some class q is predominantly present in the data at hand, say $\mathbb{P}(Y = q) = 1 - \varepsilon$, for some small $\varepsilon > 0$, then the predictor h_{maj} that always outputs $h_{\text{maj}}(x) = q$ regardless of x has a classification error lower than ε . Yet, it might be important not to classify an instance of some class p in class q : take the example of classifying mushrooms according to the categories {hallucinogen, poisonous, innocuous}, it might not be benign to predict innocuous (the majority class) instead of hallucinogen or poisonous. The framework we consider allows us, among other things, to be immune to situations where class-imbalance may occur.

We do claim that a more relevant object to consider is the *confusion matrix* which, given a binary sequence $\mathbf{s} = \{s_1 \cdots s_Q\} \in \{0, 1\}^Q$, is defined as

$$\mathcal{C}_{\mathbf{s}}(h) := \sum_{q: s_q=1} \mathbb{E}_{X|q} L(h, X, q),$$

where, given an hypothesis $h \in \mathcal{H}$, $x \in \mathcal{X}$, $y \in \mathcal{Y}$, $L(h, x, y) = (l_{ij})_{1 \leq i, j \leq Q} \in \mathbb{R}^{Q \times Q}$ is the *loss matrix* such that:

$$l_{ij} := \begin{cases} \ell_j(h, x, y) & \text{if } i = y \text{ and } i \neq j \\ 0 & \text{otherwise.} \end{cases}$$

Note that this matrix has at most one nonzero row, namely its i th row.

For a sequence $\mathbf{y} \in \mathcal{Y}^m$ of m labels and a random sequence \mathbf{X} distributed according to $D_{\mathbf{X}|\mathbf{y}}$, the conditional empirical confusion matrix $\widehat{\mathcal{C}}_{\mathbf{y}}(h, \mathbf{X})$ is

$$\widehat{\mathcal{C}}_{\mathbf{y}}(h, \mathbf{X}) := \sum_{i=1}^m \frac{1}{m_{y_i}} L(h, X_i, y_i) = \sum_{q \in \mathbf{y}} \frac{1}{m_q} \sum_{i: y_i=q} L(h, X_i, q) = \sum_{q \in \mathbf{y}} L_q(h, \mathbf{X}, \mathbf{y}),$$

where

$$L_q(h, \mathbf{X}, \mathbf{y}) := \frac{1}{m_q} \sum_{i: y_i=q} L(h, X_i, q).$$

For a random sequence $\mathbf{Z} = \{\mathbf{X}, \mathbf{Y}\}$ distributed according to D^m , the (unconditional) empirical confusion matrix is given by

$$\mathbb{E}_{\mathbf{X}|\mathbf{Y}} \widehat{\mathcal{C}}_{\mathbf{Y}}(h, \mathbf{X}) = \mathcal{C}_{s(\mathbf{Y})}(h),$$

which is a random variable, as it depends on the random sequence \mathbf{Y} . For exposition purposes it will often be more convenient to consider a fixed sequence \mathbf{y} of labels and state results on $\widehat{\mathcal{C}}_{\mathbf{y}}(h, \mathbf{X})$, noting that

$$\mathbb{E}_{\mathbf{X}|\mathbf{y}} \widehat{\mathcal{C}}_{\mathbf{y}}(h, \mathbf{X}) = \mathcal{C}_{s(\mathbf{y})}(h).$$

The slight differences between our definitions of (conditional) confusion matrices and the usual definition of a confusion matrix is that the diagonal elements are all zero and that they can accommodate any family of loss functions (and not just the 0-1 loss).

A natural objective that may be pursued in multiclass classification is to learn a classifier h with ‘small’ confusion matrix, where ‘small’ might be defined with respect to (some) matrix norm of $\mathcal{C}_s(h)$. The norm that we retain is the *operator norm* that we denote $\|\cdot\|$ from now on: recall that, for a matrix M , $\|M\|$ is computed as

$$\|M\| = \max_{\mathbf{v} \neq \mathbf{0}} \frac{\|M\mathbf{v}\|_2}{\|\mathbf{v}\|_2},$$

where $\|\cdot\|_2$ is the Euclidean norm; $\|M\|$ is merely the largest singular value of M —note that $\|M^\top\| = \|M\|$.

Not only is the operator norm a ‘natural’ norm on matrices but an important reason for working with it is that $\mathcal{C}_s(h)$ is often precisely used as an *operator* acting on the vector of prior distributions

$$\boldsymbol{\pi} = [\mathbb{P}(Y = 1) \cdots \mathbb{P}(Y = Q)]^\top.$$

Indeed, a quantity of interest is for instance the ℓ -risk $R_\ell(h)$ of h , with

$$\begin{aligned} R_\ell(h) &:= \mathbb{E}_{XY} \left\{ \sum_{q=1}^Q \ell_q(h, X, Y) \right\} = \mathbb{E}_Y \left\{ \sum_{q=1}^Q \mathbb{E}_{X|Y} \ell_q(h, X, Y) \right\} \\ &= \sum_{p,q=1}^Q \mathbb{E}_{X|p} \ell_q(h, X, p) \pi_p = \|\boldsymbol{\pi}^\top \mathcal{C}_1(h)\|_1. \end{aligned}$$

It is interesting to observe that, $\forall h, \forall \boldsymbol{\pi} \in \Lambda := \{\boldsymbol{\lambda} \in \mathbb{R}^Q : \lambda_q \geq 0, \sum_q \lambda_q = 1\}$:

$$\begin{aligned} 0 \leq R_{\ell}(h) &= \|\boldsymbol{\pi} \mathcal{C}_1(h)\|_1 = \boldsymbol{\pi}^\top \mathcal{C}_1(h) \mathbf{1} \\ &\leq \sqrt{Q} \|\boldsymbol{\pi}^\top \mathcal{C}_1(h)\|_2 = \sqrt{Q} \|\mathcal{C}_1^\top(h) \boldsymbol{\pi}\|_2 \\ &\leq \sqrt{Q} \|\mathcal{C}_1^\top(h)\| \|\boldsymbol{\pi}\|_2 \leq \sqrt{Q} \|\mathcal{C}_1^\top(h)\| = \sqrt{Q} \|\mathcal{C}_1(h)\|, \end{aligned}$$

where we have used Cauchy-Schwarz inequality in the second line, the definition of the operator norm on the third line and the fact that $\|\boldsymbol{\pi}\|_2 \leq 1$ for any $\boldsymbol{\pi} \in \Lambda$; $\mathbf{1}$ is the Q -dimensional vector where each entry is 1. Recollecting things, we just established the following proposition.

Proposition 1. $\forall h \in \mathcal{H}, R_{\ell}(h) = \|\boldsymbol{\pi}^\top \mathcal{C}_1(h)\|_1 \leq \sqrt{Q} \|\mathcal{C}_1(h)\|$.

This precisely says that the operator norm of the confusion matrix (according to our definition) provides a bound on the risk. As a consequence, bounding $\|\mathcal{C}_1(h)\|$ is a relevant way to bound the risk in a way that is independent from the class priors (since the $\mathcal{C}_1(h)$ is independent from these prior distributions as well). This is essential in class-imbalanced problems and also critical if sampling (prior) distributions are different for training and test data.

Again, we would like to insist on the fact that the confusion matrix *is the subject of our study* for its ability to provide fine-grain information on the prediction errors made by classifiers; as mentioned in the introduction, there are application domains where confusion matrices *indeed are* the measure of performance that is looked at. If needed, the norm of the confusion matrix allows us to summarize the characteristics of the classifiers in one scalar value (the larger, the worse), and it provides, as a (beneficial) “side effect”, a bound on $R_{\ell}(h)$.

3 Deriving Stability Bounds on the Confusion Matrix

One of the most prominent issues in *learning theory* is to estimate the real performance of a learning system. The usual approach consists in studying how empirical measures converge to their expectation. In the traditional settings, it often boils down to providing bounds describing how the empirical risk relates to the expected one. In this work, we show that one can use similar techniques to provide bounds on (the operator norm of) the confusion loss.

3.1 Stability

Following the early work of [7], the risk has traditionally been estimated through its empirical measure and a measure of the complexity of the hypothesis class such as the Vapnik-Chervonenkis dimension, the fat-shattering dimension or the Rademacher complexity. During the last decade, a new and successful approach based on *algorithmic stability* to provide some new bounds has emerged. One of the highlights of this approach is the focus on properties of the learning algorithm at hand, instead of the richness of hypothesis class. In essence, algorithmic stability results aim at taking advantage from the way a given algorithm actually explores the hypothesis space, which

may lead to tight bounds. The main results of [6] were obtained using the definition of *uniform stability*.

Definition 1 (Uniform stability [6]). An algorithm \mathcal{A} has uniform stability β with respect to loss function ℓ if the following holds:

$$\forall \mathbf{Z} \in \mathcal{Z}^m, \forall i \in \{1, \dots, m\}, \|\ell(\mathcal{A}_{\mathbf{Z}}, \cdot) - \ell(\mathcal{A}_{\mathbf{Z} \setminus i}, \cdot)\|_{\infty} \leq \beta.$$

In the present paper, we now focus on the generalization of stability-based results to confusion loss. We introduce the definition of *confusion stability*.

Definition 2 (Confusion stability). An algorithm \mathcal{A} is confusion stable with respect to the set of loss functions ℓ if there exists a constant $B > 0$ such that $\forall i \in \{1, \dots, m\}, \forall \mathbf{z} \in \mathcal{Z}^m$, whenever $m_q \geq 2, \forall q \in \mathcal{Y}$,

$$\sup_{x \in \mathcal{X}} \|L(\mathcal{A}_{\mathbf{z}}, x, y_i) - L(\mathcal{A}_{\mathbf{z} \setminus i}, x, y_i)\| \leq \frac{B}{m_{y_i}}.$$

From here on, q^* , m^* and β^* will stand for

$$q^* := \operatorname{argmin}_q m_q, \quad m^* := m_{q^*}, \quad \text{and} \quad \beta^* := B/m^*.$$

3.2 Noncommutative McDiarmid's Bounded Difference Inequality

Central to the result of [6] is a variation of Azuma's concentration inequality, due to [8]. It describes how a scalar function of independent random variables (the elements of our training set) concentrates around its mean, given how changing one of the random variables impacts the value of the function.

Recently there has been an extension of McDiarmid's inequality to the matrix setting [2]. For the sake of self-containedness, we recall this noncommutative bound.

Theorem 1 (Matrix bounded difference ([2], corollary 7.5)). Let H be a function that maps m variables from some space \mathcal{Z} to a self-adjoint matrix of dimension $2Q$. Consider a sequence $\{A_i\}$ of fixed self-adjoint matrices that satisfy

$$(H(z_1, \dots, z_i, \dots, z_m) - H(z_1, \dots, z'_i, \dots, z_m))^2 \preceq A_i^2, \quad (1)$$

for $z_i, z'_i \in \mathcal{Z}$ and for $i = 1, \dots, m$, where \preceq is the (partial) order on self-adjoint matrices. Then, if \mathbf{Z} is a random sequence of independent variables over \mathcal{Z} :

$$\forall t \geq 0, \mathbb{P} \{\|H(\mathbf{Z}) - \mathbb{E}_{\mathbf{Z}} H(\mathbf{Z})\| \geq t\} \leq 2Qe^{-t^2/8\sigma^2},$$

where $\sigma^2 := \|\sum_i A_i^2\|$.

The confusion matrices we deal with are not necessarily self-adjoint, as is required by the theorem. To make use of the theorem, we rely on the dilation $\mathfrak{D}(A)$ of A , with

$$\mathfrak{D}(A) := \begin{pmatrix} 0 & A \\ A^* & 0 \end{pmatrix},$$

where A^* is the adjoint of A (note that $\mathfrak{D}(A)$ is self-adjoint) and on the result (see [2])

$$\|\mathfrak{D}(A)\| = \|A\|.$$

3.3 Stability Bound

The following theorem is the main result of the paper. It says that the empirical confusion is close to the expected confusion whenever the learning algorithm at hand exhibits confusion-stability properties. This is a new flavor of the results of [6] for the case of matrix-based loss.

Theorem 2 (Confusion bound). *Let \mathcal{A} be a learning algorithm. Assume that all the loss functions under consideration take values in the range $[0; M]$. Let $\mathbf{y} \in \mathcal{Y}^m$ be a fixed sequence of labels.*

If \mathcal{A} is a confusion stable as defined in Definition 2, then, $\forall m \geq 1, \forall \delta \in (0, 1)$, the following holds, with prob. $1 - \delta$ over the random draw of $\mathbf{X} \sim D_{\mathbf{X}|\mathbf{y}}$,

$$\left\| \widehat{\mathcal{C}}_{\mathbf{y}}(\mathcal{A}, \mathbf{X}) - \mathcal{C}_{s(\mathbf{y})}(\mathcal{A}) \right\| \leq 2B \sum_q \frac{1}{m_q} + Q \sqrt{8 \ln \left(\frac{Q^2}{\delta} \right)} \left(4\sqrt{m^*} \beta^* + M \sqrt{\frac{Q}{m^*}} \right).$$

As a consequence, with probability $1 - \delta$ over the random draw of $\mathbf{Z} \sim D^m$,

$$\left\| \widehat{\mathcal{C}}_{\mathbf{Y}}(\mathcal{A}, \mathbf{X}) - \mathcal{C}_{s(\mathbf{Y})}(\mathcal{A}) \right\| \leq 2B \sum_q \frac{1}{m_q} + Q \sqrt{8 \ln \left(\frac{Q^2}{\delta} \right)} \left(4\sqrt{m^*} \beta^* + M \sqrt{\frac{Q}{m^*}} \right).$$

Proof (Sketch). The complete proof can be found in the next subsection. We here provide the skeleton of the proof. We proceed in 3 steps to get the first bound.

1. **Triangle inequality.** To start with, we know by the triangle inequality

$$\begin{aligned} \left\| \widehat{\mathcal{C}}_{\mathbf{y}}(\mathcal{A}, \mathbf{X}) - \mathcal{C}_{s(\mathbf{y})}(\mathcal{A}) \right\| &= \left\| \sum_{q \in \mathbf{y}} (L_q(\mathcal{A}_{\mathbf{Z}}, \mathbf{Z}) - \mathbb{E}_{\mathbf{X}} L_q(\mathcal{A}_{\mathbf{Z}}, \mathbf{Z})) \right\| \\ &\leq \sum_{q \in \mathbf{y}} \|L_q(\mathcal{A}_{\mathbf{Z}}, \mathbf{Z}) - \mathbb{E}_{\mathbf{X}} L_q(\mathcal{A}_{\mathbf{Z}}, \mathbf{Z})\|. \end{aligned} \quad (2)$$

Using uniform stability arguments, we bound each summand with probability $1 - \delta/Q$.

2. **Union Bound.** Then, using the union bound we get a bound on $\|\widehat{\mathcal{C}}(\mathcal{A}, \mathbf{X}) - \mathcal{C}_{s(\mathbf{y})}(\mathcal{A})\|$ that holds with probability at least $1 - \delta$.
3. **Wrap up.** Finally, recouring to a simple argument, we express the obtained bound solely with respect to m^* .

Among the three steps, the first one is the more involved and much part of the proof is devoted to address it.

To get the bound with the unconditional confusion matrix $\mathcal{C}_{s(\mathbf{Y})}(\mathcal{A})$ it suffices to observe that for any event $\mathcal{E}(\mathbf{X}, \mathbf{Y})$ that depends on \mathbf{X} and \mathbf{Y} , such that for all sequences \mathbf{y} , $\mathbb{P}_{\mathbf{X}|\mathbf{y}}\{\mathcal{E}(\mathbf{X}, \mathbf{y})\} \leq \delta$, the following holds:

$$\begin{aligned} \mathbb{P}_{\mathbf{X}\mathbf{Y}}(\mathcal{E}(\mathbf{X}, \mathbf{Y})) &= \mathbb{E}_{\mathbf{X}\mathbf{Y}} \{ \mathbb{I}_{\{\mathcal{E}(\mathbf{X}, \mathbf{Y})\}} \} = \mathbb{E}_{\mathbf{Y}} \{ \mathbb{E}_{\mathbf{X}|\mathbf{Y}} \mathbb{I}_{\{\mathcal{E}(\mathbf{X}, \mathbf{Y})\}} \} \\ &= \sum_{\mathbf{y}} \mathbb{E}_{\mathbf{X}|\mathbf{Y}} \mathbb{I}_{\{\mathcal{E}(\mathbf{X}, \mathbf{Y})\}} \mathbb{P}_{\mathbf{Y}}(\mathbf{Y} = \mathbf{y}) = \sum_{\mathbf{y}} \mathbb{P}_{\mathbf{X}|\mathbf{y}}\{\mathcal{E}(\mathbf{X}, \mathbf{y})\} \mathbb{P}_{\mathbf{Y}}(\mathbf{Y} = \mathbf{y}) \\ &\leq \sum_{\mathbf{y}} \delta \mathbb{P}_{\mathbf{Y}}(\mathbf{Y} = \mathbf{y}) = \delta, \end{aligned}$$

which gives the desired result. \square

Remark 1. If needed, it is straightforward to bound $\|\mathcal{C}_{s(\mathbf{y})}(\mathcal{A})\|$ and $\|\mathcal{C}_{s(\mathbf{Y})}(\mathcal{A})\|$ by using the triangle inequality $\|A\| - \|B\| \leq \|A - B\|$ on the stated bounds.

Remark 2. A few comments may help understand the meaning of our main theorem. First, it is expected to get a bound expressed in terms of $1/\sqrt{m^*}$, since a) $1/\sqrt{m}$ is a typical rate encountered in bounds based on m data and b) the bound cannot be better than a bound devoted to the least informed class (that would be in $1/\sqrt{m^*}$) —resampling procedures may be a strategy to consider to overcome this limit. Second, this theorem says that it is a relevant idea to try and minimize the empirical confusion matrix of a multiclass predictor provided the algorithm used is stable —as will be the case of the algorithms analyzed in the following section. Designing algorithm that minimize the norm of the confusion matrix is therefore an enticing challenge. Finally, when $Q = 2$, that is we are in a binary classification framework, Theorem 2 gives a bound on the maximum of the false-positive rate and the false-negative rate, since this the operator norm of the confusion matrix precisely corresponds to this maximum value.

3.4 Proof of Theorem 2

To ease the readability, we introduce additional notation:

$$\begin{aligned}\mathcal{L}_q &:= \mathbb{E}_{X|q} L(\mathcal{A}_{\mathbf{Z}}, X, q), & \hat{\mathcal{L}}_q &:= L_q(\mathcal{A}_{\mathbf{Z}}, \mathbf{X}, \mathbf{y}), \\ \mathcal{L}_q^i &:= \mathbb{E}_{X|q} L(\mathcal{A}_{\mathbf{Z}^i}, X, q), & \hat{\mathcal{L}}_q^i &:= L_q(\mathcal{A}_{\mathbf{Z}^i}, \mathbf{X}^i, \mathbf{y}^i), \\ \mathcal{L}_q^{\setminus i} &:= \mathbb{E}_{X|q} L(\mathcal{A}_{\mathbf{Z}^{\setminus i}}, X, q), & \hat{\mathcal{L}}_q^{\setminus i} &:= L_q(\mathcal{A}_{\mathbf{Z}^{\setminus i}}, \mathbf{X}^{\setminus i}, \mathbf{y}^{\setminus i}).\end{aligned}$$

After using the triangle inequality in (2), we need to provide a bound on each summand. To get the result, we will, for each q , fix the X_k such that $y_k \neq q$ and work with functions of m_q variables. Then, we will apply Theorem 1 for each

$$H_q(\mathbf{X}_q, \mathbf{y}_q) := \mathfrak{D}(\mathcal{L}_q) - \mathfrak{D}(\hat{\mathcal{L}}_q).$$

To do so, we prove the following lemma

Lemma 1. $\forall q, \forall i, y_i = q$

$$(H_q(\mathbf{Z}_q) - H_q(\mathbf{Z}_q^i))^2 \preceq \left(\frac{4B}{m_q} + \frac{\sqrt{QM}}{m_q} \right)^2 I.$$

Proof. This is a proof that works in 2 steps.

Note that

$$\begin{aligned}\|H_q(\mathbf{X}_q, \mathbf{y}_q) - H_q(\mathbf{X}_q^i, \mathbf{y}_q^i)\| &= \|\mathfrak{D}(\mathcal{L}_q) - \mathfrak{D}(\hat{\mathcal{L}}_q) - \mathfrak{D}(\mathcal{L}_q^i) + \mathfrak{D}(\hat{\mathcal{L}}_q^i)\| \\ &= \|\mathcal{L}_q - \hat{\mathcal{L}}_q - \mathcal{L}_q^i + \hat{\mathcal{L}}_q^i\| \leq \|\mathcal{L}_q - \mathcal{L}_q^i\| + \|\hat{\mathcal{L}}_q - \hat{\mathcal{L}}_q^i\|.\end{aligned}$$

Step 1: bounding $\|\mathcal{L}_q - \mathcal{L}_q^i\|$. We can trivially write:

$$\|\mathcal{L}_q - \mathcal{L}_q^i\| \leq \|\mathcal{L}_q - \mathcal{L}_q^{\setminus i}\| + \|\mathcal{L}_q^i - \mathcal{L}_q^{\setminus i}\|$$

Taking advantage of the stability of \mathcal{A} :

$$\begin{aligned} \|\mathcal{L}_q - \mathcal{L}_q^{\setminus i}\| &= \left\| \mathbb{E}_{X|q} [L(\mathcal{A}_{\mathbf{Z}}, X, q) - L(\mathcal{A}_{\mathbf{Z}^{\setminus i}}, X, q)] \right\| \\ &\leq \mathbb{E}_{X|q} \|L(\mathcal{A}_{\mathbf{Z}}, X, q) - L(\mathcal{A}_{\mathbf{Z}^{\setminus i}}, X, q)\| \\ &\leq \frac{B}{m_q}, \end{aligned}$$

and the same holds for $\|\mathcal{L}_q^i - \mathcal{L}_q^{\setminus i}\|$, i.e. $\|\mathcal{L}_q^i - \mathcal{L}_q^{\setminus i}\| \leq B/m_q$. Thus, we have:

$$\|\mathcal{L}_q - \mathcal{L}_q^i\| \leq \frac{2B}{m_q}. \quad (3)$$

Step 2: bounding $\|\hat{\mathcal{L}}_q - \hat{\mathcal{L}}_q^i\|$. This is a little trickier than the first step.

$$\begin{aligned} \|\hat{\mathcal{L}}_q - \hat{\mathcal{L}}_q^i\| &= \|L_q(\mathcal{A}_{\mathbf{Z}}, \mathbf{Z}) - L_q(\mathcal{A}_{\mathbf{Z}^i}, \mathbf{Z}^i)\| \\ &= \frac{1}{m_q} \left\| \sum_{k:k \neq i, y_k=q} \left(L(\mathcal{A}_{\mathbf{Z}}, X_k, q) - L(\mathcal{A}_{\mathbf{Z}^i}, X_k, q) \right) \right. \\ &\quad \left. + L(\mathcal{A}_{\mathbf{Z}}, X_i, q) - L(\mathcal{A}_{\mathbf{Z}^i}, X'_i, q) \right\| \\ &\leq \frac{1}{m_q} \left\| \sum_{k:k \neq i, y_k=q} \left(L(\mathcal{A}_{\mathbf{Z}^i}, X_k, q) - L(\mathcal{A}_{\mathbf{Z}^i}, X_k, q) \right) \right\| \\ &\quad + \frac{1}{m_q} \left\| L(\mathcal{A}_{\mathbf{Z}}, X_i, q) - L(\mathcal{A}_{\mathbf{Z}^i}, X'_i, q) \right\| \end{aligned}$$

Using the stability argument as before, we have:

$$\begin{aligned} &\left\| \sum_{k:k \neq i, y_k=q} \left(L(\mathcal{A}_{\mathbf{Z}}, X_k, q) - L(\mathcal{A}_{\mathbf{Z}^i}, X_k, q) \right) \right\| \\ &\leq \sum_{k:k \neq i, y_k=q} \|L(\mathcal{A}_{\mathbf{Z}}, X_k, q) - L(\mathcal{A}_{\mathbf{Z}^i}, X_k, q)\| \leq \sum_{k:k \neq i, y_k=q} 2 \frac{B}{m_q} \leq 2B. \end{aligned}$$

On the other hand, we observe that

$$\left\| L(\mathcal{A}_{\mathbf{Z}}, X_i, q) - L(\mathcal{A}_{\mathbf{Z}^i}, X'_i, q) \right\| \leq \sqrt{Q}M.$$

Indeed, the matrix $\Delta := L(\mathcal{A}_{\mathbf{Z}}, X_i, q) - L(\mathcal{A}_{\mathbf{Z}^i}, X'_i, q)$ is a matrix that is zero except for (possibly) its q th row, that we may call δ_q . Thus:

$$\|\Delta\| = \sup_{\mathbf{v}: \|\mathbf{v}\|_2 \leq 1} \|\Delta \mathbf{v}\|_2 = \sup_{\mathbf{v}: \|\mathbf{v}\|_2 \leq 1} \|\delta_q \cdot \mathbf{v}\| = \|\delta_q\|_2,$$

where \mathbf{v} is a vector of dimension Q . Since each of the Q elements of $\boldsymbol{\delta}_q$ is in the range $[-M; M]$, we get that $\|\boldsymbol{\delta}_q\|_2 \leq \sqrt{QM}$.

This allows us to conclude that

$$\|\hat{\mathcal{L}}_q - \hat{\mathcal{L}}_q^i\| \leq \frac{2B}{m_q} + \frac{\sqrt{QM}}{m_q} \quad (4)$$

Combining (3) and (4) we just proved that, for all i such that $y_i = q$

$$(H_q(\mathbf{Z}_q) - H_q(\mathbf{Z}_q^i))^2 \preceq \left(\frac{4B}{m_q} + \frac{\sqrt{QM}}{m_q} \right)^2 I.$$

□

We then establish the following Lemma

Lemma 2. $\forall q$,

$$\mathbb{P}_{\mathbf{X}|\mathbf{y}} \left\{ \|\mathcal{L}_q - \hat{\mathcal{L}}_q\| \geq t + \|\mathbb{E}_{\mathbf{X}|\mathbf{y}}[\mathcal{L}_q - \hat{\mathcal{L}}_q]\| \right\} \leq 2Q \exp \left\{ -\frac{t^2}{8 \left(\frac{4B}{\sqrt{m_q}} + \frac{\sqrt{QM}}{\sqrt{m_q}} \right)^2} \right\}.$$

Proof. Given the previous Lemma, Theorem 1, when applied on $H_q(\mathbf{X}_q, y_q) = \mathcal{D}(\mathcal{L}_q - \hat{\mathcal{L}}_q)$ gives

$$\sigma_q^2 = \left(\frac{4B}{m_q} + \frac{\sqrt{QM}}{\sqrt{m_q}} \right)^2$$

to give, for $t > 0$:

$$\mathbb{P}_{\mathbf{X}|\mathbf{y}} \left\{ \|\mathcal{L}_q - \hat{\mathcal{L}}_q - \mathbb{E}[\mathcal{L}_q - \hat{\mathcal{L}}_q]\| \geq t \right\} \leq 2Q \exp \left\{ -\frac{t^2}{8 \left(\frac{4B}{m_q} + \frac{\sqrt{QM}}{\sqrt{m_q}} \right)^2} \right\},$$

which, using the triangle inequality

$$\| \|A\| - \|B\| \| \leq \|A - B\|,$$

gives the result. □

Finally, we observe

Lemma 3. $\forall q$,

$$\mathbb{P}_{\mathbf{X}|\mathbf{y}} \left\{ \|\mathcal{L}_q - \hat{\mathcal{L}}_q\| \geq t + \frac{2B}{m_q} \right\} \leq 2Q \exp \left\{ -\frac{t^2}{8 \left(\frac{4B}{\sqrt{m_q}} + \frac{\sqrt{QM}}{\sqrt{m_q}} \right)^2} \right\}.$$

Proof. It suffices to show that

$$\left\| \mathbb{E}[\mathcal{L}_q - \hat{\mathcal{L}}_q] \right\| \leq \frac{2B}{m_q},$$

and to make use of the previous Lemma. We note that for any i such that $y_i = q$, and for X'_i distributed according to $D_{X|q}$:

$$\begin{aligned} \mathbb{E}_{\mathbf{X}|\mathbf{y}} \hat{\mathcal{L}}_q &= \mathbb{E}_{\mathbf{X}|\mathbf{y}} L_q(\mathcal{A}_{\mathbf{Z}}, \mathbf{X}, \mathbf{y}) = \frac{1}{m_q} \sum_{j: y_j=q} \mathbb{E}_{\mathbf{X}|\mathbf{y}} L(\mathcal{A}_{\mathbf{Z}}, X_j, q) \\ &= \frac{1}{m_q} \sum_{j: y_j=q} \mathbb{E}_{\mathbf{X}, X'_j|\mathbf{y}} L(\mathcal{A}_{\mathbf{Z}^i}, X'_j, q) = \mathbb{E}_{\mathbf{X}, X'_i|\mathbf{y}} L(\mathcal{A}_{\mathbf{Z}^i}, X'_i, q). \end{aligned}$$

Hence, using the stability argument,

$$\begin{aligned} \left\| \mathbb{E}[\mathcal{L}_q - \hat{\mathcal{L}}_q] \right\| &= \left\| \mathbb{E}_{\mathbf{X}, X'_i|\mathbf{y}} [L(\mathcal{A}_{\mathbf{Z}}, X'_i, q) - L(\mathcal{A}_{\mathbf{Z}^i}, X'_i, q)] \right\| \\ &\leq \mathbb{E}_{\mathbf{X}, X'_i|\mathbf{y}} \|L(\mathcal{A}_{\mathbf{Z}}, X'_i, q) - L(\mathcal{A}_{\mathbf{Z}^i}, X'_i, q)\| \\ &\leq \mathbb{E}_{\mathbf{X}, X'_i|\mathbf{y}} \|L(\mathcal{A}_{\mathbf{Z}}, X'_i, q) - L(\mathcal{A}_{\mathbf{Z} \setminus i}, X'_i, q)\| \\ &\quad + \mathbb{E}_{\mathbf{X}, X'_i|\mathbf{y}} \|L(\mathcal{A}_{\mathbf{Z}^i}, X'_i, q) - L(\mathcal{A}_{\mathbf{Z} \setminus i}, X'_i, q)\| \\ &\leq \frac{2B}{m_q}. \end{aligned}$$

This inequality in combination with the previous lemma provides the result. \square

We are now set to make use of a union bound argument:

$$\begin{aligned} \mathbb{P} \left\{ \exists q : \|\mathcal{L}_q - \hat{\mathcal{L}}_q\| \geq t + \frac{2B}{m_q} \right\} &\leq \sum_{q \in \mathcal{Y}} \mathbb{P} \left\{ \exists q : \|\mathcal{L}_q - \hat{\mathcal{L}}_q\| \geq t + \frac{2B}{m_q} \right\} \\ &\leq 2Q \sum_q \exp \left\{ -\frac{t^2}{8 \left(\frac{4B}{\sqrt{m_q}} + \frac{\sqrt{QM}}{\sqrt{m_q}} \right)^2} \right\} \leq 2Q^2 \max_q \exp \left\{ -\frac{t^2}{8 \left(\frac{4B}{\sqrt{m_q}} + \frac{\sqrt{QM}}{\sqrt{m_q}} \right)^2} \right\} \end{aligned}$$

According to our definition m^* , we get

$$\mathbb{P} \left\{ \exists q : \|\mathcal{L}_q - \hat{\mathcal{L}}_q\| \geq t + \frac{2B}{m_q} \right\} \leq 2Q^2 \exp \left\{ -\frac{t^2}{8 \left(\frac{4B}{\sqrt{m^*}} + \frac{\sqrt{QM}}{\sqrt{m^*}} \right)^2} \right\}.$$

Setting the right hand side to δ , gives the result of Theorem 2.

4 Analysis of existing algorithms

Now that the main result on stability bound has been established, we will investigate how existing multiclass algorithms exhibit stability properties and thus fall in the scope of our analysis. More precisely, we will analyse two well-known models for multiclass support vector machines and we will show that they may promote small confusion error. But first, we will study the more general stability of multiclass algorithms using regularization in Reproducing Kernel Hilbert Spaces (RKHS).

4.1 Hilbert Space Regularized Algorithms

Many well-known and widely-used algorithms feature a minimization of a regularized objective functions [9]. In the context of multiclass kernel machines [10, 11], this regularizer $\Omega(h)$ may take the following form:

$$\Omega(h) = \sum_q \|h_q\|_k^2.$$

where $k : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ denotes the kernel associated to the RKHS \mathcal{H} .

In order to study the stability properties of algorithms, minimizing a data-fitting term, penalized by such regularizers, in our multi-class setting, we need to introduce a minor definition that is an addition to definition 19 of [6].

Definition 3. A loss function ℓ defined on $\mathcal{H}^Q \times \mathcal{Y}$ is σ -multi-admissible if ℓ is σ -admissible with respect to any of his Q first arguments.

This allows us to come up with the following theorem.

Theorem 3. Let \mathcal{H} be a reproducing kernel Hilbert space (with kernel k) such that $\forall X \in \mathcal{X}, k(X, X) \leq \kappa^2 < +\infty$. Let L be a loss matrix, such that $\forall q \in \mathcal{Y}, \ell_q$ is σ_q -multi-admissible. And let \mathcal{A} be an algorithm such that

$$\begin{aligned} \mathcal{A}_S &= \operatorname{argmin}_{h \in \mathcal{H}^Q} \sum_q \sum_{n: y_n=q} \frac{1}{m_q} \ell_q(h, x_n, q) + \lambda \sum_q \|h_q\|_k^2. \\ &:= \operatorname{argmin}_{h \in \mathcal{H}^Q} J(h). \end{aligned}$$

Then \mathcal{A} is confusion stable with respect to the set of loss functions ℓ . Moreover, a B value defining the stability is

$$B = \max_q \frac{\sigma_q^2 Q \kappa^2}{2\lambda},$$

where κ is such that $k(X, X) \leq \kappa^2 < +\infty$

Proof (Sketch of proof). In essence the idea is to exploit Definition 3 in order to apply Theorem 22 of [6] for each loss ℓ_q . Moreover our regularizer is a sum (over q) of RKHS norms, hence the additional Q in the value of B . \square

4.2 Lee, Lin and Wahba model

One of the most well-known and well-studied model for multi-class classification, in the context of SVM, was proposed by [12]. In this work, the authors suggest the use of the following loss function.

$$\ell(h, x, y) = \sum_{q \neq y} \left(h_q(x) + \frac{1}{Q-1} \right)_+$$

Their algorithm, denoted \mathcal{A}^{LLW} , then consists in minimizing the following (penalized) functional,

$$J(h) = \frac{1}{m} \sum_{k=1}^m \sum_{q \neq y_k} \left(h_q(x_k) + \frac{1}{Q-1} \right)_+ + \lambda \sum_{q=1}^Q \|h_q\|^2,$$

with the constraint $\sum_q h_q = 0$.

We can trivially rewrite $J(h)$ as

$$J(h) = \sum_q \sum_{n: y_n=q} \frac{1}{m_q} \ell_q(h, x_n, q) + \lambda \sum_{q=1}^Q \|h_q\|^2,$$

with

$$\ell_q(h, x_n, q) = \sum_{p \neq q} \left(h_p(x_k) + \frac{1}{Q-1} \right)_+.$$

It is straightforward that for any q , ℓ_q is 1-multi-admissible. We thus can apply theorem 3 and get $B = Q\kappa^2/2\lambda$.

Lemma 4. *Let h^* denote the solution found by \mathcal{A}^{LLW} . $\forall x \in \mathcal{X}, \forall y \in \mathcal{Y}, \forall q$, we have*

$$\ell_q(h^*, x, y) \leq \frac{Q\kappa}{\sqrt{\lambda}} + 1.$$

Proof. As h^* is a minimizer of J , we have

$$J(h^*) \leq J(0) = \sum_q \sum_{n: y_n=q} \frac{1}{m_q} \ell_q(0, x_n, q) = \sum_q \sum_{n: y_n=q} \frac{1}{(Q-1)m_q} = 1.$$

As the data fitting term is non-negative, we also have

$$J(h^*) \geq \lambda \sum_q \|h_q^*\|^2.$$

Given that $h^* \in \mathcal{H}$, Cauchy-Schwarz inequality gives

$$\forall x \in \mathcal{X}, \|h_q^*\|_k \geq \frac{|h_q^*(x)|}{\kappa}.$$

Collecting things, we have

$$\forall x \in \mathcal{X}, |h_q^*(x)| \leq \frac{\kappa}{\sqrt{\lambda}}.$$

Going back to the definition of ℓ_q , we get the result. \square

Using theorem 2, it follows that, with probability $1 - \delta$,

$$\left\| \widehat{\mathcal{C}}_{\mathbf{Y}}(\mathcal{A}^{\text{LLW}}, \mathbf{X}) - \mathcal{C}_{\mathbf{s}(\mathbf{Y})}(\mathcal{A}^{\text{LLW}}) \right\| \leq \sum_q \frac{Q\kappa^2}{\lambda m_q} + \frac{\sqrt{8 \ln \left(\frac{Q^2}{\delta} \right) \left(\frac{2Q^2\kappa^2}{\lambda} + \left(\frac{Q\kappa}{\sqrt{\lambda}} + 1 \right) Q\sqrt{Q} \right)}}{\sqrt{m^*}}.$$

4.3 Weston and Watkins model

Another multiclass mode is due to [13]. They consider the following loss functions.

$$\ell(h, x, y) = \sum_{q \neq y} (1 - h_y(x) + h_q(x))_+$$

The algorithm \mathcal{A}^{WW} minimizes the following functional

$$J(h) = \frac{1}{m} \sum_{k=1}^m \sum_{q \neq y_k} (1 - h_y(x) + h_q(x))_+ + \lambda \sum_{q < p=1}^Q \|h_q - h_p\|^2,$$

This time, for $1 \leq p, q \leq Q$, we will introduce the functions $h_{pq} = h_p - h_q$. We can then rewrite $J(h)$ as

$$J(h) = \sum_q \sum_{n: y_n=q} \frac{1}{m_q} \ell_q(h, x_n, q) + \lambda \sum_{p=1}^Q \sum_{q=1}^{p-1} \|h_{pq}\|^2,$$

with

$$\ell_q(h, x_n, q) = \sum_{p \neq q} (1 - h_{pq}(x_n))_+.$$

It still is straightforward that for any q , ℓ_q is 1-multi-admissible. However, this time, our regularizer consists in the sum of $\frac{Q(Q-1)}{2} < \frac{Q^2}{2}$ norms. Applying Theorem 3 therefore gives $B = Q^2 \kappa^2 / 4\lambda$.

Lemma 5. *Let h^* denote the solution found by \mathcal{A}^{WW} . $\forall x \in \mathcal{X}, \forall y \in \mathcal{Y}, \forall q$, we have $\ell_q(h^*, x, y) \leq Q \left(1 + \kappa \sqrt{\frac{Q}{\lambda}}\right)$.*

This lemma can be proven following exactly the same techniques and reasoning as Lemma 4.

Using theorem 2, it follows that, with probability $1 - \delta$,

$$\left\| \widehat{\mathcal{C}}_{\mathbf{Y}}(\mathcal{A}^{\text{WW}}, \mathbf{X}) - \mathcal{C}_{\mathbf{s}(\mathbf{Y})}(\mathcal{A}^{\text{WW}}) \right\| \leq \sum_q \frac{Q^2 \kappa^2}{2\lambda m_q} + \frac{\sqrt{8 \ln\left(\frac{Q^2}{\delta}\right)} \left(\frac{Q^3 \kappa^2}{\lambda} + Q^2 \left(\sqrt{Q} + \kappa \sqrt{\frac{Q}{\lambda}}\right)\right)}{\sqrt{m^*}}.$$

5 Discussion and Conclusion

In this paper, we have proposed a new framework, namely the algorithmic *confusion stability*, together with new bounds to characterize the generalization properties of multiclass learning algorithms. The crux of our study is to envision the confusion matrix

as a performance measure, which differs from commonly encountered approaches that investigate generalization properties of scalar-valued performances.

A few questions that are raised by the present work are the following. Is it possible to derive confusion stable algorithms that precisely aim at controlling the norm of their confusion matrix? Are there other algorithms than those analyzed here that may be studied in our new framework? On a broader perspective: how can noncommutative concentration inequalities be of help to analyze complex settings encountered in machine learning (such as, e.g., structured prediction, operator learning)?

References

1. Recht, B.: A simpler approach to matrix completion. *Journal of Machine Learning Research* **12** (2011) 3413–3430
2. Tropp, J.A.: User-friendly tail bounds for sums of random matrices. *Foundations of Computational Mathematics* (august 2011)
3. Gosh, A., Kale, S., McAfee, P.: Who moderates the moderators?: crowdsourcing abuse detection in user-generated content. In: *Proc. of the 12th ACM conference on Electronic commerce, EC 11.* (2011) 167–176
4. Rudelson, M., Vershynin, R.: Sampling from large matrices: An approach through geometric functional analysis. *J. ACM* **54**(4) (2007)
5. Chaudhuri, K., Kakade, S., Livescu, K., Sridharan, K.: Multi-view clustering via canonical correlation analysis. In: *Proc. of the 26th Int. Conf. on Machine Learning – ICML 09.* ICML '09, New York, NY, USA, ACM (2009) 129–136
6. Bousquet, O., Elisseeff, A.: Stability and generalization. *JMLR* **2** (march 2002) 499–526
7. Vapnik, V.N.: *Estimation of Dependences Based on Empirical Data.* Springer-Verlag (1982)
8. McDiarmid, C.: On the method of bounded differences. In: *Surveys in Combinatorics.* (1989) 148–188
9. Tikhonov, A.N., Arsenin, V.Y.: *Solutions of Ill-Posed Problems.* Winston (1977)
10. Crammer, K., Singer, Y.: On the algorithmic implementation of multiclass kernel-based vector machines. *Journal of Machine Learning Research* **2** (2001) 2001
11. Cristianini, N., Shawe-Taylor, J.: *An Introduction to Support Vector Machines.* Cambridge University Press (2000)
12. Lee, Y., Lin, Y., Wahba, G.: Multicategory support vector machines. *J. of the American Statistical Association* **99** (2004) 67–81
13. Weston, J., Watkins, C.: Multi-class support vector machines. Technical report, Royal Holloway, University of London (1998)