Confusion Matrix Stability Bounds for Multiclass Classification

Pierre Machart, Liva Ralaivola

Confusion Matrix Stability Bounds for Multiclass Classification

Pierre Machart Liva Ralaivola
Aix-Marseille Univ., LIF-QARMA, CNRS, UMR 7279, F-13013, Marseille, France
{*firstname.name*}*@lif.univ-mrs.fr*

February 28, 2012

Abstract

In this paper, 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.

Keywords: Machine Learning, Stability Generalization Bounds, Confusion Matrix, Non-Commutative Concentration Inequality, Multi-Class

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 nowadays, given the widespread need of such predictors in information retrieval, web mining, bioinformatics or neuroscience.

Yet, the literature on multiclass learning is not as voluminous than that of binary classification, whereas this problem 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. More precisely, building on very recent matrix-based concentration inequalities provided by Tropp (2011) —sometimes referred to as noncommutative concentration inequalities— we establish a stability based framework for confusion-aware 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 Bousquet and Elisseeff (2002), which are designed for scalar loss functions. To the best of our knowledge, this is the first work that establishes generalization bounds for 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* procdures 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 Matrix Awareness

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

We will use $D_{X|y}$ for the conditional distribution on X given that $Y = y$; therefore, for a given sequence $\boldsymbol{y} = \{y_1, \dots, y_m\} \in \mathcal{Y}^m$, $D_{\boldsymbol{X}|\boldsymbol{y}} = \otimes_{i=1}^m D_{X|y_i}$ is the distribution of the random sample $\boldsymbol{X} = \{X_1, \dots, X_m\}$ over $\mathcal{X}^T m$ such that X_i is distributed according to $D_{X|y_i}$; for $q \in \mathcal{Y}$, and $\boldsymbol{X}$ distributed according to $D_{\boldsymbol{X}|\boldsymbol{y}}$, $\boldsymbol{X}_q = \{X_{i_1}, \dots, X_{i_{m_q}}\}$ 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 $\{Z_1, \dots Z_{i-1}, Z_i', Z_{i+1}, \dots, Z_m\}$ where Z_i' is distributed as Z_i; $\mathbf{Z}^{\setminus i}$ is the sequence $\{Z_1, \dots Z_{i-1}, Z_{i+1}, \dots, Z_m\}$ — these definitions carry directly over when conditioned on a sequence of labels $\boldsymbol{y}$ (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 qth coordinate. Also, $\boldsymbol{\ell} = (\ell_q)_{1 \le q \le Q}$ is a set of loss functions such that: $\ell_q : \mathcal{H} \times \mathcal{X} \times \mathcal{Y} \to \mathbb{R}_+$.

2.2 Confusion Matrix

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. As the definition of the confusion loss in the online learning framework is less usual and a bit more intricate than its definition in the batch scenario, we develop the discussion here within the batch setting —obviously, the argument carries over to the online learning framework.

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; 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 thant ε. Yet, it might be important not to classify an instance of some class p in class q: in the context of classifying mushrooms according to the categories {`hallucinogen`, `poisonous`, `innocuous`}, the consequence of predicting `innocuous` (the majority class) instead of `hallucinogen` or `poisonous` might be disastrous.

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

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

where, given an hypothesis $h \in \mathcal{H}$, $x \in \mathcal{X}$, $y \in \mathcal{Y}$, $L(h, x, y) = (l_{ij})_{1 \le i,j \le Q} \in \mathbb{R}^{Q \times Q}$ 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 ith row.

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

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

where

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

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

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

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

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

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 accomodate 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 $\mathcal{C}_{\boldsymbol{s}}(h)$. The norm that we retain is the *operator norm* that we denote $\|\cdot\|$ from now on: for a matrix M, $\|M\|$ is defined as

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

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

A nice reason for focusing on the operator norm is that $\mathcal{C}_{\boldsymbol{s}}(h)$ is often precisely used as an operator that acts on the vector of prior distributions

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

Indeed, a quantity of interest is for instance the *risk* $R_{\boldsymbol{\ell}}(h)$ of h, with

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

It is interesting to observe that, $\forall h, \forall \boldsymbol{\pi}$:

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

where we have used Cauchy-Schwarz inequalty 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 $\{\boldsymbol{\lambda} \in \mathbb{R}^Q : \lambda_q \geq 0, \sum_q \lambda_q = 1\}$.

In addition to support the use of the operator norm, this also says that bounding the norm of the confusion loss is a good way to control the risk of h as well (independently of the prior distribution, see discussion below).

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 matrix.

3.1 Stability

Following the early work of Vapnik (1982), the risk has traditionally been estimated through its empirical measure and a measure of the complexity of the hypothesis class such as Vapnik-Chervonenkis (VC-dim), fat-shattering dimension or 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 some properties of the learning algorithm at hand, instead of the characterization of the hypothesis class. Roughly, this makes it possible to take advantage from the knowledge on how a given algorithm actually explores the hypothesis space, often leading to tighter bounds.

The main results in Bousquet and Elisseeff (2002) were obtained using the following definition of *uniform stability*.

Definition 1 (Uniform stability Bousquet and Elisseeff (2002))**.** *An algorithm A has* uniform stability β *with respect to the loss function l if the following holds:*

$$\forall S \in \mathcal{Z}^m, \forall i \in \{1, \dots, m\}, \|l(A_S, .) - l(A_{S^{\setminus i}}, .)\|_\infty \leq \beta.$$

Following the same approach, we now focus on the generalization of such results for confusion matrices. We introduce a new definition of *confusion stability*.

Definition 2 (Confusion stability)**.** *An algorithm $\mathcal{A}$ is* confusion stable *with respect to the loss matrix L if $\forall q \in \mathcal{Y}$, there exist a β_q decreasing as $\frac{1}{m_q}$ such that $\mathcal{A}$ has β_q uniform stability with respect to the loss ℓ_q*

3.2 Non-Commutative McDiarmid

In Bousquet and Elisseeff (2002), the authors make use of some well-known concentration inequalities to derive bounds. More specifically, they use a variation of Azuma's inequality, due to McDiarmid McDiarmid (1989). It describes how a scalar function of independent random variables (the elements of our training set) normally concentrates around its mean, with variance depending on how changing one of the random variables impacts the value of the function.

Some more recent work Tropp (2011) extends McDiarmid's inequality to the matrix setting. For the sake of self-containedness, we recall this non-commutative bound.

Theorem 1 (Matrix bounded difference (Tropp (2011), corollary 7.5))**.** *Let S and S^i be defined as above, and let H be a function that maps m variables to a self-adjoint matrix of dimension Q. Consider a sequence $\{A_i\}$ of fixed self-adjoint matrices that satisfy*

$$\big(H(S) - H(S^i)\big)^2 \preccurlyeq A_k^2, \tag{1}$$

where z_i and z_i range over all possible values of the space $\mathcal{Z}_i$ it belongs to, for each index i. Then, for all $t \geq 0$,

$$\mathbb{P}\{\lambda_{max}\big(H(z) - \mathbb{E}H(z)\big) \geq t\} \leq Qe^{-t^2/8\sigma^2},$$

where $z = (Z_1, \dots, Z_m)$ and $\sigma^2 := \|\sum_i A_i^2\|$.

One may notice that H has to be applied on a function mapping to self-adjoint matrices. Unfortunately, our confusion matrices are real-valued but are not symmetric. However, we can make use of a dilation technique to overcome this.

Namely, instead of working directly on a non-self-adjoint matrix A, we will build a self-adjoint dilated matrix $\mathfrak{D}(A)$:

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

where A^* denotes the adjoint of A.

The choice of such a dilation is directly motivated by the following property, recalled in Tropp (2011).

Lemma 1. *If λ is the largest eigenvalue of $\mathfrak{D}(A)$, then λ is the largest singular value of A.*

As a direct consequence, any result on the largest eigenvalue of a dilated matrix $\mathfrak{D}(A)$ directly applies to the operator norm $\|A\|$ of A.

3.3 Stability Bound

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 $\boldsymbol{y} \in \mathcal{Y}^m$ be a fixed sequence of labels.*

If $\mathcal{A}$ has confusion stability β with respect to all loss matrices L_q, for $q \in \mathcal{Y}$, then, $\forall m \geq 1$, $\forall \delta \in (0, 1)$, the following holds, with probability $1 - \delta$ over the random draw of $\boldsymbol{X} \sim D_{\boldsymbol{X}|\boldsymbol{y}}$,

$$\left\|\widehat{\mathcal{C}}_{\boldsymbol{y}}(\mathcal{A}, \boldsymbol{X}) - \mathcal{C}_{\boldsymbol{s}(\boldsymbol{y})}(\mathcal{A})\right\| \leq 2\sum_q \beta_q + Q\sqrt{8\ln\left(\frac{Q^2}{\delta}\right)}\left(4\sqrt{m_{min}}\beta_{\min} + M\sqrt{\frac{Q}{m_{min}}}\right).$$

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

$$\left\|\widehat{\mathcal{C}}_{\boldsymbol{Y}}(\mathcal{A}, \boldsymbol{X}) - \mathcal{C}_{\boldsymbol{s}(\boldsymbol{Y})}(\mathcal{A})\right\| \leq 2\sum_q \beta_q + Q\sqrt{8\ln\left(\frac{Q^2}{\delta}\right)}\left(4\sqrt{m_{min}}\beta_{\min} + M\sqrt{\frac{Q}{m_{min}}}\right).$$

Sketch of proof. The complete proof can be found in the following subsection. We proceed in three steps to proof the first bound. To start with, we know that by the triangle inequality

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

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

Then, using the union bound we have a bound on $\|\widehat{\mathcal{C}}(\mathcal{A}, \boldsymbol{X}) - \mathcal{C}_{s(\boldsymbol{y})}(\mathcal{A})\|$ that holds with probability at least $1 - \delta$.

Finally, recoursing to a simple argument, we express the obtained bound solely with respect to $m_{\min}$.

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

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

which gives the desired result. □

Remark 1. It is straightforward to directly obtain a bound on $\|\mathcal{C}_{\boldsymbol{s}(\boldsymbol{y})}(\mathcal{A})\|$ and $\|\mathcal{C}_{\boldsymbol{s}(\boldsymbol{Y})}(\mathcal{A})\|$ by using the triangle inequality $|\|A\| - \|B\|| \leq \|A - B\|$ on the bounds given in the theorem.

3.4 Proof of Theorem 2

Proof. To ease the readability, we introduce some additional notation: $\mathbf{Z}, \mathbf{Z}^i, \mathbf{Z}^{\setminus i}$ will refer respectively to $\{\boldsymbol{X}, \boldsymbol{y}\}, \{\boldsymbol{X}^i, \boldsymbol{y}^i\}, \{\boldsymbol{X}^{\setminus i}, \boldsymbol{y}^{\setminus i}\}$, and

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

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

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

To do so, we will bound the differences

$$\|H_q(\boldsymbol{X}_q, \boldsymbol{y}_q) - H_q(\boldsymbol{X}_q^i, \boldsymbol{y}_q^i)\|.$$

Note that

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

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

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

Using the β_q-stability of $\mathcal{A}$:

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

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

$$\|\mathcal{L}_q - \mathcal{L}_q^i\| \leq 2\beta_q. \tag{3}$$

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

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

Using the β_q-stability argument as before, we have:

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

On the other hand, we observe that

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

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 qth row, that we may call $\boldsymbol{\delta}_q$. Thus:

$$\|\Delta\| = \sup_{\boldsymbol{v}:\|\boldsymbol{v}\|_2\leq 1} \|\Delta \boldsymbol{v}\|_2 = \sup_{\boldsymbol{v}:\|\boldsymbol{v}\|_2\leq 1} \|\boldsymbol{\delta}_q \cdot \boldsymbol{v}\| = \|\boldsymbol{\delta}_q\|_2,$$

where $\boldsymbol{v}$ 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 $\|\boldsymbol{\delta}_q\|_2 \leq \sqrt{Q}M$.

This allows us to conclude that

$$\|\hat{\mathcal{L}}_q - \hat{\mathcal{L}}_q^i\| \leq 2\beta_q + \frac{\sqrt{Q}M}{m_q} \tag{4}$$

Step 3: Applying Matrix McDiarmid Combining (3) and (4) that we just proved, we have that, for all i such that $y_i = q$

$$(H_q(\boldsymbol{Z}_q) - H_q(\boldsymbol{Z}_q^i))^2 \preccurlyeq \left(4\beta_q + \frac{\sqrt{Q}M}{m_q}\right)^2 I.$$

Therefore, Theorem 1 may be applied on $H_q(\boldsymbol{X}_q, y_q) = \mathfrak{D}(\mathcal{L}_q - \hat{\mathcal{L}}_q)$ with

$$\sigma_q^2 = m_q\beta_q = \left(4\sqrt{m_q}\beta_q + \frac{\sqrt{Q}M}{\sqrt{m_q}}\right)^2$$

to give, for $t > 0$:

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

which, using the triangle inequality

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

gives

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

We may want to bound $\|\mathbb{E}_{\boldsymbol{X}|\boldsymbol{y}}[\mathcal{L}_q - \hat{\mathcal{L}}_q]\|$. To do so, we note that for any i such that $y_i = q$, and for X_i' distributed according to $D_{X|q}$:

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

Hence, using the β_q stability

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

This leads to

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

Step 4. Union Bound Now, using the union bound, we have

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

According to our definition of *confusion stability* (cf. Definition 2), β_q decreases as $\frac{1}{m_q}$. Therefore

$$\mathbb{P}\left\{\exists q : \|\mathcal{L}_q - \hat{\mathcal{L}}_q\| \geq t + 2\beta_q\right\} \leq 2Q^2 \max_q \exp\left\{-\frac{t^2}{8\left(4\sqrt{m_{\min}}\beta_{\min} + \frac{\sqrt{Q}M}{\sqrt{m_{\min}}}\right)^2}\right\}$$

where $m_{\min} = \min_q m_q$ and $\beta_{\min}$ is the associated β_q. Setting the right hand side to δ, we can get, with probability $1-\delta$,

$$\sum_q \|C^q - C^q_{\text{emp}}\| \leq 2\sum_q \beta_q + Q\sqrt{8\ln\left(\frac{2Q^2}{\delta}\right)}\left(4\sqrt{m_{\min}}\beta_{\min} + M\sqrt{\frac{Q}{m_{\min}}}\right)$$

Hence the result. □

4 Analysis of existing algorithms

Now that the main result on stability bound has been established, we will investigate how some already existing multi-class algorithms display some 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 to promote small confusion error. But first, we will study the more general stability of multi-class 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 functionTikhonov and Arsen (1977). In the context of kernel machinesCristianini and Shawe-Taylor (2000), this regularizer $\Omega(h)$ may take the following form:

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

where $k : \mathcal{X}\times\mathcal{X} \to \mathbb{R}$ 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 Bousquet and Elisseeff (2002).

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 $\forall q \in \mathcal{Y}$, ℓ_q is σ_q-multi-admissible. And let $\mathcal{A}$ be an algorithm such that*

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

Then $\mathcal{A}$ is confusion stable with respect to the loss matrix L. Moreover, $\forall q \in \mathcal{Y}$, we have

$$\beta_q \leq \frac{\sigma_q^2 Q \kappa^2}{2\lambda m_q}$$

Sketch of proof. Roughly, the idea is to exploit definition 3 in order to apply the theorem 22 of Bousquet and Elisseeff (2002) for each loss ℓ_q. Moreover our regularizer is a sum (over q) of RKHS norms, hence the additional Q in the the bound on β_q. □

From now on, we will always suppose that we are working with kernels such that $k(X, X) \leq \kappa^2 < +\infty$.

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 Lee et al. (2004). In this work, the authors suggest the use of the following loss function.

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

Their algorithm, denoted $\mathcal{A}^{\text{LLW}}$, 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 $\beta_q \leq \frac{Q\kappa^2}{2\lambda m_q}$.

Lemma 2. *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

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

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

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

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. □

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

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

With regards to the β_q we obtained, one can conclude that $\mathcal{A}^{\mathrm{LLW}}$ is confusion stable.

4.3 Weston and Watkins model

One of the oldest models, when it comes to multi-class SVM is due to Weston and Watkins (1998). They consider the following loss functions.

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

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

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

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

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

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 us $\beta_q\leq\frac{Q^2\kappa^2}{4\lambda m_q}$

Lemma 3. *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 2.

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

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

With regards to the β_q we obtained, one can conclude that $\mathcal{A}^{\mathrm{WW}}$ is confusion stable.

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 (such as, e.g., the accuracy).

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? In particular, is it the case for k-nearest-neighbors, the generalization analysis of which is amenable thanks to classical stability arguments? On a broader perspective: how can noncommutative concentration inequalities be of some use in machine learning?

References

Bousquet, O. and Elisseeff, A. (2002). Stability and generalization. *Journal of Machine Learning Research*, 2:499–526.

Cristianini, N. and Shawe-Taylor, J. (2000). *An Introduction to Support Vector Machines: and Other Kernel-Based Learning Methods*. Cambridge University Press.

Lee, Y., Lin, Y., and Wahba, G. (2004). Multicategory support vector machines. *Journal of the American Statistical Association*, 99:67–81.

McDiarmid, C. (1989). On the method of bounded differences. In *Surveys in Combinatorics*, pages 148–188.

Tikhonov, A. N. and Arsenin, V. Y. (1977). *Solutions of Ill-Posed Problems*. Winston.

Tropp, J. A. (2011). User-friendly tail bounds for sums of random matrices. *Foundations of Computational Mathematics*.

Vapnik, V. N. (1982). *Estimation of Dependences Based on Empirical Data*. Springer-Verlag.

Weston, J. and Watkins, C. (1998). Multi-class support vector machines. Technical report, Royal Holloway, University of London, Department of Computer Science.