Radius-Margin Bound on the Leave-One-Out Error of the LLW-M-SVM

Yann Guermeur
LORIA-CNRS
Campus Scientifique, BP 239
54506 Vandœuvre-lès-Nancy cedex, France
(e-mail: Yann.Guermeur@loria.fr)

Emmanuel Monfrini
TELECOM & Management SudParis
9 rue Charles Fourier
91011 EVRY cedex, France
(e-mail: Emmanuel.Monfrini@it-sudparis.eu)

January 26, 2009

Abstract

Using a support vector machine (SVM) requires to set the values of two types of hyperparameters: the soft margin parameter C and the parameters of the kernel. To perform this model selection task, the method of choice is cross-validation. Its leave-one-out variant is known to produce an estimator of the generalization error which is almost unbiased. Its major drawback rests in its time requirement. To overcome this difficulty, several upper bounds on the leave-one-out error of the pattern recognition SVM have been derived. Among those bounds, the most popular one is probably the radius-margin bound. In this report, we establish a generalized radius-margin bound dedicated to the multi-class SVM of Lee, Lin and Wahba.

Keywords: M-SVMs, model selection, leave-one-out error, radius-margin bound

1 Introduction

Using a SVM [3, 6] requires to set the values of two types of hyperparameters: the soft margin parameter C and the parameters of the kernel. To perform this model selection task, several approaches are available (see for instance [13, 16]). The solution of choice consists in applying a cross-validation procedure. Among those procedures, the leave-one-out one appears especially attractive, since it is known to produce an estimator of the generalization error which is almost unbiased [15]. The seamy side of things is that it is highly time consuming. This is the reason why, in recent years, a number of upper bounds on the leave-one-out error of the (standard) pattern recognition SVM have been proposed in literature (see [5] for a survey). Among those bounds, the tightest one is the *span bound* [21]. However, the results of Chapelle and co-workers presented in [5] show that another bound, the *radius-margin* one [20], achieves equivalent performance for model selection while being far simpler to compute. This is the reason why it is currently the most popular bound. In this report, we establish a generalized radius-margin bound on the leave-one-out error of the hard margin version of the multi-class SVM (M-SVM) of Lee, Lin and Wahba [14].

The organization of this paper is as follows. Section 2 presents the M-SVMs, by describing their common architecture and the general form taken by their different training algorithms. Section 3 focuses on the M-SVM of Lee, Lin and Wahba. Section 4 is devoted to the formulation and proof of the corresponding multi-class radius-margin bound. At last, we draw conclusions and outline our ongoing research in Section 5.

2 Multi-Class SVMs

Like the SVMs, the M-SVMs are large margin classifiers which are devised in the framework of Vapnik's statistical learning theory [20].

2.1 Formalization of the learning problem

We are interested here in multi-class pattern recognition problems. Formally, we consider the case of Q-category classification problems with $3 \leq Q < \infty$, but our results extend to the case of dichotomies. Each object is represented by its description $x \in \mathcal{X}$ and the set $\mathcal{Y}$ of the categories y can be identified with the set of indexes of the categories: $[\![1, Q]\!]$. We assume that the link between objects and categories can be described by an unknown probability measure P on the product space $\mathcal{X} \times \mathcal{Y}$. The aim of the learning problem consists in selecting in a set $\mathcal{G}$ of functions $g = (g_k)_{1 \leq k \leq Q}$ from $\mathcal{X}$ into $\mathbb{R}^Q$ a function classifying data in an optimal way. The criterion which is to be optimized must be specified. The function g assigns $x \in \mathcal{X}$ to the category l if and only if $g_l(x) > \max_{k \neq l} g_k(x)$. In case of ex æquo, x is assigned to a dummy category denoted by $*$. Let f be the decision function (from $\mathcal{X}$ into $\mathcal{Y} \bigcup \{*\}$) associated with g. With these definitions at hand, the objective function to be minimized is the probability of error $P(f(X) \neq Y)$. The optimization process, called *training*, is based on empirical data. More precisely, we assume that there exists a random pair $(X, Y) \in \mathcal{X} \times \mathcal{Y}$, distributed according to P, and we are provided with a m-sample $D_m = ((X_i, Y_i))_{1 \leq i \leq m}$ of independent copies of (X, Y).

There are two questions raised by such problems: how to properly choose the class of functions $\mathcal{G}$ and how to determine the best candidate g^* in this class, using only D_m. This report focuses on the first question, named *model selection*, in the particular case when the model considered is a M-SVM. The second question, named *function selection*, is addressed for instance in [11].

2.2 Architecture and training algorithms

M-SVMs, like all the SVMs, belong to the family of *kernel machines* [17]. As such, they operate on a class of functions induced by a positive semidefinite function/kernel. This calls for the formulation of some definitions and basic results. For the sake of simplicity, we consider real-valued functions only, although the general form of these definitions and results involves complex-valued functions.

Definition 1 (Positive semidefinite (positive type) function) *A real-valued function κ on $\mathcal{X}^2$ is called a* positive semidefinite function *(or a* positive type function*) if it is symmetric and*

$$\forall n \in \mathbb{N}^*,\ \forall \left(x_i\right)_{1 \leq i \leq n} \in \mathcal{X}^n,\ \forall \left(a_i\right)_{1 \leq i \leq n} \in \mathbb{R}^n,\ \sum_{i=1}^{n} \sum_{j=1}^{n} a_i a_j \kappa \left(x_i, x_j\right) \geq 0.$$

Definition 2 (Reproducing kernel Hilbert space [2]) *Let $(\mathbf{H}, \langle \cdot, \cdot \rangle_{\mathbf{H}})$ be a Hilbert space of real-valued functions on $\mathcal{X}$. A real-valued function κ on $\mathcal{X}^2$ is a* reproducing kernel *of $\mathbf{H}$ if and only if*

1. *$\forall x \in \mathcal{X},\ \kappa_x = \kappa\left(x, \cdot\right) \in \mathbf{H}$;*
2. *$\forall x \in \mathcal{X}, \forall h \in \mathbf{H},\ \langle h, \kappa_x \rangle_{\mathbf{H}} = h(x)$ (reproducing property).*

A Hilbert space of real-valued functions which possesses a reproducing kernel is called a reproducing kernel Hilbert space *(RKHS) or a* proper Hilbert space*.*

The connection between positive semidefinite functions and RKHSs is provided by the Moore-Aronszajn theorem.

Theorem 1 (Moore-Aronszajn theorem [1]) *Let κ be a real-valued positive semidefinite function on $\mathcal{X}^2$. There exists only one Hilbert space $(\mathbf{H}, \langle \cdot, \cdot \rangle_{\mathbf{H}})$ of real-valued functions on $\mathcal{X}$ with κ as reproducing kernel. The subspace $\mathbf{H}_0$ of $\mathbf{H}$ spanned by the functions κ_x is dense in $\mathbf{H}$ and $\mathbf{H}$ is the set of functions on $\mathcal{X}$ which are pointwise limits of Cauchy sequences in $\mathbf{H}_0$ with the inner product*

$$\langle h, h' \rangle_{\mathbf{H}_0} = \sum_{i=1}^{n} \sum_{j=1}^{n'} a_i a'_j \kappa \left(x_i, x'_j\right)$$

where $h = \sum_{i=1}^{n} a_i \kappa_{x_i}$ and $h' = \sum_{j=1}^{n'} a'_j \kappa_{x'_j}$.

Proposition 1 *Let κ be a real-valued positive semidefinite function on $\mathcal{X}^2$. There exists a map Φ from $\mathcal{X}$ into a Hilbert space $\left(E_{\Phi(\mathcal{X})}, \langle \cdot, \cdot \rangle\right)$ such that:*

$$\forall (x, x') \in \mathcal{X}^2,\ \kappa(x, x') = \langle \Phi(x), \Phi(x') \rangle. \tag{1}$$

In the sequel, such a map Φ will be called a *feature map* and $E_{\Phi(\mathcal{X})}$ a *feature space*. Taking advantage of the fact that the value of the inner product is the same in all the feature spaces (since it only depends on the choice of the kernel κ), we will also make the slight abuse of language consisting in calling Φ *the feature map* and $E_{\Phi(\mathcal{X})}$ *the feature space.* Let κ be a real-valued positive semidefinite kernel on $\mathcal{X}^2$ and let $(\mathbf{H}_\kappa, \langle \cdot, \cdot \rangle_{\mathbf{H}_\kappa})$ be the RKHS spanned by κ. Let $\bar{\mathcal{H}} = (\mathbf{H}_\kappa, \langle \cdot, \cdot \rangle_{\mathbf{H}_\kappa})^Q$ and let $\mathcal{H} = ((\mathbf{H}_\kappa, \langle \cdot, \cdot \rangle_{\mathbf{H}_\kappa}) + \{1\})^Q$. By construction, $\mathcal{H}$ is the class of vector-valued functions $h = \left(h_k\right)_{1 \leq k \leq Q}$ on $\mathcal{X}$ such that their component functions are finite affine combinations of the form

$$h_k(\cdot) = \sum_{i=1}^{m_k} \beta_{ik} \kappa \left(x_{ik}, \cdot\right) + b_k$$

where the x_{ik} are elements of $\mathcal{X}$ (the β_{ik} and b_k are scalars), as well as the limits of these functions as the sets $\{x_{ik} : 1 \leq i \leq m_k\}$ become dense in $\mathcal{X}$, in the norm induced by the inner product $\langle \cdot, \cdot \rangle_{\mathbf{H}_\kappa}$ (see also [22]). Due to Equation 1, $\mathcal{H}$ can also be seen as a multivariate affine model on $\Phi(\mathcal{X})$. Functions h can then be rewritten as

$$h(\cdot) = (\langle w_k, \cdot \rangle + b_k)_{1 \leq k \leq Q}$$

where the vectors w_k are elements of $E_{\Phi(\mathcal{X})}$. They are thus described by the pair $(\mathbf{w}, \mathbf{b})$ with $\mathbf{w} = (w_k)_{1 \leq k \leq Q} \in E_{\Phi(\mathcal{X})}^Q$ and $\mathbf{b} = (b_k)_{1 \leq k \leq Q} \in \mathbb{R}^Q$. As a consequence, $\bar{\mathcal{H}}$ can be seen as a multivariate linear model on $\Phi(\mathcal{X})$, endowed with a norm $\| \cdot \|_{\bar{\mathcal{H}}}$ given by:

$$\forall \bar{h} \in \bar{\mathcal{H}}, \ \left\| \bar{h} \right\|_{\bar{\mathcal{H}}} = \sqrt{\sum_{k=1}^{Q} \|w_k\|^2} = \|\mathbf{w}\|,$$

where $\|w_k\| = \sqrt{\langle w_k, w_k \rangle}$. With these definitions, theorems and propositions at hand, a generic definition of the M-SVMs can be formulated as follows.

Definition 3 (M-SVM, Definition 42 in [11]) *Let* $((x_i, y_i))_{1 \leq i \leq m} \in (\mathcal{X} \times [\![1, Q]\!])^m$ *and* $\lambda \in \mathbb{R}_+^*$. *A* Q-category M-SVM *is a large margin discriminant model obtained by minimizing over the hyperplane* $\sum_{k=1}^{Q} h_k = 0$ *of* $\mathcal{H}$ *a penalized risk* $J_{M\text{-}SVM}$ *of the form:*

$$J_{M\text{-}SVM}(h) = \sum_{i=1}^{m} \ell_{M\text{-}SVM}(y_i, h(x_i)) + \lambda \left\| \bar{h} \right\|_{\bar{\mathcal{H}}}^2$$

where the data fit component involves a loss function $\ell_{M\text{-}SVM}$ *which is convex.*

Definition 4 (Hard and soft margin M-SVM) *If a M-SVM is trained subject to the constraint that the data fit component is null (*$\sum_{i=1}^{m} \ell_{M\text{-}SVM}(y_i, h(x_i)) = 0$*), it is called a* hard margin *M-SVM. Otherwise, it is called a* soft margin *M-SVM.*

Three main models of M-SVMs can be found in literature (see [10] for a survey). The oldest one is the model of Weston and Watkins [23, 20, 4], which corresponds to the loss function ℓ_{WW} given by:

$$\ell_{\mathrm{WW}}(y, h(x)) = \sum_{k \neq y} (1 - h_y(x) + h_k(x))_+,$$

where the *hinge loss* function $(\cdot)_+$ is the function $\max(0, \cdot)$. The second one is due to Crammer and Singer [7] and corresponds to the loss function ℓ_{CS} given by:

$$\ell_{\mathrm{CS}}(y, \bar{h}(x)) = \left(1 - \bar{h}_y(x) + \max_{k \neq y} \bar{h}_k(x)\right)_+.$$

The most recent model is the one of Lee, Lin and Wahba [14], which corresponds to the loss function ℓ_{LLW} given by:

$$\ell_{\mathrm{LLW}}(y, h(x)) = \sum_{k \neq y} \left(h_k(x) + \frac{1}{Q-1} \right)_+. \tag{2}$$

Among the three models, the M-SVM of Lee, Lin and Wahba is the only one that implements asymptotically the Bayes decision rule. It is *Fisher consistent* [24, 19].

2.3 Geometrical margins

From a geometrical point of view, the algorithms described above select functions h^* (sets of the form $\{(w_k^*, b_k^*) : 1 \leq k \leq Q\}$) associated with sets of separating hyperplanes that tend to maximize globally the $\binom{Q}{2}$ *margins* between the different categories. If these margins are defined as in the bi-class case, their analytical expression is more complex.

Definition 5 (Geometrical margins, Definition 7 in [10]) *Let n be a positive integer and let $d_n = \{(x_i, y_i) : 1 \leq i \leq n\}$ be a set of n examples (belonging to $\mathcal{X} \times \mathcal{Y}$). If a function h in $\mathcal{H}$ classifies these examples without error, then its* margin between categories k and l *(computed with respect to d_n), $\gamma_{kl}(h)$, is defined as the smallest distance of a point of d_n either in k or l to the hyperplane separating those categories. Let us denote*

$$d(h) = \min_{1 \leq k < l \leq Q} \left\{ \min \left[\min_{i : y_i = k} \left(h_k(x_i) - h_l(x_i) \right), \min_{j : y_j = l} \left(h_l(x_j) - h_k(x_j) \right) \right] \right\}$$

and for $1 \leq k < l \leq Q$, let $d_{kl}(h)$ be

$$d_{kl}(h) = \frac{1}{d(h)} \min \left[\min_{i : y_i = k} \left(h_k(x_i) - h_l(x_i) \right), \min_{j : y_j = l} \left(h_l(x_j) - h_k(x_j) \right) \right] - 1.$$

Then we have

$$\gamma_{kl}(h) = d(h) \frac{1 + d_{kl}(h)}{\|w_k - w_l\|}.$$

Remark 1 *By definition, if $h \in \mathcal{H}$ classifies the examples of d_n without error, then*

$$\min_{1 \leq k < l \leq Q} d_{kl}(h) = 0.$$

However, for the hard margin versions of the three main models of M-SVMs, the assumption that all the values of the parameters $d_{kl}(h^*)$ are equal to 0 cannot be made a priori.

In the case of the M-SVMs (satisfying $\sum_{k=1}^{Q} w_k = 0$), the connection between the geometrical margins and the penalizer of $J_{\text{M-SVM}}$ is given by the following equation:

$$\sum_{k<l} \|w_k - w_l\|^2 = Q \sum_{k=1}^{Q} \|w_k\|^2, \tag{3}$$

the proof of which can for instance be found in Chapter 2 of [10].

3 The M-SVM of Lee, Lin and Wahba (LLW-M-SVM)

We now present in more detail the M-SVM for which the generalized radius-margin bound will be established.

3.1 Training algorithms

The substitution in Definition 3 of $\ell_{\text{M-SVM}}$ with the expression of the loss function ℓ_{LLW} given by Equation 2 provides us with the expressions of the quadratic programming (QP) problems corresponding to the training algorithms of the hard margin and soft margin versions of the M-SVM of Lee, Lin and Wahba.

Problem 1 (Hard margin M-SVM, primal formulation)

$$\min_{\mathbf{w}, \mathbf{b}} J_{HM}(\mathbf{w}, \mathbf{b})$$

$$s.t. \begin{cases} \forall i \in [\![1, m]\!], \forall k \in [\![1, Q]\!] \setminus \{y_i\}, \quad \langle w_k, \Phi(x_i) \rangle + b_k \leq -\frac{1}{Q-1} \\ \sum_{k=1}^{Q} w_k = 0 \\ \sum_{k=1}^{Q} b_k = 0 \end{cases}$$

where

$$J_{HM}(\mathbf{w}, \mathbf{b}) = \frac{1}{2} \sum_{k=1}^{Q} \|w_k\|^2.$$

Problem 2 (Soft margin M-SVM, primal formulation)

$$\min_{\mathbf{w},\mathbf{b},\xi} J_{SM}(\mathbf{w},\mathbf{b},\xi)$$

$$s.t. \begin{cases} \forall i \in [\![1,m]\!], \forall k \in [\![1,Q]\!] \setminus \{y_i\}, & \langle w_k, \Phi(x_i)\rangle + b_k \leq -\frac{1}{Q-1} + \xi_{ik} \\ \forall i \in [\![1,m]\!], \forall k \in [\![1,Q]\!] \setminus \{y_i\}, & \xi_{ik} \geq 0 \\ \sum_{k=1}^{Q} w_k = 0 \\ \sum_{k=1}^{Q} b_k = 0 \end{cases}$$

where

$$J_{SM}(\mathbf{w},\mathbf{b},\xi) = \frac{1}{2}\sum_{k=1}^{Q} \|w_k\|^2 + C\sum_{i=1}^{m}\sum_{k \neq y_i} \xi_{ik}.$$

In Problem 2, the ξ_{ik} are *slack variables* introduced in order to relax the constraints of correct classification. For convenience of notation, the vector ξ of these variables is represented as follows : $\xi = (\xi_{ik})_{1 \leq i \leq m, 1 \leq k \leq Q} \in \mathbb{R}_+^{Qm}$. ξ_{ik} is thus its component of index $(i-1)Q+k$ and the ξ_{iy_i} are dummy variables, all equal to 0. Using the notation e_n to designate the vector of $\mathbb{R}^n$ such that all its components are equal to e, we have thus $(\xi_{iy_i})_{1 \leq i \leq m} = 0_m$. The coefficient C, which characterizes the trade-off between prediction accuracy on the training set and smoothness of the solution, can be expressed in terms of the regularization coefficient λ as follows: $C = (2\lambda)^{-1}$. It is called the *soft margin parameter*. Instead of directly solving Problems 1 and 2, one usually solves their Wolfe dual [8]. We now derive the dual problem of Problem 2. Giving the details of the implementation of the Lagrangian duality will provide us with partial results which will prove useful in the sequel.

Let $\alpha = (\alpha_{ik})_{1 \leq i \leq m, 1 \leq k \leq Q} \in \mathbb{R}_+^{Qm}$ be the vector of Lagrange multipliers associated with the constraints of good classification. Similarly, let $\beta = (\beta_{ik})_{1 \leq i \leq m, 1 \leq k \leq Q} \in \mathbb{R}_+^{Qm}$ be the vector of Lagrange multipliers associated with the constraints of nonnegativity of the slack variables. These vectors are built according to the same principle as vector ξ. Let $\gamma \in E_{\Phi(\mathcal{X})}$ be the Lagrange multiplier associated with the constraint $\sum_{k=1}^{Q} w_k = 0$ and $\delta \in \mathbb{R}$ the Lagrange multiplier associated with the constraint $\sum_{k=1}^{Q} b_k = 0$. The Lagrangian function of Problem 2 is given by:

$$L(\mathbf{w},\mathbf{b},\xi,\alpha,\beta,\gamma,\delta) =$$

$$\frac{1}{2}\sum_{k=1}^{Q} \|w_k\|^2 + C\sum_{i=1}^{m}\sum_{k=1}^{Q} \xi_{ik} + \sum_{i=1}^{m}\sum_{k=1}^{Q} \alpha_{ik}\left(\langle w_k, \Phi(x_i)\rangle + b_k + \frac{1}{Q-1} - \xi_{ik}\right) - \sum_{i=1}^{m}\sum_{k=1}^{Q} \beta_{ik}\xi_{ik}$$

$$-\langle \gamma, \sum_{k=1}^{Q} w_k \rangle - \delta\sum_{k=1}^{Q} b_k. \tag{4}$$

Setting the gradient of L with respect to w_k equal to the null vector provides us with Q alternative expressions for the optimal value of vector γ:

$$\forall k \in [\![1,Q]\!], \ \ \gamma^* = w_k^* + \sum_{i=1}^{m} \alpha_{ik}^* \Phi(x_i). \tag{5}$$

Since by hypothesis, $\sum_{k=1}^{Q} w_k^* = 0$, summing over the index k provides us with the expression of γ^* as a function of dual variables only:

$$\gamma^* = \frac{1}{Q}\sum_{i=1}^{m}\sum_{k=1}^{Q} \alpha_{ik}^* \Phi(x_i).$$

By substitution into (5), we get the expression of the vectors w_k at the optimum:

$$\forall k \in [\![1,Q]\!], \ \ w_k^* = \sum_{i=1}^{m}\sum_{l=1}^{Q} \left(\frac{1}{Q} - \delta_{k,l}\right) \alpha_{il}^* \Phi(x_i), \tag{6}$$

where $\delta_{k,l}$ is the Kronecker symbol.

Let us now set the gradient of L with respect to $\mathbf{b}$ equal to the null vector. We get

$$\forall k \in [\![1, Q]\!], \;\; \delta^* = \sum_{i=1}^{m} \alpha_{ik}^*$$

and thus

$$\forall k \in [\![1, Q]\!], \;\; \sum_{i=1}^{m} \sum_{l=1}^{Q} \left(\frac{1}{Q} - \delta_{k,l} \right) \alpha_{il}^* = 0. \tag{7}$$

Given the constraint $\sum_{k=1}^{Q} b_k = 0$,

$$\sum_{i=1}^{m} \sum_{k=1}^{Q} \alpha_{ik}^* b_k^* = \sum_{k=1}^{Q} b_k^* \sum_{i=1}^{m} \alpha_{ik}^* = \delta^* \sum_{k=1}^{Q} b_k^* = 0. \tag{8}$$

Setting the gradient of L with respect to ξ equal to the null vector gives:

$$\forall i \in [\![1, m]\!], \forall k \in [\![1, Q]\!] \setminus \{y_i\}, \;\; \alpha_{ik}^* + \beta_{ik}^* = C. \tag{9}$$

By application of (6),

$$\sum_{k=1}^{Q} \|w_k^*\|^2 = \sum_{k=1}^{Q} \langle \sum_{i=1}^{m} \sum_{l=1}^{Q} \left(\frac{1}{Q} - \delta_{k,l} \right) \alpha_{il}^* \Phi(x_i), \sum_{j=1}^{m} \sum_{n=1}^{Q} \left(\frac{1}{Q} - \delta_{k,n} \right) \alpha_{jn}^* \Phi(x_j) \rangle$$

$$= \sum_{i=1}^{m} \sum_{j=1}^{m} \sum_{l=1}^{Q} \sum_{n=1}^{Q} \left\{ \sum_{k=1}^{Q} \left(\frac{1}{Q} - \delta_{k,l} \right) \left(\frac{1}{Q} - \delta_{k,n} \right) \right\} \alpha_{il}^* \alpha_{jn}^* \langle \Phi(x_i), \Phi(x_j) \rangle$$

$$= \sum_{i=1}^{m} \sum_{j=1}^{m} \sum_{l=1}^{Q} \sum_{n=1}^{Q} \left(\delta_{l,n} - \frac{1}{Q} \right) \alpha_{il}^* \alpha_{jn}^* \kappa(x_i, x_j). \tag{10}$$

Still by application of (6),

$$\sum_{i=1}^{m} \sum_{k=1}^{Q} \alpha_{ik}^* \langle w_k^*, \Phi(x_i) \rangle = \sum_{i=1}^{m} \sum_{k=1}^{Q} \alpha_{ik}^* \langle \sum_{j=1}^{m} \sum_{l=1}^{Q} \left(\frac{1}{Q} - \delta_{k,l} \right) \alpha_{jl}^* \Phi(x_j), \Phi(x_i) \rangle$$

$$= \sum_{i=1}^{m} \sum_{j=1}^{m} \sum_{k=1}^{Q} \sum_{l=1}^{Q} \left(\frac{1}{Q} - \delta_{k,l} \right) \alpha_{ik}^* \alpha_{jl}^* \kappa(x_i, x_j). \tag{11}$$

Combining (10) and (11) gives:

$$\frac{1}{2} \sum_{k=1}^{Q} \|w_k^*\|^2 + \sum_{i=1}^{m} \sum_{k=1}^{Q} \alpha_{ik}^* \langle w_k^*, \Phi(x_i) \rangle = -\frac{1}{2} \sum_{k=1}^{Q} \|w_k^*\|^2$$

$$= -\frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} \sum_{k=1}^{Q} \sum_{l=1}^{Q} \left(\delta_{k,l} - \frac{1}{Q} \right) \alpha_{ik}^* \alpha_{jl}^* \kappa(x_i, x_j). \tag{12}$$

Extending to the case of matrices the double subscript notation used to designate the general terms of the vectors α, β and ξ, let us define H as the matrix of $\mathcal{M}_{Qm,Qm}(\mathbb{R})$ of general term:

$$h_{ik,jl} = \left(\delta_{k,l} - \frac{1}{Q} \right) \kappa(x_i, x_j).$$

With these notations at hand, reporting (8), (9) and (12) in (4) provides us with an algebraic expression of the Lagrangian function at the optimum where the primal variables have been eliminated. This provides us in turn with the following expression for the objective function of the Wolfe dual of Problem 2:

$$J_{\mathrm{LLW,d}}(\alpha) = -\frac{1}{2}\alpha^T H \alpha + \frac{1}{Q-1} 1_{Qm}^T \alpha.$$

The constraints of this problem are derived from Equations 7 and 9. The Wolfe dual of Problem 2 is thus:

Problem 3 (Soft margin M-SVM, dual formulation)

$$\max_{\alpha} J_{LLW,d}(\alpha)$$

$$s.t. \begin{cases} \forall i \in [\![1, m]\!], \forall k \in [\![1, Q]\!] \setminus \{y_i\}, & 0 \leq \alpha_{ik} \leq C \\ \forall k \in [\![1, Q-1]\!], & \sum_{i=1}^m \sum_{l=1}^Q \left(\frac{1}{Q} - \delta_{k,l}\right) \alpha_{il} = 0 \end{cases}$$

where

$$J_{LLW,d}(\alpha) = -\frac{1}{2}\alpha^T H \alpha + \frac{1}{Q-1} 1_{Qm}^T \alpha,$$

with the general term of the Hessian matrix H being

$$h_{ik,jl} = \left(\delta_{k,l} - \frac{1}{Q}\right) \kappa(x_i, x_j).$$

With slight modifications, the derivation above can be adapted to express the Wolfe dual of Problem 1. This leads to:

Problem 4 (Hard margin M-SVM, dual formulation)

$$\max_{\alpha} J_{LLW,d}(\alpha)$$

$$s.t. \begin{cases} \forall i \in [\![1, m]\!], \forall k \in [\![1, Q]\!] \setminus \{y_i\}, & \alpha_{ik} \geq 0 \\ \forall k \in [\![1, Q-1]\!], & \sum_{i=1}^m \sum_{l=1}^Q \left(\frac{1}{Q} - \delta_{k,l}\right) \alpha_{il} = 0 \end{cases} .$$

3.2 Geometrical margins

The geometrical margins of the hard margin Q-category M-SVM of Lee, Lin and Wahba can be characterized thanks to three propositions among which the two last will prove useful to establish the radius-margin bound.

Proposition 2 *Let us consider a hard margin Q-category M-SVM of Lee, Lin and Wahba. Then,*

$$d(h^*) \geq \frac{Q}{Q-1}.$$

Proof First, note that if $h \in \mathcal{H}$ classifies the examples of the set $\{(x_i, y_i) : 1 \leq i \leq n\}$ without error, then $d(h) = \min_{1 \leq i \leq n} \min_{k \neq y_i} (h_{y_i}(x_i) - h_k(x_i))$. By application of the formula giving ℓ_{LLW},

$$\forall i \in [\![1, m]\!], \forall k \in [\![1, Q]\!] \setminus \{y_i\}, \ \ h_k^*(x_i) \leq -\frac{1}{Q-1}.$$

Since $\sum_{k=1}^Q h_k^* = 0$, this implies that

$$\forall i \in [\![1, m]\!], \ \ h_{y_i}^*(x_i) \geq 1$$

and thus $d(h^*) \geq \frac{Q}{Q-1}$. ■

Proposition 3 *For the hard margin Q-category M-SVM of Lee, Lin and Wahba trained on* $\{(x_i, y_i) : 1 \leq i \leq m\}$*, in the non-trivial case when* $\alpha^* \neq 0$*, there exists a mapping* $\mathcal{I}$ *from* $[\![1, Q]\!]$ *to* $[\![1, m]\!]$ *such that*

$$\forall k \in [\![1, Q]\!], \;\; h_k^* \left(x_{\mathcal{I}(k)} \right) = -\frac{1}{Q-1}.$$

Proof This proposition results readily from the Kuhn-Tucker optimality conditions and the form taken by the constraints of Problem 4. Indeed, if $\alpha^* \neq 0$, then for all k, there exists at least one dual variable α_{ik}^* which is positive. ■

Proposition 4 *For the hard margin Q-category M-SVM of Lee, Lin and Wahba, we have*

$$\frac{d\left(h^*\right)^2}{Q} \sum_{k<l} \left(\frac{1 + d_{kl}\left(h^*\right)}{\gamma_{kl}\left(h^*\right)} \right)^2 = \sum_{k=1}^{Q} \|w_k^*\|^2 = {\alpha^*}^T H \alpha^* = \frac{1}{Q-1} 1_{Qm}^T \alpha^*.$$

Proof

- $\frac{d(h^*)^2}{Q} \sum_{k<l} \left(\frac{1+d_{kl}(h^*)}{\gamma_{kl}(h^*)} \right)^2 = \sum_{k=1}^{Q} \|w_k^*\|^2$

  This equation is a direct consequence of Definition 5 and Equation 3.

- $\sum_{k=1}^{Q} \|w_k^*\|^2 = {\alpha^*}^T H \alpha^*$

  This is a direct consequence of Equation 12 and the definition of matrix H.

- ${\alpha^*}^T H \alpha^* = \frac{1}{Q-1} 1_{Qm}^T \alpha^*$

  The general term of the gradient $\nabla J_{\text{LLW,d}}\left(\alpha^*\right) = -H\alpha^* + \frac{1}{Q-1} 1_{Qm}$ is $\langle w_k^*, \Phi(x_i) \rangle + \frac{1}{Q-1}$. Thus, the Kuhn-Tucker optimality conditions imply that

  $$\sum_{i=1}^{m} \sum_{k=1}^{Q} \alpha_{ik}^* \left(\langle w_k^*, \Phi(x_i) \rangle + b_k^* + \frac{1}{Q-1} \right) = -{\alpha^*}^T H \alpha^* + \frac{1}{Q-1} 1_{Qm}^T \alpha^* + \sum_{i=1}^{m} \sum_{k=1}^{Q} \alpha_{ik}^* b_k^* = 0.$$

  By application of Equation 8, the right-hand side of this equation simplifies into ${\alpha^*}^T H \alpha^* = \frac{1}{Q-1} 1_{Qm}^T \alpha^*$.

■

4 Multi-Class Radius-Margin Bound on the Leave-One-Out Error of the M-SVM

Like its bi-class counterpart, our multi-class radius-margin bound is based on a key lemma.

4.1 Multi-class key lemma

Lemma 1 (Multi-class key lemma) *Let us consider a hard margin Q-category M-SVM of Lee, Lin and Wahba on a domain* $\mathcal{X}$*. Let* $d_m = \{(x_i, y_i) : 1 \leq i \leq m\}$ *be its training set. Consider now the same machine trained on* $d_m \setminus \{(x_p, y_p)\}$*. If it makes an error on* (x_p, y_p)*, then the inequality*

$$\max_{k \in [\![1, Q]\!]} \alpha_{pk}^* \geq \frac{1}{Q(Q-1)\mathcal{D}_m^2}$$

holds, where $\mathcal{D}_m$ *is the diameter of the smallest sphere of the feature space containing the set* $\{\Phi(x_i) : 1 \leq i \leq m\}$*.*

Proof Let $h^p \in \mathcal{H}$ be the optimal solution when the machine is trained on $d_m \setminus \{(x_p, y_p)\}$. Accordingly, let us denote by $(\mathbf{w}^p, \mathbf{b}^p)$ the couple characterizing the optimal hyperplanes and by $\alpha^p = (\alpha^p_{ik}) \in \mathbb{R}_+^{Qm}$ the corresponding vector of the dual variables, with $\left(\alpha^p_{pk}\right)_{1 \leq k \leq Q} = 0_Q$. This representation is used in order to simplify the simultaneous handling of both M-SVMs. Indeed, α^p is an optimal solution of Problem 4 under the additional constraint $(\alpha_{pk})_{1 \leq k \leq Q} = 0_Q$. Let us define two more vectors in $\mathbb{R}_+^{Qm}$, $\lambda^p = (\lambda^p_{ik})_{1 \leq i \leq m, 1 \leq k \leq Q}$ and $\mu^p = (\mu^p_{ik})_{1 \leq i \leq m, 1 \leq k \leq Q}$. λ^p exhibits additional properties so that the vector $\alpha^* - \lambda^p$ is a feasible solution of Problem 4 under the additional constraint that $\left(\alpha^*_{pk} - \lambda^p_{pk}\right)_{1 \leq k \leq Q} = 0_Q$, i.e., $\alpha^* - \lambda^p$ satisfies the same constraints as α^p. We have thus

$$\forall i \in [\![1, m]\!] \setminus \{p\}, \forall k \in [\![1, Q]\!] \setminus \{y_i\}, \;\; \alpha^*_{ik} - \lambda^p_{ik} \geq 0 \Longleftrightarrow \lambda^p_{ik} \leq \alpha^*_{ik}.$$

We deduce from the equality constraints of Problem 4 that:

$$\forall k \in [\![1, Q]\!], \;\; \sum_{i=1}^m \sum_{l=1}^Q \left(\frac{1}{Q} - \delta_{k,l} \right) (\alpha^*_{il} - \lambda^p_{il}) = 0 \Longleftrightarrow \sum_{i=1}^m \sum_{l=1}^Q \left(\frac{1}{Q} - \delta_{k,l} \right) \lambda^p_{il} = 0.$$

To sum up, vector λ^p satisfies the following constraints:

$$\begin{cases} \forall k \in [\![1, Q]\!], \;\; \lambda^p_{pk} = \alpha^*_{pk} \\ \forall i \in [\![1, m]\!] \setminus \{p\}, \forall k \in [\![1, Q]\!] \setminus \{y_i\}, \;\; 0 \leq \lambda^p_{ik} \leq \alpha^*_{ik} \\ \forall k \in [\![1, Q-1]\!], \;\; \sum_{i=1}^m \sum_{l=1}^Q \left(\frac{1}{Q} - \delta_{k,l} \right) \lambda^p_{il} = 0 \end{cases} . \tag{13}$$

Note that the domain defined by these constraints is a subset of the feasible set of Problem 4 (vector λ^p is a feasible solution of Problem 4). The properties of vector μ^p are such that $\alpha^p + K_1 \mu^p$ satisfies the same constraints as α^*, where K_1 is a positive scalar the value of which will be specified in the sequel. We have thus:

$$\forall i \in [\![1, m]\!], \;\; \alpha^p_{iy_i} + K_1 \mu^p_{iy_i} = 0 \Longleftrightarrow \mu^p_{iy_i} = 0.$$

Moreover, we have

$$\forall i \in [\![1, m]\!], \forall k \in [\![1, Q]\!] \setminus \{y_i\}, \;\; \mu^p_{ik} \geq 0 \Longrightarrow \alpha^p_{ik} + K_1 \mu^p_{ik} \geq 0.$$

Finally,

$$\forall k \in [\![1, Q]\!], \;\; \sum_{i=1}^m \sum_{l=1}^Q \left(\frac{1}{Q} - \delta_{k,l} \right) (\alpha^p_{il} + K_1 \mu^p_{il}) = 0 \Longleftrightarrow \sum_{i=1}^m \sum_{l=1}^Q \left(\frac{1}{Q} - \delta_{k,l} \right) \mu^p_{il} = 0.$$

To sum up, vector μ^p is a feasible solution of Problem 4. In the sequel, for the sake of simplicity, we write J in place of $J_{\text{LLW,d}}$. By construction of vectors λ^p and μ^p, we have $J(\alpha^* - \lambda^p) \leq J(\alpha^p)$ and $J\left(\alpha^p + K_1 \mu^p\right) \leq J(\alpha^*)$. Hence,

$$J(\alpha^*) - J(\alpha^* - \lambda^p) \geq J(\alpha^*) - J(\alpha^p) \geq J\left(\alpha^p + K_1 \mu^p\right) - J(\alpha^p). \tag{14}$$

The expression of the first term is

$$J(\alpha^*) - J(\alpha^* - \lambda^p) = \frac{1}{2} \lambda^{pT} H \lambda^p + \nabla J(\alpha^*)^T \lambda^p.$$

Since α^* and λ^p are respectively an optimal and a feasible solution of Problem 4, then necessarily,

$$\nabla J(\alpha^*)^T \lambda^p \leq 0.$$

This becomes obvious when one thinks about the principle of the Frank-Wolfe algorithm [9]. As a consequence,

$$J(\alpha^*) - J(\alpha^* - \lambda^p) \leq \frac{1}{2} \lambda^{pT} H \lambda^p$$

and equivalently, in view of Equations 6 and 10 (where α^* has been replaced with λ^p), as well as the definition of H,

$$J(\alpha^*) - J(\alpha^* - \lambda^p) \leq \frac{1}{2} \sum_{k=1}^{Q} \left\| \sum_{i=1}^{m} \sum_{l=1}^{Q} \left(\frac{1}{Q} - \delta_{k,l} \right) \lambda_{il}^p \Phi(x_i) \right\|^2 . \tag{15}$$

We now turn to the right-hand side of (14). The line of reasoning already used for the left-hand side gives:

$$J\left(\alpha^p + K_1 \mu^p\right) - J(\alpha^p) = K_1 \nabla J(\alpha^p)^T \mu^p - \frac{K_1^2}{2} \sum_{k=1}^{Q} \left\| \sum_{i=1}^{m} \sum_{l=1}^{Q} \left(\frac{1}{Q} - \delta_{k,l} \right) \mu_{il}^p \Phi(x_i) \right\|^2 . \tag{16}$$

By hypothesis, the M-SVM trained on $d_m \backslash \{(x_p, y_p)\}$ does not classify x_p correctly. This means that there exists $n \in [\![1, Q]\!] \setminus \{y_p\}$ such that $h_n^p\left(x_p\right) \geq 0$. Furthermore, α^p is not an optimal solution of Problem 4. Since μ^p is a feasible solution of the same problem, it can be built in such a way that $\nabla J(\alpha^p)^T \mu^p > 0$ (it defines a direction of ascent). These observations being made, neglecting the case $\alpha^p = 0$ as a degenerate one, we apply Proposition 3 to build a vector μ^p with adequate properties. Thus, let $\mathcal{I}$ be a mapping from $[\![1, Q]\!]$ to $[\![1, m]\!] \setminus \{p\}$ such that

$$\forall k \in [\![1, Q]\!], \;\; h_k^p\left(x_{\mathcal{I}(k)}\right) = -\frac{1}{Q-1}.$$

For $K_2 \in \mathbb{R}_+^*$, let μ^p be the vector of $\mathbb{R}_+^{Qm}$ that only differs from the null vector in the following way:

$$\begin{cases} \mu_{pn}^p = K_2 \\ \forall k \in [\![1, Q]\!] \setminus \{n\}, \;\; \mu_{\mathcal{I}(k)k}^p = K_2 \end{cases} .$$

Obviously, this solution satisfies the constraints of Problem 4. With this definition of vector μ^p, the inner product $\nabla J(\alpha^p)^T \mu^p$ simplifies as follows:

$$\nabla J(\alpha^p)^T \mu^p = \sum_{i=1}^{m} \sum_{k=1}^{Q} \mu_{ik}^p \left(\langle w_k^p, \Phi(x_i) \rangle + \frac{1}{Q-1} \right)$$

$$= K_2 \left\{ \langle w_n^p, \Phi(x_p) \rangle + \frac{1}{Q-1} + \sum_{k \neq n} \left(\langle w_k^p, \Phi\left(x_{\mathcal{I}(k)}\right) \rangle + \frac{1}{Q-1} \right) \right\}$$

$$= K_2 \left\{ h_n^p(x_p) + \frac{1}{Q-1} - \sum_{k=1}^{Q} b_k^p \right\} .$$

As a consequence,

$$\nabla J(\alpha^p)^T \mu^p \geq \frac{K_2}{Q-1}.$$

By substitution into Equation 16, we get

$$J\left(\alpha^p + K_1 \mu^p\right) - J(\alpha^p) \geq \frac{K_1 K_2}{Q-1} - \frac{K_1^2}{2} \sum_{k=1}^{Q} \left\| \sum_{i=1}^{m} \sum_{l=1}^{Q} \left(\frac{1}{Q} - \delta_{k,l} \right) \mu_{il}^p \Phi(x_i) \right\|^2 . \tag{17}$$

Combining (14), (15) and (17) finally gives

$$\frac{1}{2} \sum_{k=1}^{Q} \left\| \sum_{i=1}^{m} \sum_{l=1}^{Q} \left(\frac{1}{Q} - \delta_{k,l} \right) \lambda_{il}^p \Phi(x_i) \right\|^2 \geq$$

$$\frac{K_1 K_2}{Q-1} - \frac{K_1^2}{2} \sum_{k=1}^{Q} \left\| \sum_{i=1}^{m} \sum_{l=1}^{Q} \left(\frac{1}{Q} - \delta_{k,l} \right) \mu_{il}^p \Phi(x_i) \right\|^2 . \tag{18}$$

Let $\nu^p = (\nu^p_{ik})_{1 \leq i \leq m, 1 \leq k \leq Q}$ be the vector of $\mathbb{R}_+^{Qm}$ such that $\mu^p = K_2 \nu^p$. The value of the scalar $K = K_1 K_2$ maximizing the right-hand side of (18) is:

$$K^* = \frac{\frac{1}{Q-1}}{\sum_{k=1}^{Q} \left\| \sum_{i=1}^{m} \sum_{l=1}^{Q} \left(\frac{1}{Q} - \delta_{k,l} \right) \nu^p_{il} \Phi(x_i) \right\|^2}.$$

By substitution in (18), this implies that

$$(Q-1)^2 \sum_{k=1}^{Q} \left\| \sum_{i=1}^{m} \sum_{l=1}^{Q} \left(\frac{1}{Q} - \delta_{k,l} \right) \lambda^p_{il} \Phi(x_i) \right\|^2 \sum_{k=1}^{Q} \left\| \sum_{i=1}^{m} \sum_{l=1}^{Q} \left(\frac{1}{Q} - \delta_{k,l} \right) \nu^p_{il} \Phi(x_i) \right\|^2 \geq 1.$$

For $\eta = (\eta_{ik})_{1 \leq i \leq m, 1 \leq k \leq Q} \in \mathbb{R}^{Qm}$, let $S(\eta) = \frac{1}{Q} \sum_{i=1}^{m} \sum_{k=1}^{Q} \eta^p_{ik}$. Given the equality constraints satisfied by vector λ^p, the quadratic form $\lambda^{pT} H \lambda^p$ can be rewritten as

$$\sum_{k=1}^{Q} \left\| \frac{1}{Q} \sum_{i=1}^{m} \sum_{l=1}^{Q} \lambda^p_{il} \Phi(x_i) - \sum_{i=1}^{m} \lambda^p_{ik} \Phi(x_i) \right\|^2 =$$

$$S(\lambda^p)^2 \sum_{k=1}^{Q} \left\| \operatorname{conv}\{\Phi(x_i) : 1 \leq i \leq m\} - \operatorname{conv}_k \{\Phi(x_i) : 1 \leq i \leq m\} \right\|^2$$

where $\operatorname{conv}\{\Phi(x_i) : 1 \leq i \leq m\}$ and the $\operatorname{conv}_k \{\Phi(x_i) : 1 \leq i \leq m\}$ are convex combinations of the $\Phi(x_i)$. As a consequence,

$$\forall k \in [\![1, Q]\!], \;\; \left\| \operatorname{conv}\{\Phi(x_i) : 1 \leq i \leq m\} - \operatorname{conv}_k \{\Phi(x_i) : 1 \leq i \leq m\} \right\|^2 \leq \mathcal{D}_m^2.$$

Since the same reasoning applies to ν^p, we get:

$$(Q-1)^2 Q^2 S(\lambda^p)^2 S(\nu^p)^2 \mathcal{D}_m^4 \geq 1. \tag{19}$$

By construction, $S(\nu^p) = 1$. We now construct a vector λ^p minimizing the objective function S. First, note that due to the equality constraints satisfied by this vector,

$$\forall (k,l) \in [\![1, Q]\!]^2, \;\; \sum_{i=1}^{m} \lambda^p_{ik} = \sum_{i=1}^{m} \lambda^p_{il}.$$

This implies that

$$\forall k \in [\![1, Q]\!], \;\; \sum_{i=1}^{m} \lambda^p_{ik} \geq \max_{l \in [\![1, Q]\!]} \alpha^*_{pl}$$

and thus

$$\min_{\lambda^p} S(\lambda^p) \geq \max_{l \in [\![1, Q]\!]} \alpha^*_{pl}.$$

Obviously, the nature of the function S calls for the choice of minimal values for the components λ^p_{ik}, which is coherent with the box constraints in (13). Thus, there exists a vector λ^{p^*} which is a minimizer of S subject to the set of constraints (13) such that

$$\forall k \in [\![1, Q]\!], \;\; \sum_{i=1}^{m} \lambda^{p^*}_{ik} = \max_{l \in [\![1, Q]\!]} \alpha^*_{pl},$$

i.e., $S\left(\lambda^{p^*}\right) = \max_{l \in [\![1, Q]\!]} \alpha^*_{pl}$. The substitution of the values of $S(\nu^p)$ and $S\left(\lambda^{p^*}\right)$ in (19) provides us with

$$\left(\max_{k \in [\![1, Q]\!]} \alpha^*_{pk} \right)^2 \geq \frac{1}{(Q-1)^2 Q^2 \mathcal{D}_m^4}.$$

Taking the square root of both sides concludes the proof of the lemma. ■

4.2 Multi-class radius-margin bound

The multi-class radius-margin bound is a direct consequence of Lemma 1.

Theorem 2 (Multi-class radius-margin bound) *Let us consider a hard margin Q-category M-SVM of Lee, Lin and Wahba on a domain $\mathcal{X}$. Let $d_m = \{(x_i, y_i) : 1 \leq i \leq m\}$ be its training set, $\mathcal{L}_m$ the number of errors resulting from applying a leave-one-out cross-validation procedure to this machine, and $\mathcal{D}_m$ the diameter of the smallest sphere of the feature space containing the set $\{\Phi(x_i) : 1 \leq i \leq m\}$. Then, using the notations of Definition 5, the following upper bound holds true:*

$$\mathcal{L}_m \leq (Q-1)^2 \mathcal{D}_m^2 d\left(h^*\right)^2 \sum_{k<l} \left(\frac{1 + d_{kl}\left(h^*\right)}{\gamma_{kl}\left(h^*\right)} \right)^2 .$$

Proof Lemma 1 exhibits a non-trivial lower bound on $\max_{k \in [\![1,Q]\!]} \alpha^*_{pk}$ when the machine trained on the set $d_m \setminus \{(x_p, y_p)\}$ makes an error on (x_p, y_p), i.e., when (x_p, y_p) contributes to $\mathcal{L}_m$. As a consequence,

$$1_{Qm}^T \alpha^* \geq \sum_{i=1}^{m} \max_{k \in [\![1,Q]\!]} \alpha^*_{ik} \geq \frac{\mathcal{L}_m}{Q(Q-1)\mathcal{D}_m^2}. \tag{20}$$

According to Proposition 4,

$$1_{Qm}^T \alpha^* = \frac{Q-1}{Q} d\left(h^*\right)^2 \sum_{k<l} \left(\frac{1 + d_{kl}\left(h^*\right)}{\gamma_{kl}\left(h^*\right)} \right)^2 .$$

A substitution in (20) thus provides us with the announced result. ■

5 Conclusions and Ongoing Research

In this report, we have established a generalization of Vapnik's radius-margin bound dedicated to the (hard margin) M-SVM of Lee, Lin and Wahba. In doing so, we have highlighted different features of the M-SVMs which make their study intrinsically more difficult than the one of bi-class pattern recognition SVMs. For instance, the formula expressing the geometrical margins as a function of the vector of dual variables α^* (Proposition 4) is far more complicated than its bi-class counterpart. This work, which comes after our Vapnik-Chervonenkis theory of the large margin multi-category classifiers [11] and our characterization of the Rademacher complexity of the M-SVMs [12], thus provides us with new arguments suggesting that the study of multi-category classification should be tackled independently of the one of dichotomy computation.

An open question of central importance is the possibility to use our bound to set the value of the soft margin parameter C. This question can be reformulated as follows: is there a variant of the soft margin M-SVM of Lee, Lin and Wahba such that its training algorithm is equivalent to the training algorithm of a hard margin machine obtained by a simple change of kernel? In the bi-class case, it is well known that the answer is positive, and the corresponding variant is the 2-norm SVM (see for instance Chapter 7 in [18]). Finding the answer in the multi-class case is the subject of an ongoing research, as well as the derivation of radius-margin bounds suitable for the two other M-SVMs.

Acknowledgments

The work of E. Monfrini was supported by the Decrypthon program of the "Association Française contre les Myopathies" (AFM), the CNRS and IBM. The authors would like to thank M. Bertrand for carefully reading this manuscript.

References

[1] N. Aronszajn. Theory of reproducing kernels. *Transactions of the American Mathematical Society*, 68(3):337–404, 1950.

[2] A. Berlinet and C. Thomas-Agnan. *Reproducing Kernel Hilbert Spaces in Probability and Statistics.* Kluwer Academic Publishers, Boston, 2004.

[3] B.E. Boser, I.M. Guyon, and V.N. Vapnik. A training algorithm for optimal margin classifiers. In *COLT'92*, pages 144–152, 1992.

[4] E.J. Bredensteiner and K.P. Bennett. Multicategory classification by support vector machines. *Computational Optimization and Applications*, 12(1/3):53–79, 1999.

[5] O. Chapelle, V.N. Vapnik, O. Bousquet, and S. Mukherjee. Choosing multiple parameters for support vector machines. *Machine Learning*, 46(1):131–159, 2002.

[6] C. Cortes and V.N. Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, 1995.

[7] K. Crammer and Y. Singer. On the algorithmic implementation of multiclass kernel-based vector machines. *Journal of Machine Learning Research*, 2:265–292, 2001.

[8] R. Fletcher. *Practical Methods of Optimization.* John Wiley & Sons, Chichester, second edition, 1987.

[9] M. Frank and P. Wolfe. An algorithm for quadratic programming. *Naval Research Logistics Quarterly*, 3:95–110, 1956.

[10] Y. Guermeur. *SVM multiclasses, théorie et applications.* Habilitation à diriger des recherches, UHP, 2007. (in French).

[11] Y. Guermeur. VC theory of large margin multi-category classifiers. *Journal of Machine Learning Research*, 8:2551–2594, 2007.

[12] Y. Guermeur. Sample complexity of classifiers taking values in $\mathbb{R}^Q$, application to multi-class SVMs. *Communications in Statistics*, 2009. (to appear).

[13] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning - Data Mining, Inference, and Prediction.* Springer, New York, 2001.

[14] Y. Lee, Y. Lin, and G. Wahba. Multicategory support vector machines: Theory and application to the classification of microarray data and satellite radiance data. *Journal of the American Statistical Association*, 99(465):67–81, 2004.

[15] A. Luntz and V. Brailovsky. On estimation of characters obtained in statistical procedure of recognition. *Technicheskaya Kibernetica*, 3, 1969. (in Russian).

[16] P. Massart. Concentrations inequalities and model selection. In *Ecole d'Eté de Probabilités de Saint-Flour XXXIII*, LNM. Springer-Verlag, 2003.

[17] B. Schölkopf and A.J. Smola. *Learning with Kernels - Support Vector Machines, Regularization, Optimization, and Beyond.* The MIT Press, Cambridge, MA, 2002.

[18] J. Shawe-Taylor and N. Cristianini. *Kernel Methods for Pattern Analysis.* Cambridge University Press, Cambridge, 2004.

[19] A. Tewari and P.L. Bartlett. On the consistency of multiclass classification methods. *Journal of Machine Learning Research*, 8:1007–1025, 2007.

[20] V.N. Vapnik. *Statistical Learning Theory.* John Wiley & Sons, Inc., New York, 1998.

[21] V.N. Vapnik and O. Chapelle. Bounds on error expectation for support vector machines. *Neural Computation*, 12(9):2013–2036, 2000.

[22] G. Wahba. Support vector machines, reproducing kernel Hilbert spaces, and randomized GACV. In B. Schölkopf, C.J.C. Burges, and A.J. Smola, editors, *Advances in Kernel Methods, Support Vector Learning*, chapter 6, pages 69–88. The MIT Press, Cambridge, MA, 1999.

[23] J. Weston and C. Watkins. Multi-class support vector machines. Technical Report CSD-TR-98-04, Royal Holloway, University of London, Department of Computer Science, 1998.

[24] T. Zhang. Statistical analysis of some multi-category large margin classification methods. *Journal of Machine Learning Research*, 5:1225–1251, 2004.