A Quadratic Loss Multi-Class SVM for which a Radius-Margin Bound Applies

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

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

September 15, 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 requirements in terms of computational time. 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. It applies to the hard margin machine, and, by extension, to the 2-norm SVM. In this article, we introduce a variant of the multi-class SVM of Lee, Lin and Wahba: the M-SVM^2. This quadratic loss machine can be seen as a direct extension of the 2-norm SVM to the multi-class case. For this machine, a generalized radius-margin bound is then established.

Keywords: Multi-class SVMs, model selection, leave-one-out cross-validation error, radius-margin bounds

1 Introduction

Using a SVM [3, 7] 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 [18, 24]). 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 [23]. 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* [30]. However, the results of Chapelle and co-workers presented in [5] show that another bound, the *radius-margin* one [29], achieves equivalent performance for model selection while being far simpler to compute. These results are corroborated by those of several comparative studies, among which [9]. As a consequence, this bound, which applies to the hard margin machine and, by extension, to the 2-norm SVM (see for instance Chapter 7 in [27]), is currently the most popular one. Several variants have been proposed, for instance in [6]. During the last few years, several multi-class SVMs (M-SVMs) have been introduced by different teams (see [13] for a survey). However, to the best of our knowledge, literature only proposes a single multi-class extension of the radius-margin bound. This bound, introduced in [32, 33], makes use of the bi-class bound in the framework of the one-versus-one decomposition method. As such, it does not represent a direct generalization of the initial result to a M-SVM, and the authors state that "such a theoretical generalization of this bound is not that straightforward because this bound is rooted in the theoretical basis of binary SVMs."

In this article, a new multi-class SVM is introduced: the M-SVM^2. It can be seen either as a quadratic loss variant of the M-SVM of Lee, Lin and Wahba (LLW-M-SVM) [22] or as a multi-class extension of the 2-norm SVM. A generalized radius-margin bound on the leave-one-out error of the hard margin version of the LLW-M-SVM is then established and assessed. This provides us with a differentiable objective function to perform model selection for the M-SVM^2.

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 and Section 4 introduces the M-SVM^2. Section 5 is devoted to the formulation, proof and analysis of the corresponding multi-class radius-margin bound. At last, we draw conclusions and outline our ongoing research in Section 6.

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 [29].

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 $3 \leqslant Q < \infty$, 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 indices of the categories: $[\![1, Q]\!]$. We assume that the link between descriptions and categories can be described by an unknown probability measure P on the product space $\mathcal{X} \times \mathcal{Y}$. The learning problem then consists in selecting in a set $\mathcal{G}$ of functions $g = (g_k)_{1 \leqslant k \leqslant Q}$ 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, ideally, the objective function to be minimized over $\mathcal{G}$ is the probability of error $P\left(f\left(X\right) \neq Y\right)$. In practice, since P is unknown, other criteria are used and the optimization process, called *training*, is based on empirical data. More precisely, we assume that there exists a random pair (couple) $(X, Y) \in \mathcal{X} \times \mathcal{Y}$ distributed according to P, and we are provided with a m-sample $D_m = \left(\left(X_i, Y_i\right)\right)_{1 \leqslant i \leqslant m}$ of independent copies of (X, Y). Those copies form the *training set*.

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 article 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 [14].

2.2 Architecture and training algorithms

M-SVMs, like all the SVMs, belong to the family of *kernel machines* [26]. 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*

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

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 corresponding RKHS. 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 $h = (h_k)_{1 \leqslant k \leqslant Q}$ 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 $\{x_{ik} : 1 \leqslant i \leqslant m_k\}$ become dense in $\mathcal{X}$, in the norm induced by the inner product $\langle \cdot, \cdot \rangle_{\mathbf{H}_\kappa}$ (see also [31]). Due to Equation 1, $\mathcal{H}$ can alternatively be seen as a multivariate affine model on $\Phi\left(\mathcal{X}\right)$. Functions h can then be rewritten as

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

where the vectors w_k are elements of $E_{\Phi(\mathcal{X})}$. They are thus described by the pair $(\mathbf{w}, \mathbf{b})$ with $\mathbf{w} = (w_k)_{1 \leqslant k \leqslant Q} \in E_{\Phi(\mathcal{X})}^Q$ and $\mathbf{b} = (b_k)_{1 \leqslant k \leqslant Q} \in \mathbb{R}^Q$. As a consequence, $\bar{\mathcal{H}}$ can be seen as a multivariate linear model on $\Phi\left(\mathcal{X}\right)$, 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 [14]) *Let* $((x_i, y_i))_{1 \leqslant i \leqslant m} \in (\mathcal{X} \times [\![1, Q]\!])^m$ *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:*

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

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

The M-SVMs thus differ according to the nature of the function $\ell_{\text{M-SVM}}$ which corresponds to a multi-class extension of the hinge loss function.

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* ($\sum_{i=1}^{m} \ell_{M\text{-}SVM}\left(y_i, h\left(x_i\right)\right) = 0$), *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 [13] for a survey). The first one in chronological order is the model of Weston and Watkins [34, 29, 4]. Following the common usage

in machine learning, we denote by $(\cdot)_+$ the truncate function $\max(0, \cdot)$. The loss function ℓ_{WW} of the M-SVM of Weston and Watkins is then given by:

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

The second machine is due to Crammer and Singer [8] and corresponds to the loss function ℓ_{CS} defined as:

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

The most recent model is the one of Lee, Lin and Wahba [22]. Its loss function ℓ_{LLW} is given by:

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

Among the three models, the M-SVM of Lee, Lin and Wahba is the only one that implements asymptotically the theoretically optimal classification rule, the so-called *Bayes decision rule*. It is *Fisher consistent* [22, 35, 28].

2.3 Geometrical margins

From a geometrical point of view, the algorithms described above select functions h^* (sets of the form $\{(w_k^*, b_k^*) : 1 \leqslant k \leqslant Q\}$) 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 [13]) *Let n be a positive integer and let $d_n = \{(x_i, y_i) : 1 \leqslant i \leqslant n\}$ 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 for any pair of distinct categories (k, l), its* margin between 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*

$$d(h) = \min_{1 \leqslant k < l \leqslant Q} \left\{ \min_{i : y_i \in \{k,l\}} |h_k(x_i) - h_l(x_i)| \right\}$$

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

$$d_{kl}(h) = \frac{1}{d(h)} \min_{i : y_i \in \{k,l\}} |h_k(x_i) - h_l(x_i)| - 1.$$

Then we have

$$\forall (k, l) : 1 \leqslant k < l \leqslant Q, \;\; \gamma_{kl}(h) = \gamma_{lk}(h) = d(h) \frac{1 + d_{kl}(h)}{\|w_k - w_l\|}.$$

Remark 1 *The positivity of $d(h)$ is a direct consequence of the fact that the decision function takes the value $*$ in case of ex æquo. By definition, if $h \in \mathcal{H}$ classifies the examples of d_n without error, then*

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

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 [13].

3 The M-SVM of Lee, Lin and Wahba

We now present in more detail the LLW-M-SVM, from which the M-SVM^2 is derived. The reason for this reminder is self-completeness: some of the formulas established in this section will be used in the presentation of the new machine and the proof of the radius-margin bound. We refer the reader to [10] for an introduction to the basic notions of optimization used in the sequel.

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 LLW-M-SVM.

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

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

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

where

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

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

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

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

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: $\xi = (\xi_{ik})_{1\leqslant i\leqslant m, 1\leqslant k\leqslant Q} \in \mathbb{R}_+^{Qm}$. 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 $(\xi_{iy_i})_{1\leqslant i\leqslant m} = 0_m$. The coefficient C, which characterizes the trade-off between the prediction accuracy (on the training set) and the smoothness of the minimizer h^*, 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. We now derive the dual problem of Problem 2. One of the specificities of the LLW-M-SVM compared to the other two M-SVMs rests in the fact that the primal formulation of its training algorithm must incorporate explicitly the sum-to-0 constraint $\sum_{k=1}^{Q} w_k = 0$. In the framework of the implementation of the Lagrangian duality, this raises a difficulty since the feature space can be infinite dimensional. To overcome this difficulty, Lee and her co-authors reformulate the primal problem by making use of a representer theorem [22, 21]. Their approach is the most

elegant one. However, in what follows, since our aim is simply to reestablish some useful formulas, we handle the aforementioned constraint directly irrespective of the dimensionality of $E_{\Phi(\mathcal{X})}$.

Let $\alpha = (\alpha_{ik})_{1 \leqslant i \leqslant m, 1 \leqslant k \leqslant Q} \in \mathbb{R}_+^{Qm}$ be the vector of Lagrange multipliers associated with the constraints of good classification. Similarly, let $\beta = (\beta_{ik})_{1 \leqslant i \leqslant m, 1 \leqslant k \leqslant Q} \in \mathbb{R}_+^{Qm}$ 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:

$$L_1(\mathbf{w}, \mathbf{b}, \xi, \alpha, \beta, \gamma, \delta) =$$
$$\frac{1}{2}\sum_{k=1}^{Q}\|w_k\|^2 + C\sum_{i=1}^{m}\sum_{k=1}^{Q}\xi_{ik} + \sum_{i=1}^{m}\sum_{k=1}^{Q}\alpha_{ik}\left(\langle w_k, \Phi(x_i)\rangle + b_k + \frac{1}{Q-1} - \xi_{ik}\right) - \sum_{i=1}^{m}\sum_{k=1}^{Q}\beta_{ik}\xi_{ik}$$
$$-\langle \gamma, \sum_{k=1}^{Q} w_k\rangle - \delta\sum_{k=1}^{Q} b_k. \tag{4}$$

Setting the gradient of L_1 with respect to w_k equal to the null vector provides us with Q alternative expressions for the optimal value of vector γ:

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

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:

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

where $\delta_{k,l}$ is the Kronecker symbol. Let us now set the gradient of L_1 with respect to $\mathbf{b}$ equal to the null vector. We get

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

and thus

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

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_1 with respect to ξ equal to the null vector gives:

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

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:

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

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

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

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

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

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 LLW-M-SVM, dual formulation)

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

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

3.2 Geometrical margins

The geometrical margins of the hard margin Q-category LLW-M-SVM 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 LLW-M-SVM. Then,*

$$d\left(h^*\right) \geqslant \frac{Q}{Q-1}.$$

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

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

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

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

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

Proposition 3 *For the hard margin Q-category LLW-M-SVM trained on $\{(x_i, y_i) : 1 \leqslant i \leqslant m\}$, 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 Karush-Kuhn-Tucker (KKT) 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 LLW-M-SVM, 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 $\nabla J_{\mathrm{LLW,d}}(\alpha^*)$ is

  $$\forall i \in [\![1, m]\!],\ \forall k \in [\![1, Q]\!] \backslash \{y_i\},\ \frac{\partial}{\partial \alpha_{ik}} J_{\mathrm{LLW,d}}(\alpha^*) = -(H\alpha^*)_{ik} + \frac{1}{Q-1} = \langle w_k^*, \Phi(x_i) \rangle + \frac{1}{Q-1}.$$

  By application of the KKT complementary conditions

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

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

■

4 The M-SVM^2

4.1 Quadratic loss multi-class SVMs: motive and principle

The M-SVMs presented in Section 2.2 share a common feature with the standard pattern recognition SVM: the contribution of the slack variables to their objective functions is linear. Let ξ be the vector of these variables. In the cases of the M-SVMs of Weston and Watkins and Lee, Lin and Wahba, we have $\xi = (\xi_{ik})_{1 \leqslant i \leqslant m, 1 \leqslant k \leqslant Q}$ with $(\xi_{iy_i})_{1 \leqslant i \leqslant m} = 0_m$, and in the case of the model of Crammer and Singer, it is simply $\xi = (\xi_i)_{1 \leqslant i \leqslant m}$. In both cases, the contribution to the objective function is $C\|\xi\|_1$. In the bi-class case, there exists a variant of the standard SVM which is known as the 2-*norm SVM* since for this machine, the empirical contribution to the objective function is $C\|\xi\|_2^2$. Its main advantage, underlined for instance in the Chapter 7 of [27], is that its training algorithm can be expressed, after an appropriate change of kernel, as the training algorithm of a hard margin machine. As a consequence, its leave-one-out cross-validation error can be upper bounded thanks to the radius-margin bound.

Unfortunately, a naive extension of the 2-norm SVM to the multi-class case, resulting from substituting in the objective function of either of the three M-SVMs $\|\xi\|_1$ with $\|\xi\|_2^2$, does not preserve this property. Section 2.4.1.4 of [13] gives detailed explanations about that point. The strategy that we propose to exhibit interesting multi-class generalizations of the 2-norm SVM consists in studying the class of *quadratic loss M-SVMs*, i.e., the class of extensions of the M-SVMs such that the contribution of the slack variables is a quadratic form:

$$C\xi^T M \xi = C \sum_{i=1}^{m} \sum_{j=1}^{m} \sum_{k=1}^{Q} \sum_{l=1}^{Q} m_{ik,jl} \xi_{ik} \xi_{jl}$$

where $M = (m_{ik,jl})_{1 \leqslant i,j \leqslant m, 1 \leqslant k,l \leqslant Q}$ is such that its submatrix M' obtained by suppressing the rows and columns whose indices respectively satisfy $k = y_i$ and $l = y_j$ is symmetric positive definite. The constraints on M correspond to necessary and sufficient conditions for $\xi^T M \xi$ to be a norm of ξ.

4.2 The M-SVM^2 as a multi-class generalization of the 2-norm SVM

In this section, we establish that the idea introduced above provides us with a solution to the problem of interest when the M-SVM used is the one of Lee, Lin and Wahba and M is the block diagonal matrix of general term

$$m_{ik,jl} = (1 - \delta_{y_i,k}) \left(1 - \delta_{y_j,l}\right) \delta_{i,j} \left(\delta_{k,l} + 1\right).$$

We first note that the corresponding matrix M' is actually symmetric positive definite. Indeed, it can be rewritten as follows:

$$M' = I_m \otimes \left(\delta_{k,l} + 1\right)_{1 \leqslant k,l \leqslant Q-1}, \tag{13}$$

where I_m designates the identity matrix of size m and $\otimes$ denotes the Kronecker product. Its spectrum is thus identical to the one of the matrix $\left(\delta_{k,l} + 1\right)_{1 \leqslant k,l \leqslant Q-1}$, i.e., made up of two positive eigenvalues: 1 and Q. The corresponding machine is named M-SVM^2. Its training algorithm is given by the following QP problem.

Problem 5 (M-SVM^2, primal formulation)

$$\min_{\mathbf{w},\mathbf{b},\xi} J_{M\text{-}SVM^2}\left(\mathbf{w},\mathbf{b},\xi\right)$$

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

where

$$J_{M\text{-}SVM^2}\left(\mathbf{w},\mathbf{b},\xi\right) = \frac{1}{2}\sum_{k=1}^{Q} \|w_k\|^2 + C \sum_{i=1}^{m}\sum_{j=1}^{m}\sum_{k=1}^{Q}\sum_{l=1}^{Q} (1 - \delta_{y_i,k}) \left(1 - \delta_{y_j,l}\right) \delta_{i,j} \left(\delta_{k,l} + 1\right) \xi_{ik}\xi_{jl}.$$

Keeping the notations of the preceding sections, the expression of the Lagrangian function associated with this problem is:

$$L_2\left(\mathbf{w},\mathbf{b},\xi,\alpha,\gamma,\delta\right) =$$
$$\frac{1}{2}\sum_{k=1}^{Q} \|w_k\|^2 + C\xi^T M \xi + \sum_{i=1}^{m}\sum_{k=1}^{Q} \alpha_{ik}\left(\langle w_k, \Phi(x_i)\rangle + b_k + \frac{1}{Q-1} - \xi_{ik}\right)$$
$$-\langle \gamma, \sum_{k=1}^{Q} w_k\rangle - \delta \sum_{k=1}^{Q} b_k. \tag{14}$$

Setting the gradient of L_2 with respect to ξ equal to the null vector gives

$$2CM\xi^* = \alpha^*. \tag{15}$$

The coefficient $(1 - \delta_{y_i,k})\left(1 - \delta_{y_j,l}\right)$ has been introduced in the general term of the matrix M so as to verify:

$$\forall i \in [\![1,m]\!],\ \ 2C\left(M\xi\right)_{iy_i} = \alpha_{iy_i} = 0.$$

It springs from (15) that

$$C{\xi^*}^T M \xi^* - {\alpha^*}^T \xi^* = -C{\xi^*}^T M \xi^*. \tag{16}$$

Using the same reasoning that we used to derive the objective function of Problem 3 and (16), at the optimum, (14) simplifies into

$$L_2\left(\mathbf{w}^*,\mathbf{b}^*,\xi^*,\alpha^*,\gamma^*,\delta^*\right) = -\frac{1}{2}{\alpha^*}^T H \alpha^* - C{\xi^*}^T M \xi^* + \frac{1}{Q-1} 1_{Qm}^T \alpha^*.$$

Proving that the M-SVM^2 exhibits the same property as the 2-norm SVM amounts to exhibiting a kernel κ' such that

$$C\xi^{*T}M\xi^* = \frac{1}{2}\alpha^{*T}H'\alpha^* \tag{17}$$

with the general term of the matrix H' being:

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

Combining (15) and (17) gives:

$$\frac{1}{2}\alpha^{*T}H'\alpha^* = 2C^2\xi^{*T}M^TH'M\xi^* = C\xi^{*T}M\xi^*.$$

After some algebra, we get the general term of the matrix $M^TH'M$, which is

$$\left(1 - \delta_{y_i,k}\right)\left(1 - \delta_{y_j,l}\right)\left(\delta_{k,l} + 1\right)\kappa'\left(x_i, x_j\right).$$

Thus, $2C\xi^{*T}M^TH'M\xi^* = \xi^{*T}M\xi^*$ provided that

$$\forall (i,j) \in [\![1,m]\!]^2, \quad \kappa'\left(x_i, x_j\right) = \frac{1}{2C}\delta_{i,j}.$$

This expression of the second kernel is precisely the one obtained in the case of the 2-norm SVM. With this definition of κ', the objective function of the dual problem simplifies into

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

where the matrix $\tilde{H}$ is deduced from H by substituting to the kernel κ the kernel $\tilde{\kappa}$ equal to $\kappa + \kappa'$ ($\tilde{H} = H + H'$). Since $\nabla_{\mathbf{b}}L_2\left(\mathbf{w}, \mathbf{b}, \xi, \alpha, \gamma, \delta\right) = \nabla_{\mathbf{b}}L_1\left(\mathbf{w}, \mathbf{b}, \xi, \alpha, \beta, \gamma, \delta\right)$, the equality constraints of the dual are still given by (7). On the contrary, the only inequality constraints correspond to the nonnegativity of the Lagrange multipliers α_{ik}. Thus, the Wolfe dual of Problem 5 is:

Problem 6 (M-SVM^2, dual formulation)

$$\max_{\alpha} J_{\textit{M-SVM}^2,d}(\alpha)$$

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

where

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

with the general term of the Hessian matrix $\tilde{H}$ being

$$\tilde{h}_{ik,jl} = \left(\delta_{k,l} - \frac{1}{Q}\right)\left(\kappa(x_i, x_j) + \frac{1}{2C}\delta_{i,j}\right).$$

This problem is precisely Problem 4 with $\kappa + \kappa'$ as kernel, which establishes that for the M-SVM^2, as for the 2-norm SVM, a radius-margin bound can be used to choose the soft margin parameter C. By application of Proposition 4 and (17), at the optimum,

$$J_{\text{M-SVM}^2}\left(\mathbf{w}^*, \mathbf{b}^*, \xi^*\right) = \frac{1}{2}\sum_{k=1}^Q \|w_k^*\|^2 + C\xi^{*T}M\xi^* = \frac{1}{2}\alpha^{*T}H\alpha^* + \frac{1}{2}\alpha^{*T}H'\alpha^* = \frac{1}{2}\alpha^{*T}\tilde{H}\alpha^*.$$

Once more by application of Proposition 4,

$$J_{\text{M-SVM}^2,\text{d}}\left(\alpha^*\right) = -\frac{1}{2}\alpha^{*T}\tilde{H}\alpha^* + \frac{1}{Q-1}1_{Qm}^T\alpha^* = \frac{1}{2}\alpha^{*T}\tilde{H}\alpha^*.$$

This enables us to check that $J_{\text{M-SVM}^2}\left(\mathbf{w}^*, \mathbf{b}^*, \xi^*\right) = J_{\text{M-SVM}^2,\text{d}}\left(\alpha^*\right)$.

4.3 Properties and implementation of the M-SVM^2

Contrary to the training algorithm of the standard pattern recognition SVM, the training algorithm of the 2-norm SVM does not incorporate explicitly the constraints of nonnegativity of the slack variables. This is just useless. Indeed, these constraints are actually satisfied by the optimal solution, for which the expression of the slack variables as a function of the (nonnegative) dual variables is simply:

$$\forall i \in [\![1, m]\!], \;\; \xi_i^* = \frac{1}{2C} \alpha_i^*.$$

Problem 5 does not incorporate the constraints of nonnegativity of the slack variables either. In that case however, this makes a significant difference since some of these variables can be negative. At the optimum, their expression can be deduced from (13) and (15), by inverting matrix M'.

$$M'^{-1} = I_m \otimes \left((\delta_{k,l} + 1)_{1 \leqslant k,l \leqslant Q-1} \right)^{-1} = I_m \otimes \left(\delta_{k,l} - \frac{1}{Q} \right)_{1 \leqslant k,l \leqslant Q-1}.$$

We then get

$$\forall i \in [\![1, m]\!], \; \forall k \in [\![1, Q]\!] \setminus \{y_i\}, \;\; \xi_{ik}^* = \frac{1}{2C} \sum_{l=1}^{Q} \left(\delta_{k,l} - \frac{1}{Q} \right) \alpha_{il}^* \tag{18}$$

or equivalently:

$$\forall i \in [\![1, m]\!], \; \forall k \in [\![1, Q]\!] \setminus \{y_i\}, \;\; \xi_{ik}^* = \left(H' \alpha^* \right)_{ik}. \tag{19}$$

The optimal values of the slack variables are only positive on average, since applying on (18) a summation over both indices gives

$$1_{Qm}^T \xi^* = \frac{1}{2CQ} 1_{Qm}^T \alpha^*.$$

The relaxation of the constraints of nonnegativity of the slack variables alters the meaning of the constraints of good classification, although the global connection between a small value of the norm on ξ and a small training error is preserved. We conjecture that for any of the three M-SVMs, no choice of the matrix M can give rise to a machine such that its Wolfe dual problem is the one of a hard margin machine and its slack variables are all nonnegative.

To solve Problem 6, we implemented the Frank-Wolfe algorithm [11] in the same way as we did in [16] to train the M-SVM of Weston and Watkins. The corresponding piece of software is available at the following address: `http://www.loria.fr/~guermeur/M_SVM_2.tar.gz`. The computation of the primal variables and the values taken by the component functions h_k as a function of the data and the dual variables calls for some explanations. At any iteration of the gradient ascent, the expression of the linear part of the model is simply deduced from (6):

$$\forall x \in \mathcal{X}, \; \forall k \in [\![1, Q]\!], \;\; \bar{h}_k(x) = \langle w_k, \Phi(x) \rangle = \sum_{i=1}^{m} \sum_{l=1}^{Q} \left(\frac{1}{Q} - \delta_{k,l} \right) \alpha_{il} \kappa(x_i, x).$$

This expression can be reformulated in the case when x belongs to the training set:

$$\forall i \in [\![1, m]\!], \; \forall k \in [\![1, Q]\!], \;\; \bar{h}_k(x_i) = -\left(H\alpha \right)_{ik}.$$

This is useful indeed, as the computation of the vector $H\alpha$ can also appear as a step in the computation of the dual objective function. The difficulty rests in the computation of the vectors $\mathbf{b}$ and ξ. In the case of the LLW-M-SVM, the KKT complementary conditions imply that at the optimum:

$$\forall i \in [\![1, m]\!], \; \forall k \in [\![1, Q]\!] \setminus \{y_i\}, \;\; \alpha_{ik}^* \in (0, C) \Longrightarrow \langle w_k^*, \Phi(x_i) \rangle + b_k^* = -\frac{1}{Q-1},$$

i.e.,

$$\forall i \in [\![1,m]\!],\ \forall k \in [\![1,Q]\!] \setminus \{y_i\},\ \ \alpha^*_{ik} \in (0,C) \Longrightarrow b^*_k = -\frac{\partial}{\partial \alpha_{ik}} J_{\text{LLW,d}}(\alpha^*).$$

This last formula can also be used before the optimum is reached, simply to obtain a "sensible" (but suboptimal) value for $\mathbf{b}$. Let us define the sets $\mathcal{S}_k$ as follows:

$$\forall k \in [\![1,Q]\!],\ \ \mathcal{S}_k = \{i \in [\![1,m]\!] :\ \alpha^*_{ik} \in (0,C)\}.$$

Setting

$$\forall k \in [\![1,Q]\!],\ \ b'_k = -\frac{1}{|\mathcal{S}_k|} \sum_{i \in \mathcal{S}_k} \frac{\partial}{\partial \alpha_{ik}} J_{\text{LLW,d}}(\alpha)$$

and

$$\forall k \in [\![1,Q]\!],\ \ b_k = b'_k - \frac{1}{Q} \sum_{k=1}^{Q} b'_k$$

provides us in turn with a value for the vector ξ thanks to the formula

$$\forall i \in [\![1,m]\!],\ \forall k \in [\![1,Q]\!] \setminus \{y_i\},\ \ \xi_{ik} = \left(\frac{\partial}{\partial \alpha_{ik}} J_{\text{LLW,d}}(\alpha) + b_k \right)_+. \tag{20}$$

Plugging this expression of vector ξ in the formula giving J_{SM}, one readily obtains an upper bound on the value of the primal objective function (at any step of the gradient ascent). Let $\mathbf{b}^*(\mathbf{w})$ and $\xi^*(\mathbf{w})$ respectively denote the optimal values of $\mathbf{b}$ and ξ corresponding to $\mathbf{w}$ (or equivalently α).

$$\forall i \in [\![1,m]\!],\ \forall k \in [\![1,Q]\!] \setminus \{y_i\},\ \ \xi^*_{ik}(\mathbf{w}) = \left(\frac{\partial}{\partial \alpha_{ik}} J_{\text{LLW,d}}(\alpha) + b^*_k(\mathbf{w}) \right)_+.$$

We have precisely

$$J_{\text{LLW,d}}(\alpha) \leqslant J_{\text{LLW,d}}(\alpha^*) = J_{\text{SM}}(\mathbf{w}^*, \mathbf{b}^*, \xi^*) \leqslant J_{\text{SM}}(\mathbf{w}, \mathbf{b}^*(\mathbf{w}), \xi^*(\mathbf{w})) \leqslant J_{\text{SM}}(\mathbf{w}, \mathbf{b}, \xi),$$

with the limit of $J_{\text{SM}}(\mathbf{w}, \mathbf{b}, \xi)$ as the number of gradient steps increases being $J_{\text{SM}}(\mathbf{w}^*, \mathbf{b}^*, \xi^*)$, which makes it possible to specify a stopping criterion for training based on the value of $J_{\text{SM}}(\mathbf{w}, \mathbf{b}, \xi) - J_{\text{LLW,d}}(\alpha)$. This criterion, by the way, can be an early stopping one. Going back to the M-SVM^2, once more, the KKT complementary conditions provide us with the value of $\mathbf{b}^*$. We get

$$\forall i \in [\![1,m]\!],\ \forall k \in [\![1,Q]\!] \setminus \{y_i\},\ \ \alpha^*_{ik} > 0 \Longrightarrow \langle w^*_k, \Phi(x_i) \rangle + b^*_k = -\frac{1}{Q-1} + \xi^*_{ik}.$$

Making use of (19), this can be reformulated as:

$$\forall i \in [\![1,m]\!],\ \forall k \in [\![1,Q]\!] \setminus \{y_i\},\ \ \alpha^*_{ik} > 0 \Longrightarrow -(H\alpha^*)_{ik} + b^*_k = -\frac{1}{Q-1} + (H'\alpha^*)_{ik}$$

and finally

$$\forall i \in [\![1,m]\!],\ \forall k \in [\![1,Q]\!] \setminus \{y_i\},\ \ \alpha^*_{ik} > 0 \Longrightarrow b^*_k = \left(\tilde{H}\alpha^*\right)_{ik} - \frac{1}{Q-1} = -\frac{\partial}{\partial \alpha_{ik}} J_{\text{M-SVM}^2\text{,d}}(\alpha^*).$$

Obviously, the derivation of this formula is one step shorter if one considers the hard margin machine instead of the M-SVM^2. As in the case of the LLW-M-SVM, it can be used to derive a value for vector $\mathbf{b}$, but this does not give rise directly to a value for vector ξ (and thus to an upper bound on the minimum value of the primal objective function at $\mathbf{w}$, $J_{\text{M-SVM}^2}(\mathbf{w}, \mathbf{b}^*(\mathbf{w}), \xi^*(\mathbf{w}))$), since there is no analytical expression for vector ξ. The reason why there is no equivalent to (20) for the M-SVM^2 is precisely the relaxation of the constraints of nonnegativity of the slack variables.

No help can be expected from considering the hard margin machine instead of the M-SVM^2. The value of its primal objective function:

$$J_{\mathrm{HM}}\left(\tilde{\mathbf{w}}, \mathbf{b}\right) = \frac{1}{2}\sum_{k=1}^{Q}\|\tilde{w}_k\|^2 = \frac{1}{2}\alpha^T\tilde{H}\alpha$$

cannot be used as an upper bound on $J_{\mathrm{LLW,d}}\left(\alpha^*\right) = \frac{1}{2}\alpha^{*T}\tilde{H}\alpha^*$, because the vector α does not necessarily correspond to a feasible solution of the primal (hard margin) problem (Problem 1 with $\tilde{\kappa}$ as kernel). As a consequence, it does not provide us either with an upper bound on $J_{\text{M-SVM}^2}\left(\mathbf{w}, \mathbf{b}^*\left(\mathbf{w}\right), \xi^*\left(\mathbf{w}\right)\right)$. As a matter of fact, the null vector is a feasible solution of Problem 6, whereas it is associated with a function in $\mathcal{H}$ taking a constant value equal to $\mathbf{b}$. To sum up, for a given value of α corresponding to a feasible solution of Problem 6, obtaining an upper bound on $J_{\text{M-SVM}^2}\left(\mathbf{w}, \mathbf{b}^*\left(\mathbf{w}\right), \xi^*\left(\mathbf{w}\right)\right)$ useful to decide to stop training requires to solve an additional optimization problem with $\mathbf{b}$ and ξ as vectors of parameters (or at least ξ). Among the criteria remaining to characterize the vicinity of the optimal solution is the convergence of $\alpha^T\tilde{H}\alpha$ and $\frac{1}{Q-1}1_{Qm}^T\alpha$ towards an identical value (see Proposition 4).

5 Radius-Margin Bound on the Leave-One-Out Cross-Validation Error of the Hard Margin LLW-M-SVM

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

5.1 Multi-class key lemma

Lemma 1 (Multi-class key lemma) *Let us consider a hard margin Q-category LLW-M-SVM on a domain $\mathcal{X}$. Let $d_m = \{(x_i, y_i) : 1 \leqslant i \leqslant m\}$ 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*

$$\max_{1\leqslant k\leqslant Q}\alpha_{pk}^* \geqslant \frac{Q}{(Q-1)^3\mathcal{D}_m^2}$$

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

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 $\alpha^p = (\alpha_{ik}^p) \in \mathbb{R}_+^{Qm}$ the corresponding vector of the dual variables, with $\left(\alpha_{pk}^p\right)_{1\leqslant k\leqslant Q} = 0_Q$. 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 $(\alpha_{pk})_{1\leqslant k\leqslant Q} = 0_Q$. Let us define two more vectors in $\mathbb{R}_+^{Qm}$: $\lambda^p = (\lambda_{ik}^p)_{1\leqslant i\leqslant m, 1\leqslant k\leqslant Q}$ and $\mu^p = (\mu_{ik}^p)_{1\leqslant i\leqslant m, 1\leqslant k\leqslant Q}$. λ^p exhibits additional properties so that the vector $\alpha^* - \lambda^p$ is a feasible solution of Problem 4 under the additional constraint that $\left(\alpha_{pk}^* - \lambda_{pk}^p\right)_{1\leqslant k\leqslant Q} = 0_Q$, i.e., $\alpha^* - \lambda^p$ satisfies the same constraints as α^p. We have thus

$$\forall i \in [\![1, m]\!] \setminus \{p\},\ \forall k \in [\![1, Q]\!] \setminus \{y_i\},\ \ \alpha_{ik}^* - \lambda_{ik}^p \geqslant 0 \Longleftrightarrow \lambda_{ik}^p \leqslant \alpha_{ik}^*.$$

We deduce from the equality constraints of Problem 4 that:

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

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

$$\begin{cases} \forall k \in [\![1, Q]\!], \;\; \lambda^p_{pk} = \alpha^*_{pk} \\ \forall i \in [\![1, m]\!] \setminus \{p\}\,, \; \forall k \in [\![1, Q]\!] \setminus \{y_i\}\,, \;\; 0 \leqslant \lambda^p_{ik} \leqslant \alpha^*_{ik} \\ \forall k \in [\![1, Q-1]\!], \;\; \sum_{i=1}^{m} \sum_{l=1}^{Q} \left(\frac{1}{Q} - \delta_{k,l}\right) \lambda^p_{il} = 0 \end{cases} . \tag{21}$$

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

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

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 $J_{\mathrm{LLW,d}}$. By construction of vectors λ^p and μ^p, we have $J(\alpha^* - \lambda^p) \leqslant J(\alpha^p)$ and $J\left(\alpha^p + K_1\mu^p\right) \leqslant J(\alpha^*)$. Hence,

$$J(\alpha^*) - J(\alpha^* - \lambda^p) \geqslant J(\alpha^*) - J(\alpha^p) \geqslant J\left(\alpha^p + K_1\mu^p\right) - J(\alpha^p). \tag{22}$$

The expression of the first term is

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

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

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

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

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

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

$$J(\alpha^*) - J(\alpha^* - \lambda^p) \leqslant \frac{1}{2} \sum_{k=1}^{Q} \left\| \sum_{i=1}^{m} \sum_{l=1}^{Q} \left(\frac{1}{Q} - \delta_{k,l}\right) \lambda^p_{il} \Phi(x_i) \right\|^2 . \tag{23}$$

We now turn to the right-hand side of (22). The line of reasoning already used for the left-hand side gives:

$$J\left(\alpha^p + K_1\mu^p\right) - J(\alpha^p) = K_1 \nabla J(\alpha^p)^T \mu^p - \frac{K_1^2}{2} \sum_{k=1}^{Q} \left\| \sum_{i=1}^{m} \sum_{l=1}^{Q} \left(\frac{1}{Q} - \delta_{k,l}\right) \mu^p_{il} \Phi(x_i) \right\|^2 . \tag{24}$$

By hypothesis, the M-SVM trained on $d_m \setminus \{(x_p, y_p)\}$ does not classify x_p correctly. This means that there exists $n \in [\![1, Q]\!] \setminus \{y_p\}$ such that $h^p_n(x_p) \geqslant 0$. 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,

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

By substitution into Equation 24, we get

$$J\left(\alpha^p + K_1 \mu^p\right) - J(\alpha^p) \geqslant \frac{K_1 K_2}{Q-1} - \frac{K_1^2}{2} \sum_{k=1}^{Q} \left\| \sum_{i=1}^{m} \sum_{l=1}^{Q} \left(\frac{1}{Q} - \delta_{k,l} \right) \mu_{il}^p \Phi(x_i) \right\|^2. \tag{25}$$

Combining (22), (23) and (25) finally gives

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

$$\frac{K_1 K_2}{Q-1} - \frac{K_1^2}{2} \sum_{k=1}^{Q} \left\| \sum_{i=1}^{m} \sum_{l=1}^{Q} \left(\frac{1}{Q} - \delta_{k,l} \right) \mu_{il}^p \Phi(x_i) \right\|^2. \tag{26}$$

Let $\nu^p = (\nu_{ik}^p)_{1 \leqslant i \leqslant m, 1 \leqslant k \leqslant Q}$ 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 (26) is:

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

By substitution in (26), this implies that

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

The quadratic form ${\lambda^p}^T H \lambda^p$ can be rewritten as

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

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

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

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

For $\eta = (\eta_{ik})_{1\leqslant i\leqslant m, 1\leqslant k\leqslant Q} \in \mathbb{R}^{Qm}$, let $S(\eta) = \frac{1}{Q} \sum_{i=1}^{m} \sum_{k=1}^{Q} \eta_{ik}^p$. Due to the equality constraints satisfied by λ^p,

$$\forall k \in [\![1, Q]\!], \ \sum_{i=1}^{m} \lambda_{ik}^p = S\left(\lambda^p\right).$$

Since $\lambda^p \in \mathbb{R}_+^{Qm}$, by construction,

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

$$\frac{S\left(\lambda^p\right)^2}{Q^2} \sum_{k=1}^{Q} \left\| \sum_{l=1,l\neq k}^{Q} \left(\mathrm{conv}_l \left\{\Phi(x_i) : 1 \leqslant i \leqslant m\right\} - \mathrm{conv}_k \left\{\Phi(x_i) : 1 \leqslant i \leqslant m\right\} \right) \right\|^2$$

where the $\mathrm{conv}_l \left\{\Phi(x_i) : 1 \leqslant i \leqslant m\right\}$ are convex combinations of the $\Phi(x_i)$. As a consequence,

$$\forall (k,l) \in [\![1, Q]\!]^2, \ \left\| \mathrm{conv}_l \left\{\Phi(x_i) : 1 \leqslant i \leqslant m\right\} - \mathrm{conv}_k \left\{\Phi(x_i) : 1 \leqslant i \leqslant m\right\} \right\|^2 \leqslant \mathcal{D}_m^2$$

and

$$\sum_{k=1}^{Q} \left\| \sum_{i=1}^{m} \sum_{l=1}^{Q} \left(\frac{1}{Q} - \delta_{k,l} \right) \lambda_{il}^p \Phi(x_i) \right\|^2 \leqslant \frac{(Q-1)^2}{Q} S\left(\lambda^p\right)^2 \mathcal{D}_m^2.$$

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

$$\frac{(Q-1)^6}{Q^2} S\left(\lambda^p\right)^2 S\left(\nu^p\right)^2 \mathcal{D}_m^4 \geqslant 1. \tag{27}$$

By construction, $S\left(\nu^p\right) = 1$. We now construct a vector λ^p minimizing the objective function S. Since $\forall k \in [\![1, Q]\!], \ \lambda_{pk}^p = \alpha_{pk}^*$,

$$\forall k \in [\![1, Q]\!], \ \sum_{i=1}^{m} \lambda_{ik}^p \geqslant \alpha_{pk}^*.$$

But since

$$\forall (k,l) \in [\![1, Q]\!]^2, \ \sum_{i=1}^{m} \lambda_{ik}^p = \sum_{i=1}^{m} \lambda_{il}^p = S\left(\lambda^p\right),$$

we have further

$$\min_{\lambda^p} S\left(\lambda^p\right) \geqslant \max_{1\leqslant l\leqslant Q} \alpha_{pl}^*.$$

Obviously, the nature of the function S calls for the choice of minimal values for the components λ_{ik}^p, which is coherent with the box constraints in (21). Thus, there exists a vector λ^{p^*} which is a minimizer of S subject to the set of constraints (21) such that

$$\forall k \in [\![1, Q]\!], \ \sum_{i=1}^{m} \lambda_{ik}^{p^*} = \max_{1 \leqslant l \leqslant Q} \alpha_{pl}^*,$$

i.e., $S\left(\lambda^{p^*}\right) = \max_{1 \leqslant l \leqslant Q} \alpha_{pl}^*$. The substitution of the values of $S\left(\nu^p\right)$ and $S\left(\lambda^{p^*}\right)$ in (27) provides us with

$$\left(\max_{1 \leqslant k \leqslant Q} \alpha_{pk}^*\right)^2 \geqslant \frac{Q^2}{(Q-1)^6 \mathcal{D}_m^4}.$$

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

5.2 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 LLW-M-SVM on a domain $\mathcal{X}$. Let $d_m = \{(x_i, y_i) : 1 \leqslant i \leqslant m\}$ 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 enclosing the set $\{\Phi(x_i) : 1 \leqslant i \leqslant m\}$. Then, using the notations of Definition 5, the following upper bound holds true:*

$$\mathcal{L}_m \leqslant \frac{(Q-1)^4}{Q^2} \mathcal{D}_m^2 d\left(h^*\right)^2 \sum_{k<l} \left(\frac{1 + d_{kl}\left(h^*\right)}{\gamma_{kl}\left(h^*\right)}\right)^2. \tag{28}$$

Proof Let $\mathcal{M}\left(d_m\right)$ be the subset of d_m made up of the examples misclassified by the cross-validation procedure ($\left|\mathcal{M}\left(d_m\right)\right| = \mathcal{L}_m$). Lemma 1 exhibits a non-trivial lower bound on $\max_{1 \leqslant k \leqslant Q} \alpha_{pk}^*$ when (x_p, y_p) belongs to $\mathcal{M}\left(d_m\right)$. As a consequence,

$$\sum_{i:(x_i,y_i) \in \mathcal{M}(d_m)} \max_{1 \leqslant k \leqslant Q} \alpha_{ik}^* \geqslant \frac{Q \mathcal{L}_m}{(Q-1)^3 \mathcal{D}_m^2}$$

and thus

$$1_{Qm}^T \alpha^* = \sum_{i=1}^{m} \sum_{k=1}^{Q} \alpha_{ik}^* \geqslant \sum_{i=1}^{m} \max_{1 \leqslant k \leqslant Q} \alpha_{ik}^* \geqslant \frac{Q \mathcal{L}_m}{(Q-1)^3 \mathcal{D}_m^2}. \tag{29}$$

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 (29) thus provides us with the announced result. ■

5.3 Discussion

When $Q = 2$, Equation 2 implies that $d\left(h^*\right) = 1 + \frac{1}{Q-1} = \frac{Q}{Q-1} = 2$. Thus, $\frac{(Q-1)^4}{Q^2} d\left(h^*\right)^2 = 1$. Furthermore, since $d_{12}\left(h^*\right) = 0$, the sum $\sum_{k<l} \left(\frac{1+d_{kl}(h^*)}{\gamma_{kl}(h^*)}\right)^2$ simplifies into $\frac{1}{\gamma^2}$. This means that the expression of the multi-class radius-margin bound simplifies into the one of the standard bi-class radius-margin bound:

$$\mathcal{L}_m \leqslant \left(\frac{\mathcal{D}_m}{\gamma}\right)^2.$$

The formulation of Theorem 2 is the one involving the radius (diameter) and the geometrical margins, so that it appears clearly as a multi-class generalization of the bi-class radius-margin bound. However, in the multi-class case, upper bounding $\sum_{i:(x_i,y_i)\in\mathcal{M}(d_m)} \max_{1\leqslant k\leqslant Q} \alpha_{ik}^*$ by $\sum_{i=1}^m \sum_{k=1}^Q \alpha_{ik}^*$ is useless. A sharper bound is available:

$$\sum_{i:(x_i,y_i)\in\mathcal{M}(d_m)} \max_{1\leqslant k\leqslant Q} \alpha_{ik}^* \leqslant \sum_{i=1}^m \max_{1\leqslant k\leqslant Q} \alpha_{ik}^*$$

and there is no need to make use of Proposition 4. Consequently, we can get a tighter bound on the leave-one-out cross-validation error:

$$\mathcal{L}_m \leqslant \frac{(Q-1)^3}{Q} \mathcal{D}_m^2 \sum_{i=1}^m \max_{1\leqslant k\leqslant Q} \alpha_{ik}^*. \tag{30}$$

If (30) is a tighter bound, the bound of the lemma is the one to be used for model selection, since it is the one that can be derived with respect to the hyperparameters, in the same way as in the bi-class case [5].

The comparison with the radius-margin bound introduced in [33] is also enlightening. This bound is dedicated to the one-versus-one decomposition strategy under the rule of max wins. [20, 19]. More precisely, it appears as a direct consequence of the application of the bi-class radius-margin bound in this framework. However, it applies to all the multi-class discriminant models based on SVMs and for which the bi-class radii and margins can be computed.

Theorem 3 (Model selection criterion I in [33]) *Let us consider a Q-category one-versus-one decomposition method involving $\binom{Q}{2}$ hard margin bi-class SVMs on a domain $\mathcal{X}$. For $1 \leqslant k < l \leqslant Q$, let (w_{kl}^*, b_{kl}^*) be the couple characterizing the machine discriminating categories k and l and γ_{kl}^* its geometrical margin ($\gamma_{kl}^* = \frac{1}{\|w_{kl}^*\|}$). Let $\mathcal{D}_{kl}$ be the diameter of the smallest sphere of the feature space enclosing the set $\{\Phi(x_i) : y_i \in \{k, l\}\}$. Then, the following upper bound holds true:*

$$\mathcal{L}_m \leqslant \sum_{k<l} \left(\frac{\mathcal{D}_{kl}}{\gamma_{kl}^*}\right)^2. \tag{31}$$

If we concentrate on the terms corresponding to a radius or a margin, then (28) and (31) share the same structure. Two arguments favour the second bound. First, by definition, we have

$$\forall (k,l) : 1 \leqslant k < l \leqslant Q, \;\; \mathcal{D}_{kl} \leqslant \mathcal{D}_m.$$

Furthermore, since the one-versus-one strategy maximizes each bi-class margin independently of the others, one can expect that $\forall (k,l) : 1 \leqslant k < l \leqslant Q$, $\gamma_{kl}^* \geqslant \gamma_{kl}(h^*)$. However, the comparison becomes far more complicated if (30) is used in place of (28). An argument in favour of (30) is that if we do not take into account the numbers of support vectors, its computation involves fewer dual variables than the computation of (28). All in all, the most useful bound could simply correspond to the most efficient strategy, either the single-machine one or the one resulting from the one-versus-one decomposition, as a function of the problem at hand. In that respect, it is currently admitted that no multi-class discriminant model based on SVMs is uniformly superior to the others [19, 12, 25].

6 Conclusions and Ongoing Research

In this article, we have introduced a new multi-class SVM: the M-SVM^2. This quadratic loss extension of the M-SVM of Lee, Lin and Wahba is the first M-SVM exhibiting the main property of the 2-norm SVM: its training algorithm can be expressed, after an appropriate change of kernel, as the training algorithm of a hard margin machine (LLW-M-SVM). As in the bi-class case, one

can take advantage of this property by making use of a radius-margin bound as objective function for the model selection procedure. The derivation of the corresponding bound is the second main contribution of the article. This study has 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. Coming after our Vapnik-Chervonenkis theory of the large margin multi-category classifiers [14] and our characterization of the Rademacher complexity of the M-SVMs [15], it provides us with new arguments backing our thesis that the study of multi-category classification should be tackled independently of the one of dichotomy computation.

The evaluation of the M-SVM^2 and its bound must be carried out in a systematic way. This represents a significant amount of work considering the number of M-SVMs (or even decomposition methods) and model selection methods that can be used for the comparative experiments. Obviously, a tuning criterion of particular interest for the comparison is the generalized approximate cross-validation (GACV) [22]. The computational complexity of those experiments should be kept reasonable thanks to the use of algorithms devised to fit the entire regularization path at a cost exceeding only slightly the one of one training of the corresponding machine. The first of those algorithms, dedicated to the standard bi-class SVM, was proposed in [17]. Its extension dedicated to the LLW-M-SVM is described in [21]. The extension to the M-SVM^2 is the subject of an ongoing research.

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 Y. Lee for providing them with additional information regarding her work. Thanks are also due to M. Bertrand and R. Bonidal 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] K.-M. Chung, W.-C. Kao, C.-L. Sun, L.-L. Wang, and C.-J. Lin. Radius margin bounds for support vector machines with the RBF kernel. *Neural Computation*, 15(11):2643–2681, 2003.

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

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

[9] K. Duan, S.S. Keerthi, and A.N. Poo. Evaluation of simple performance measures for tuning SVM hyperparameters. *Neurocomputing*, 51:41–59, 2003.

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

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

[12] J. Fürnkranz. Round robin classification. *Journal of Machine Learning Research*, 2:721–747, 2002.

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

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

[15] Y. Guermeur. Sample complexity of classifiers taking values in $\mathbb{R}^Q$, application to multi-class SVMs. *Communications in Statistics - Theory and Methods*, 39(3):543–557, 2010.

[16] Y. Guermeur, G. Pollastri, A. Elisseeff, D. Zelus, H. Paugam-Moisy, and P. Baldi. Combining protein secondary structure prediction models with ensemble methods of optimal complexity. *Neurocomputing*, 56C:305–327, 2004.

[17] T. Hastie, S. Rosset, R. Tibshirani, and J. Zhu. The entire regularization path for the support vector machine. *Journal of Machine Learning Research*, 5:1391–1415, 2004.

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

[19] C.-W. Hsu and C.-J. Lin. A comparison of methods for multi-class support vector machines. *IEEE Transactions on Neural Networks*, 13(2):415–425, 2002.

[20] S. Knerr, L. Personnaz, and G. Dreyfus. Single-layer learning revisited: A stepwise procedure for building and training a neural network. In F. Fogelman-Soulié and J. Hérault, editors, *Neurocomputing: Algorithms, Architectures and Applications*, volume F68 of *NATO ASI Series*, pages 41–50. Springer-Verlag, 1990.

[21] Y. Lee and Z. Cui. Characterizing the solution path of multicategory support vector machines. *Statistica Sinica*, 16(2):391–409, 2006.

[22] 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.

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

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

[25] R. Rifkin and A. Klautau. In defense of one-vs-all classification. *Journal of Machine Learning Research*, 5:101–141, 2004.

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

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

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

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

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

[31] 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.

[32] L. Wang, P. Xue, and K.L. Chan. Generalized radius-margin bounds for model selection in multi-class SVMs. Technical report, School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore, 639798, 2005.

[33] L. Wang, P. Xue, and K.L. Chan. Two criteria for model selection in multiclass support vector machines. *IEEE Transactions on Systems, Man, and Cybernetics - Part B*, 38(6):1432–1448, 2008.

[34] 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.

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