A Generic Model of Multi-class Support Vector Machine

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

Keywords: multi-class support vector machines, model selection.

Biographical Statement: Y. Guermeur received a French "diplôme d'ingénieur" from the IIE in 1991. He received a PhD in Computer Science from the University Paris 6 in 1997 and the "Habilitation à Diriger des Recherches" (HDR) from the University Nancy 1 in 2007. Permanent researcher at CNRS since 2003, he is currently at the head of the ABC research team in the LORIA laboratory. His research interests include machine learning and computational biology.

Abstract

Roughly speaking, there is one main model of pattern recognition support vector machine, with several variants of lower popularity. On the contrary, among the different multi-class support vector machines which can be found in literature, none is clearly favoured. On the one hand, they exhibit distinct statistical properties. On the other hand, multiple comparative studies between multi-class support vector machines and decomposition methods have highlighted the fact that in practice, each model has its advantages and drawbacks. In this article, we introduce a generic model of multi-class support vector machine. It provides the first unifying definition of all the machines of this kind published so far. This contribution makes it possible to devise new machines meeting specific requirements as well as to analyse globally the statistical properties of the multi-class support vector machines.

1 Introduction

Among all the statistical models developed for pattern recognition, a great many are based on a principle that does not change fundamentally with the number of categories. Basically, they make no difference between dichotomies and polytomies. Things are more complex in the case of the support vector machines (SVMs). Initially, Cortes and Vapnik (1995) devised a class of machines dedicated to the computation of dichotomies. Since then, the attention of the community has focused almost exclusively on one element of this class: the 1-norm SVM. Although variants exist that exhibit appealing properties, such as the 2-norm SVM (Cortes and Vapnik, 1995) or the least squares SVM (LS-SVM) (Suykens and Vandewalle, 1999) their use has remained marginal so far. The first studies dealing with the use of SVMs for multi-category classification, performed by Schölkopf et al. (1995); Vapnik (1995), report results obtained with decomposition methods involving the 1-norm SVM. Multi-class support vector machines (M-SVMs) were only introduced three years later by Weston and Watkins (1998).

During the last decade, many M-SVMs and decomposition methods involving bi-class SVMs have been introduced and evaluated (see Guermeur, 2007a; Liu, 2007, for a survey). Currently, the attention of the community is focused on four main models of M-SVMs: the model of Weston and Watkins (1998), the one of Crammer and Singer (2001), the one of Lee et al. (2004), and the M-SVM^2 (Guermeur and Monfrini, 2011). Although they operate on the same class of functions and their learning problems all extend in a straightforward way the one of bi-class SVMs (precisely the 1-norm and the 2-norm ones),

they exhibit distinct properties. In recent years, several comparative studies between M-SVMs and decomposition methods have been published (see for instance Guermeur, 2002; Hsu and Lin, 2002). In short, they establish that in practice, no model is uniformly superior or inferior to the others with respect to the standard criteria: prediction accuracy, sparsity, computational complexity, etc. The behaviours observed are different, which is in accordance with what was predicted by the theory.

This article introduces a generic model of M-SVM. To the best of our knowledge, this model provides the first unifying definition of all the machines of this kind published so far. It is based on the concept of reproducing kernel Hilbert space of vector-valued functions and locates the M-SVMs in the framework of Tikhonov's regularization theory (Tikhonov and Arsenin, 1977). Our unifying definition makes it possible to devise new machines meeting specific requirements as well as to study globally the statistical properties of the M-SVMs. The first option is illustrated with an investigation of a new class of M-SVMs of particular interest from the point of view of model selection: the class of quadratic loss M-SVMs.

The organization of the paper is as follows. Section 2 introduces the new generic model of M-SVM. The four main M-SVMs are then presented as instances of this model. Section 3 is devoted to the study of the subclass of quadratic loss M-SVMs. At last, we draw conclusions and outline our ongoing research in Section 4.

2 Generic model of multi-class support vector machine

We are interested here in Q-category pattern recognition problems with $3 \leqslant Q < +\infty$. 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, i.e., the set of the integers ranging from 1 to Q, hereafter denoted by $[\![1, Q]\!]$. The assignment of the descriptions to the categories is performed by means of a *classifier*, i.e., a function on $\mathcal{X}$ taking values in $\mathbb{R}^Q$. For such a function g, the corresponding *decision rule* d_g is defined as follows:

$$\forall x \in \mathcal{X}, \quad \begin{cases} \text{if } \exists k \in [\![1, Q]\!] : g_k(x) > \max_{l \neq k} g_l(x), \text{ then } d_g(x) = k \\ \text{else } d_g(x) = * \end{cases} \tag{1}$$

where $*$ denotes a dummy category introduced to deal with the cases of ex æquo. Like all the SVMs, the M-SVMs belong to the class of *kernel machines* (Schölkopf and Smola, 2002; Shawe-Taylor and Cristianini, 2004), which implies that the family of classifiers on

which they operate is induced by a positive type function/kernel (Berlinet and Thomas-Agnan, 2004). In what follows, κ designates a real-valued kernel on $\mathcal{X}^2$ and $\left(\mathbf{H}_\kappa, \langle \cdot, \cdot \rangle_{\mathbf{H}_\kappa}\right)$ the corresponding reproducing kernel Hilbert space (RKHS) (Berlinet and Thomas-Agnan, 2004). For all x in $\mathcal{X}$, κ_x is the element of $\mathbf{H}_\kappa$ such that for all x' in $\mathcal{X}$, $\kappa_x(x') = \kappa(x, x')$. For a given pair (Q, κ), it appears that all the M-SVMs published so far operate on the same class of functions, hereafter denoted by $\mathcal{H}_{\kappa,Q}$ (or simply $\mathcal{H}$). Our generic model shares this property.

2.1 Definition

Since the Q-category M-SVMs all operate on the same class of $\mathbb{R}^Q$-valued functions induced by a kernel, it appears appropriate to base a unifying definition of these machines on an extended definition of the RKHSs dedicated to the case of vector-valued functions. The main benefit of this choice is to highlight the fact that Tikhonov's regularization theory provides a natural theoretical framework for their study. This is all the more useful as the geometrical concept at the basis of the bi-class SVMs, the maximum margin hyperplane (Vapnik, 1982), does not extend nicely to the multi-class case (see for instance Section 2.4.1 in Guermeur, 2007a). The literature provides us with several suitable extensions of the concept of RKHS (see for instance Micchelli and Pontil, 2005). We adopt the one introduced by Wahba (1992).

Definition 1 (RKHS of $\mathbb{R}^Q$-valued functions, after Section 6 in Wahba, 1992) *Let $\tilde{\kappa}$ be a real-valued positive type function on $(\mathcal{X} \times [\![1, Q]\!])^2$. For each (x, k) in $\mathcal{X} \times [\![1, Q]\!]$, let us define the $\mathbb{R}^Q$-valued function $\tilde{\kappa}_{x,k}^{(Q)}$ on $\mathcal{X}$ by the formula*

$$\tilde{\kappa}_{x,k}^{(Q)}(\cdot) = \left(\tilde{\kappa}\left((x, k), (\cdot, l)\right)\right)_{1 \leqslant l \leqslant Q}. \tag{2}$$

The RKHS of $\mathbb{R}^Q$-valued functions $\left(\mathbf{H}_{\tilde{\kappa}^{(Q)}}, \langle \cdot, \cdot \rangle_{\mathbf{H}_{\tilde{\kappa}^{(Q)}}}\right)$ *consists of the linear manifold of all finite linear combinations of functions of the form (2) as (x, k) varies in $\mathcal{X} \times [\![1, Q]\!]$, and its closure with respect to the inner product*

$$\forall \left(x, x'\right) \in \mathcal{X}^2,\ \forall (k, l) \in [\![1, Q]\!]^2,\ \left\langle \tilde{\kappa}_{x,k}^{(Q)}, \tilde{\kappa}_{x',l}^{(Q)} \right\rangle_{\mathbf{H}_{\tilde{\kappa}^{(Q)}}} = \tilde{\kappa}\left((x, k), \left(x', l\right)\right).$$

We can now define the RKHS of $\mathbb{R}^Q$-valued functions at the basis of the Q-category M-SVMs as follows.

Definition 2 (RKHS $\mathbf{H}_{\kappa,Q}$) *Let κ be a real-valued positive type function on $\mathcal{X}^2$. Using the notations of Definition 1, the RKHS of $\mathbb{R}^Q$-valued functions at the basis of a Q-category M-SVM whose kernel is κ, $\left(\mathbf{H}_{\kappa,Q}, \langle\cdot,\cdot\rangle_{\mathbf{H}_{\kappa,Q}}\right)$, is the RKHS $\left(\mathbf{H}_{\tilde{\kappa}(Q)}, \langle\cdot,\cdot\rangle_{\mathbf{H}_{\tilde{\kappa}(Q)}}\right)$ corresponding to the following choice for the kernel $\tilde{\kappa}$:*

$$\forall \left(x,x'\right) \in \mathcal{X}^2,\ \forall \left(k,l\right) \in [\![1,Q]\!]^2,\ \ \tilde{\kappa}\left(\left(x,k\right),\left(x',l\right)\right) = \delta_{k,l}\kappa\left(x,x'\right)$$

where δ is the Kronecker symbol.

The function $\tilde{\kappa}$ involved in Definition 2 actually meets the hypotheses of Definition 1 ($\tilde{\kappa}$ is a kernel on $\left(\mathcal{X}\times[\![1,Q]\!]\right)^2$) since it is the tensor product of a kernel on $\mathcal{X}^2$ and a kernel on $[\![1,Q]\!]^2$ (see for instance Proposition 13.6 in Schölkopf and Smola, 2002). By reasoning on adequately designed Cauchy sequences, one can easily establish that the definition of $\mathbf{H}_{\kappa,Q}$ can be reformulated simply as a function of $\mathbf{H}_\kappa$.

Proposition 1 (Alternative characterization of $\mathbf{H}_{\kappa,Q}$) *Let κ be a real-valued positive type function on $\mathcal{X}^2$ and let $\left(\mathbf{H}_\kappa, \langle\cdot,\cdot\rangle_{\mathbf{H}_\kappa}\right)$ be the corresponding RKHS. Then, $\mathbf{H}_{\kappa,Q} = \mathbf{H}_\kappa^Q$. Furthermore, the inner product of $\mathbf{H}_{\kappa,Q}$ can be expressed as a function of the inner product of $\mathbf{H}_\kappa$ as follows:*

$$\forall \left(\bar{h},\bar{h}'\right) \in \mathbf{H}_{\kappa,Q}^2,\ \bar{h} = \left(\bar{h}_k\right)_{1\leqslant k\leqslant Q},\ \bar{h}' = \left(\bar{h}'_k\right)_{1\leqslant k\leqslant Q},\ \ \left\langle \bar{h},\bar{h}'\right\rangle_{\mathbf{H}_{\kappa,Q}} = \sum_{k=1}^{Q} \left\langle \bar{h}_k,\bar{h}'_k\right\rangle_{\mathbf{H}_\kappa}.$$

Definition 3 (Class of functions $\mathcal{H}_{\kappa,Q}$) *Let κ be a real-valued positive type function on $\mathcal{X}^2$ and let $\mathbf{H}_{\kappa,Q}$ be the RKHS of $\mathbb{R}^Q$-valued functions derived from κ according to Definition 2. Let $\{1\}$ be the one-dimensional space of real-valued constant functions on $\mathcal{X}$. The class of functions at the basis of a Q-category M-SVM whose kernel is κ is*

$$\mathcal{H}_{\kappa,Q} = \mathbf{H}_{\kappa,Q} \oplus \{1\}^Q = \left(\mathbf{H}_\kappa \oplus \{1\}\right)^Q.$$

The functions in $\mathcal{H}_{\kappa,Q}$ can also be seen as multivariate affine functions on $\mathbf{H}_\kappa$. Indeed, due to the reproducing property,

$$\forall h \in \mathcal{H}_{\kappa,Q},\ \forall x \in \mathcal{X},\ \ h\left(x\right) = \bar{h}\left(x\right) + b = \left(\left\langle \bar{h}_k,\kappa_x\right\rangle_{\mathbf{H}_\kappa} + b_k\right)_{1\leqslant k\leqslant Q},$$

where the function $\bar{h} = \left(\bar{h}_k\right)_{1\leqslant k\leqslant Q}$ is an element of $\mathbf{H}_\kappa^Q$ and $b = \left(b_k\right)_{1\leqslant k\leqslant Q} \in \mathbb{R}^Q$. Note that it is also possible to endow the vector space $\mathcal{H}_{\kappa,Q}$ with a structure of RKHS (of vector-valued functions). However, this is useless in the context of this study, since the norm of interest is the one on $\mathbf{H}_{\kappa,Q}$. Let $m \in \mathbb{N}^*$. For a given sequence of examples

$d_m = ((x_i, y_i))_{1 \leqslant i \leqslant m}$ in $(\mathcal{X} \times [\![1, Q]\!])^m$, we denote $\mathbb{R}^{Qm}(d_m)$ the subset of $\mathbb{R}^{Qm}$ made up of the vectors $v = (v_t)_{1 \leqslant t \leqslant Qm}$ satisfying:

$$\left(v_{(i-1)Q+y_i}\right)_{1 \leqslant i \leqslant m} = 0_m. \tag{3}$$

Similarly, $\mathbb{R}_+^{Qm}(d_m) = \mathbb{R}^{Qm}(d_m) \cap \mathbb{R}_+^{Qm}$. Furthermore, for the sake of simplicity, the components of the vectors of $\mathbb{R}^{Qm}(d_m)$ are written with two indices, i.e., v_{ik} in place of $v_{(i-1)Q+k}$, for i in $[\![1, m]\!]$ and k in $[\![1, Q]\!]$. As a consequence, (3) simplifies into $(v_{iy_i})_{1 \leqslant i \leqslant m} = 0_m$. For n in $\mathbb{N}^*$, let $\mathcal{M}_{n,n}(\mathbb{R})$ be the algebra of $n \times n$ matrices over $\mathbb{R}$. Let $\mathcal{M}_{Qm,Qm}(d_m)$ be the subset of $\mathcal{M}_{Qm,Qm}(\mathbb{R})$ made up of the matrices $M = (m_{tu})_{1 \leqslant t,u \leqslant Qm}$ satisfying:

$$\forall j \in [\![1, m]\!], \quad \left(m_{t,(j-1)Q+y_j}\right)_{1 \leqslant t \leqslant Qm} = 0_{Qm}.$$

Once more for the sake of simplicity, the components of the matrices of $\mathcal{M}_{Qm,Qm}(d_m)$ are written with four indices, i.e., $m_{ik,jl}$ in place of $m_{(i-1)Q+k,(j-1)Q+l}$, for (i, j) in $[\![1, m]\!]^2$ and (k, l) in $[\![1, Q]\!]^2$. With these definitions, propositions, and notations at hand, the generic model of M-SVM that we propose is defined as follows.

Definition 4 (New generic model of M-SVM) *Let $\mathcal{X}$ be a non empty set and $Q \in \mathbb{N} \setminus [\![0, 2]\!]$. Let κ be a real-valued positive type function on $\mathcal{X}^2$. Let $\mathbf{H}_{\kappa,Q}$ and $\mathcal{H}_{\kappa,Q}$ be the two classes of functions induced by κ according to Definitions 2 and 3. Let $P_{\mathbf{H}_{\kappa,Q}}$ be the orthogonal projection operator from $\mathcal{H}_{\kappa,Q}$ onto $\mathbf{H}_{\kappa,Q}$. For $m \in \mathbb{N}^*$, let $d_m = ((x_i, y_i))_{1 \leqslant i \leqslant m} \in (\mathcal{X} \times [\![1, Q]\!])^m$ and $\xi \in \mathbb{R}^{Qm}(d_m)$. A Q-*category M-SVM *with kernel κ and training set d_m is a large margin discriminant model trained by solving a convex quadratic programming problem of the form*

Problem 1 (Learning problem of an M-SVM, primal formulation)

$$\min_{h,\xi} \left\{ \|M\xi\|_p^p + \lambda \left\| P_{\mathbf{H}_{\kappa,Q}} h \right\|_{\mathbf{H}_{\kappa,Q}}^2 \right\}$$

$$s.t. \begin{cases} \forall i \in [\![1, m]\!],\ \forall k \in [\![1, Q]\!] \setminus \{y_i\}, \quad K_1 h_{y_i}(x_i) - h_k(x_i) \geqslant K_2 - \xi_{ik} \\ \forall i \in [\![1, m]\!],\ \forall (k, l) \in ([\![1, Q]\!] \setminus \{y_i\})^2, \quad K_3 (\xi_{ik} - \xi_{il}) = 0 \\ \forall i \in [\![1, m]\!],\ \forall k \in [\![1, Q]\!] \setminus \{y_i\}, \quad (2 - p)\xi_{ik} \geqslant 0 \\ (1 - K_1) \sum_{k=1}^{Q} h_k = 0 \end{cases}$$

where $\lambda \in \mathbb{R}_+^*$, $M \in \mathcal{M}_{Qm,Qm}(d_m)$ is a matrix of rank $(Q-1)m$, $p \in \{1, 2\}$, $(K_1, K_3) \in \{0, 1\}^2$, and $K_2 \in \mathbb{R}_+^*$. If $p = 1$, then M is a diagonal matrix.

Definition 5 (Hard and soft margin M-SVM) *If an M-SVM is trained subject to the constraint that the value of the data fit functional of its objective function is null, i.e.,* $\xi = 0_{Qm}$*, it is called a* hard margin *M-SVM. Otherwise, it is called a* soft margin *M-SVM.*

2.2 Motivations of the new definition

So far, there was basically one single unifying definition of the M-SVMs, which had been introduced independently, with minor differences, by several researchers (see for instance Zou et al., 2006; Guermeur, 2007b). Using the notations of this article, it can be formulated as follows:

Definition 6 (Standard definition of an M-SVM) *Let* $\mathcal{X}$ *be a non empty set and* $Q \in \mathbb{N}\backslash[\![0,2]\!]$*. Let* κ *be a real-valued positive type function on* $\mathcal{X}^2$ *and* $\mathcal{H}_{\kappa,Q} = (\mathbf{H}_\kappa \oplus \{1\})^Q$*. For* $m \in \mathbb{N}^*$*, let* $d_m = ((x_i, y_i))_{1 \leqslant i \leqslant m} \in (\mathcal{X} \times [\![1, Q]\!])^m$*. A* Q-category M-SVM *with kernel* κ *and training set* d_m *is a large margin discriminant model trained by solving a convex programming problem of the form*

Problem 2 (Learning problem of an M-SVM, primal formulation)

$$\min_{h \in \mathcal{H}_{\kappa,Q}} \left\{ \sum_{i=1}^{m} \ell_{\text{M-SVM}}\left(y_i, h\left(x_i\right)\right) + \lambda \sum_{k=1}^{Q} \left\|\bar{h}_k\right\|_{\mathbf{H}_\kappa}^2 \right\}$$

$$s.t. \quad \sum_{k=1}^{Q} h_k = 0$$

where $\ell_{\text{M-SVM}}$ *is a convex loss function.*

Definition 4 has been introduced to encompass models that do not fit in Definition 6. In what follows, we call *quadratic loss M-SVMs* the M-SVMs for which $p = 2$. The instances of our generic model whose learning problems cannot be reformulated as instances of Problem 2 are the quadratic loss M-SVMs for which M is not diagonal. When a machine of this kind is considered, solving Problem 1 for a given function h in $\mathcal{H}_{\kappa,Q}$ (feasible but not necessarily optimal), still amounts to solving a convex quadratic programming problem in ξ. On the contrary, one can easily check that when $p = 1$, or $p = 2$ and M is diagonal, there is an analytical expression of the optimal value of ξ as a function of h for any feasible h, which implies that Problem 1 can be reformulated as an instance of Problem 2. Switching from the standard definition to our generic model could be justified by the sole fact that our generic model covers all the M-SVMs published so far, including the M-SVM^2, precisely a quadratic loss M-SVM for which M is not diagonal (see Section 2.4). Furthermore, we

will see in Section 3 that the class of quadratic loss M-SVMs includes other models with appealing properties.

The first set of constraints of Problem 1 corresponds to the constraints of good classification. They are derived from the expression of the decision rule given by (1) and make use, in the case when $K_1 = 0$, of the sum-to-0 constraint $\sum_{k=1}^{Q} h_k = 0$. In Guermeur (2002), we highlighted this equality constraint, which had remained implicit previously. This allowed us to establish that most of the M-SVMs that had been published at that time were simply alternate formulations of the model of Weston and Watkins. It has multiple consequences, among which the fact that the 1-norm SVM and the 2-norm SVM are embedded in both multi-class extensions. Indeed, the equation of the separating hyperplane of a bi-class SVM, $\bar{h}(x) + b = 0$, with $\bar{h} \in \mathbf{H}_\kappa$ and $b \in \mathbb{R}$, can be rewritten as follows:

$$\bar{h}(x) + b = \bar{h}_1(x) - \bar{h}_2(x) + b_1 - b_2 = 2\left(\bar{h}_1(x) + b_1\right) = 0,$$

with $\bar{h}_1 = -\bar{h}_2 = \frac{1}{2}\bar{h}$ and $b_1 = -b_2 = \frac{1}{2}b$. If the constraint $\sum_{k=1}^{Q} h_k = 0$ is introduced explicitly only in the case when $K_1 = 0$, it is for the sake of parsimony. Indeed, it is satisfied in all cases (irrespective of the value of K_1), as will be established in Section 2.3. As usual, slack variables are introduced to relax the constraints of good classification, and make it possible to tolerate some misclassifications. Given the definition of the decision rule associated with a classifier, a description x_i is correctly classified by h if and only if the $Q-1$ differences $h_{y_i}(x_i) - h_k(x_i)$ for $k \neq y_i$ are positive. This can be accounted for by using either $Q-1$ slack variables ξ_{ik} or only one slack variable ξ_i per training example. The second set of constraints implements the second option (for $K_3 = 1$). As in the bi-class case, the constraints of nonnegativity of the slack variables are only introduced in the case when $p = 1$ (the data fit term in the objective function is then linear in the slack variables). This choice is discussed in Section 3.4. The other term of the objective function, the penalizer $\left\|P_{\mathbf{H}_{\kappa,Q}} h\right\|^2_{\mathbf{H}_{\kappa,Q}} = \left\|\bar{h}\right\|^2_{\mathbf{H}_{\kappa,Q}} = \sum_{k=1}^{Q} \left\|\bar{h}_k\right\|^2_{\mathbf{H}_\kappa}$, appears as a direct extension of its bi-class counterpart thanks to the introduction of the RKHS $\mathbf{H}_{\kappa,Q}$. All in all, the specificity of Definition 4 compared to the bi-class one rests in the presence of the matrix M. Its role differs as a function of the value of the parameter p. If $p = 1$, then the diagonal terms $m_{ik,ik}$ define different "misclassification costs" for each training example. The introduction of these additional degrees of freedom provides us with a multi-class extension of the scheme introduced by Veropoulos et al. (1999) for adjusting the sensitivity and specificity of the 1-norm SVM and the 2-norm SVM. This extension subsumes the one introduced by Lee et al. (2004). The role played by the matrix M when

$p = 2$ will be highlighted in Section 3. In the sequel, for the sake of simplicity, when no confusion is possible, $\mathcal{H}$ and $\bar{\mathcal{H}}$ will be used respectively in place of $\mathcal{H}_{\kappa,Q}$ and $\mathbf{H}_{\kappa,Q}$.

2.3 Wolfe dual of Problem 1

The theory of RKHSs ensures that the minimizer of Problem 1 lies in a finite dimensional space, even when $\mathbf{H}_\kappa$ is an infinite dimensional vector space. The simplest way to make use of this essential property consists in solving Problem 1 through its Wolfe dual. Applying the Lagrangian duality here raises no difficulty precisely because $\mathbf{H}_\kappa$ is a Hilbert space. It can be identified with its topological dual so that the constraint $\sum_{k=1}^{Q} h_k = 0$ can be split into two constraints, namely $\sum_{k=1}^{Q} \bar{h}_k = 0$ and $\sum_{k=1}^{Q} b_k = 0$, associated with Lagrange multipliers respectively belonging to $\mathbf{H}_\kappa$ (or its topological dual) and $\mathbb{R}$. The precise set of hypotheses that enables us to derive the dual of Problem 1 the way we do in this section can be found in Chapter 3 of Bonnans (2006).

Without loss of generality, in the case when $p = 1$, we can assume that the diagonal elements of matrix M are nonnegative, i.e., $(m_{ik,ik})_{1 \leqslant i \leqslant m, 1 \leqslant k \leqslant Q} \in \mathbb{R}_+^{Qm}(d_m)$. Once more for notational simplicity, in the case when $K_3 = 1$, the vector of slack variables is written $\xi = \left(\xi_i \left(1 - \delta_{y_i,k}\right)_{1 \leqslant k \leqslant Q}\right)_{1 \leqslant i \leqslant m}$ and the constraints of Problem 1 are adapted in consequence. Let $\alpha = (\alpha_{ik})_{1 \leqslant i \leqslant m, 1 \leqslant k \leqslant Q} \in \mathbb{R}_+^{Qm}(d_m)$ be the vector of Lagrange multipliers associated with the constraints of good classification. Let β be the vector of Lagrange multipliers associated with the constraints of nonnegativity of the slack variables. If $K_3 = 0$, then $\beta = (\beta_{ik})_{1 \leqslant i \leqslant m, 1 \leqslant k \leqslant Q} \in \mathbb{R}_+^{Qm}(d_m)$, otherwise $\beta = (\beta_i)_{1 \leqslant i \leqslant m} \in \mathbb{R}_+^m$. $\gamma \in \mathbf{H}_\kappa$ and $\delta \in \mathbb{R}$ are the Lagrange multipliers respectively associated with the constraints $\sum_{k=1}^{Q} \bar{h}_k = 0$ and $\sum_{k=1}^{Q} b_k = 0$. With these notations at hand, the Lagrangian function of Problem 1 is given by:

$$L(h, \xi, \alpha, \beta, \gamma, \delta) =$$

$$\|M\xi\|_p^p + \lambda \|P_{\bar{\mathcal{H}}} h\|_{\bar{\mathcal{H}}}^2 - \sum_{i=1}^{m} \sum_{k \neq y_i} \alpha_{ik} \left\{K_1 h_{y_i}(x_i) - h_k(x_i) - K_2 + K_3 \xi_i + (1 - K_3)\, \xi_{ik}\right\}$$

$$- (2 - p) \sum_{i=1}^{m} \left\{ K_3 \beta_i \xi_i + (1 - K_3) \sum_{k \neq y_i} \beta_{ik} \xi_{ik} \right\} - (1 - K_1) \left\langle \gamma, \sum_{k=1}^{Q} \bar{h}_k \right\rangle_{\mathbf{H}_\kappa} - (1 - K_1)\, \delta \sum_{k=1}^{Q} b_k.$$

By application of the reproducing property,

$$\forall k \in [\![1, Q]\!], \quad \nabla_{\bar{h}_k} L(h, \xi, \alpha, \beta, \gamma, \delta) = 2\lambda \bar{h}_k - K_1 \sum_{\{i : y_i = k\}} \sum_{l \neq k} \alpha_{il} \kappa_{x_i} + \sum_{\{i : y_i \neq k\}} \alpha_{ik} \kappa_{x_i} - (1 - K_1)\, \gamma.$$

Thus, at the optimum,

$$\forall k \in [\![1, Q]\!],\ \ (1 - K_1)\,\gamma^* = 2\lambda \bar{h}_k^* - K_1 \sum_{\{i:y_i=k\}} \sum_{l \neq k} \alpha_{il}^* \kappa_{x_i} + \sum_{\{i:y_i \neq k\}} \alpha_{ik}^* \kappa_{x_i}. \tag{4}$$

Summing over the index k gives:

$$(1 - K_1)\,\gamma^* = \frac{2\lambda}{Q} \sum_{k=1}^{Q} \bar{h}_k^* + \frac{1 - K_1}{Q} \sum_{i=1}^{m} \sum_{k \neq y_i} \alpha_{ik}^* \kappa_{x_i}.$$

A direct consequence of this equation is that even when $K_1 = 1$,

$$\sum_{k=1}^{Q} \bar{h}_k^* = 0.$$

By substitution into (4), we get

$$\begin{aligned}\forall k \in [\![1, Q]\!],\ \ \bar{h}_k^* &= \frac{1}{2\lambda} \left\{ K_1 \sum_{\{i:y_i=k\}} \sum_{l \neq k} \alpha_{il}^* \kappa_{x_i} + (1 - K_1)\,\frac{1}{Q} \sum_{i=1}^{m} \sum_{l \neq y_i} \alpha_{il}^* \kappa_{x_i} - \sum_{\{i:y_i \neq k\}} \alpha_{ik}^* \kappa_{x_i} \right\} \\ &= \frac{1}{2\lambda} \sum_{i=1}^{m} \sum_{l \neq y_i} \left\{ K_1 \delta_{y_i,k} + (1 - K_1)\,\frac{1}{Q} - \delta_{k,l} \right\} \alpha_{il}^* \kappa_{x_i}.\end{aligned}$$

Taking into account the fact that $(\alpha_{iy_i})_{1 \leqslant i \leqslant m} = 0_m$, this simplifies into

$$\forall k \in [\![1, Q]\!],\ \ \bar{h}_k^* = \frac{1}{2\lambda} \sum_{i=1}^{m} \sum_{l=1}^{Q} \left\{ K_1 \delta_{y_i,k} + (1 - K_1)\,\frac{1}{Q} - \delta_{k,l} \right\} \alpha_{il}^* \kappa_{x_i}. \tag{5}$$

$$\forall k \in [\![1, Q]\!],\ \ \frac{\partial}{\partial b_k} L\,(h, \xi, \alpha, \beta, \gamma, \delta) = -K_1 \sum_{\{i:y_i=k\}} \sum_{l \neq k} \alpha_{il} + \sum_{\{i:y_i \neq k\}} \alpha_{ik} - (1 - K_1)\,\delta.$$

Thus, at the optimum,

$$\forall k \in [\![1, Q]\!],\ \ (1 - K_1)\,\delta^* = -K_1 \sum_{\{i:y_i=k\}} \sum_{l \neq k} \alpha_{il}^* + \sum_{\{i:y_i \neq k\}} \alpha_{ik}^*. \tag{6}$$

A summation over the index k gives:

$$(1 - K_1)\,\delta^* = \frac{1 - K_1}{Q} \sum_{i=1}^{m} \sum_{k \neq y_i} \alpha_{ik}^*.$$

By substitution into (6), taking once more into account the fact that $(\alpha_{iy_i})_{1 \leqslant i \leqslant m} = 0_m$, we get

$$\forall k \in [\![1, Q]\!],\ \ \sum_{i=1}^{m} \sum_{l=1}^{Q} \left\{ K_1 \delta_{y_i,k} + (1 - K_1)\,\frac{1}{Q} - \delta_{k,l} \right\} \alpha_{il}^* = 0. \tag{7}$$

To compute the gradient of the Lagrangian function with respect to vector ξ, we distinguish the four cases corresponding to the possible values of (K_3, p).

- $(K_3 = 0) \wedge (p = 1)$

$$\forall i \in [\![1, m]\!],\ \forall k \in [\![1, Q]\!] \setminus \{y_i\},\ \ \frac{\partial}{\partial \xi_{ik}} L(h, \xi, \alpha, \beta, \gamma, \delta) = m_{ik,ik} - \alpha_{ik} - \beta_{ik}.$$

As a consequence, we get

$$\forall i \in [\![1, m]\!],\ \forall k \in [\![1, Q]\!] \setminus \{y_i\},\ \ \alpha^*_{ik} + \beta^*_{ik} = m_{ik,ik}. \tag{8}$$

- $(K_3 = 0) \wedge (p = 2)$

$$\forall i \in [\![1, m]\!],\ \forall k \in [\![1, Q]\!] \setminus \{y_i\},\ \ \frac{\partial}{\partial \xi_{ik}} L(h, \xi, \alpha, \beta, \gamma, \delta) = 2\left(M^T M \xi\right)_{ik} - \alpha_{ik}.$$

Thus,

$$\forall i \in [\![1, m]\!],\ \forall k \in [\![1, Q]\!] \setminus \{y_i\},\ \ 2\left(M^T M \xi^*\right)_{ik} = \alpha^*_{ik}. \tag{9}$$

Note that since $M \in \mathcal{M}_{Qm,Qm}(d_m)$, we also have

$$\forall i \in [\![1, m]\!],\ \ \left(M^T M \xi^*\right)_{iy_i} = 0,$$

so that (9) can be extended to:

$$2M^T M \xi^* = \alpha^*. \tag{10}$$

- $(K_3 = 1) \wedge (p = 1)$

$$\forall i \in [\![1, m]\!],\ \ \frac{\partial}{\partial \xi_i} L(h, \xi, \alpha, \beta, \gamma, \delta) = \sum_{k \neq y_i} m_{ik,ik} - \sum_{k \neq y_i} \alpha_{ik} - \beta_i.$$

As a consequence,

$$\forall i \in [\![1, m]\!],\ \ \sum_{k \neq y_i} \alpha^*_{ik} + \beta^*_i = \sum_{k \neq y_i} m_{ik,ik}. \tag{11}$$

- $(K_3 = 1) \wedge (p = 2)$

$$\forall i \in [\![1, m]\!],\ \ \frac{\partial}{\partial \xi_i} L(h, \xi, \alpha, \beta, \gamma, \delta) = 2 \sum_{k \neq y_i} \left(M^T M \xi\right)_{ik} - \sum_{k \neq y_i} \alpha_{ik}.$$

Thus,

$$\forall i \in [\![1, m]\!],\ \ 2 \sum_{k \neq y_i} \left(M^T M \xi^*\right)_{ik} = \sum_{k \neq y_i} \alpha^*_{ik}. \tag{12}$$

At the optimum, the terms of the Lagrangian function involving vector b vanish, i.e.,

$$-\sum_{i=1}^{m}\sum_{k\neq y_i}\alpha_{ik}^{*}\left\{K_1 b_{y_i}^{*}-b_k^{*}\right\}=0. \tag{13}$$

Indeed,

$$-\sum_{i=1}^{m}\sum_{k\neq y_i}\alpha_{ik}^{*}\left\{K_1 b_{y_i}^{*}-b_k^{*}\right\}=\sum_{k=1}^{Q}b_k^{*}\left\{-K_1\sum_{\{i:y_i=k\}}\sum_{l\neq k}\alpha_{il}^{*}+\sum_{\{i:y_i\neq k\}}\alpha_{ik}^{*}\right\},$$

and due to (6), the right-hand side of this equation can be rewritten as follows:

$$\sum_{k=1}^{Q}b_k^{*}\left\{-K_1\sum_{\{i:y_i=k\}}\sum_{l\neq k}\alpha_{il}^{*}+\sum_{\{i:y_i\neq k\}}\alpha_{ik}^{*}\right\}=\sum_{k=1}^{Q}b_k^{*}\left(1-K_1\right)\delta^{*}=0.$$

By application of (5),

$$\lambda\left\|P_{\bar{\mathcal{H}}}h^{*}\right\|_{\bar{\mathcal{H}}}^{2}=$$

$$\frac{1}{4\lambda}\sum_{k=1}^{Q}\sum_{i=1}^{m}\sum_{j=1}^{m}\sum_{l=1}^{Q}\sum_{n=1}^{Q}\left\{K_1\delta_{y_i,k}+\left(1-K_1\right)\frac{1}{Q}-\delta_{k,l}\right\}\left\{K_1\delta_{y_j,k}+\left(1-K_1\right)\frac{1}{Q}-\delta_{k,n}\right\}\kappa\left(x_i,x_j\right)\alpha_{il}^{*}\alpha_{jn}^{*}.$$

Since

$$\sum_{k=1}^{Q}\left\{K_1\delta_{y_i,k}+\left(1-K_1\right)\frac{1}{Q}-\delta_{k,l}\right\}\left\{K_1\delta_{y_j,k}+\left(1-K_1\right)\frac{1}{Q}-\delta_{k,n}\right\}=$$

$$K_1\delta_{y_i,y_j}-K_1\delta_{y_i,n}+\left(1-K_1\right)\frac{1}{Q}-\left(1-K_1\right)\frac{1}{Q}-K_1\delta_{y_j,l}-\left(1-K_1\right)\frac{1}{Q}+\delta_{l,n}=$$

$$K_1\left(\delta_{y_i,y_j}-\delta_{y_i,n}-\delta_{y_j,l}\right)-\left(1-K_1\right)\frac{1}{Q}+\delta_{l,n},$$

the expression of the penalizer at the optimum simplifies into

$$\lambda\left\|P_{\bar{\mathcal{H}}}h^{*}\right\|_{\bar{\mathcal{H}}}^{2}=$$

$$\frac{1}{4\lambda}\sum_{i=1}^{m}\sum_{j=1}^{m}\sum_{k=1}^{Q}\sum_{l=1}^{Q}\left\{K_1\left(\delta_{y_i,y_j}-\delta_{y_i,l}-\delta_{y_j,k}\right)-\left(1-K_1\right)\frac{1}{Q}+\delta_{k,l}\right\}\kappa\left(x_i,x_j\right)\alpha_{ik}^{*}\alpha_{jl}^{*}. \tag{14}$$

Making use of the reproducing property, one obtains

$$\sum_{i=1}^{m}\sum_{k\neq y_i}\alpha_{ik}^{*}\left\{K_1\bar{h}_{y_i}^{*}(x_i)-\bar{h}_k^{*}(x_i)\right\}=\sum_{k=1}^{Q}\left\langle\bar{h}_k^{*},K_1\sum_{\{i:y_i=k\}}\sum_{l\neq k}\alpha_{il}^{*}\kappa_{x_i}-\sum_{\{i:y_i\neq k\}}\alpha_{ik}^{*}\kappa_{x_i}\right\rangle_{\mathbf{H}_\kappa}.$$

By application of (4), we get

$$\sum_{i=1}^{m}\sum_{k\neq y_i}\alpha_{ik}^{*}\left\{K_1\bar{h}_{y_i}^{*}(x_i)-\bar{h}_k^{*}(x_i)\right\}=\sum_{k=1}^{Q}\left\langle\bar{h}_k^{*},2\lambda\bar{h}_k^{*}-\left(1-K_1\right)\gamma^{*}\right\rangle_{\mathbf{H}_\kappa}$$

$$= 2\lambda \left\| P_{\bar{\mathcal{H}}} h^* \right\|_{\bar{\mathcal{H}}}^2 - \left\langle (1 - K_1) \sum_{k=1}^{Q} \bar{h}_k^*, \gamma^* \right\rangle_{\mathbf{H}_\kappa}.$$

Since $(1 - K_1) \sum_{k=1}^{Q} \bar{h}_k^* = 0$ according to the constraints of Problem 1 (we have even established a stronger result, namely $\sum_{k=1}^{Q} \bar{h}_k^* = 0$), this simplifies into

$$\sum_{i=1}^{m} \sum_{k \neq y_i} \alpha_{ik}^* \left\{ K_1 \bar{h}_{y_i}^*(x_i) - \bar{h}_k^*(x_i) \right\} = 2\lambda \left\| P_{\bar{\mathcal{H}}} h^* \right\|_{\bar{\mathcal{H}}}^2. \tag{15}$$

Let $H \in \mathcal{M}_{Qm,Qm}(\mathbb{R})$ be the matrix of general term:

$$\forall (i,j,k,l) \in [\![1,m]\!] \times [\![1,m]\!] \times [\![1,Q]\!] \times [\![1,Q]\!],$$

$$h_{ik,jl} = \left\{ K_1 \left(\delta_{y_i,y_j} - \delta_{y_i,l} - \delta_{y_j,k} \right) - (1 - K_1) \frac{1}{Q} + \delta_{k,l} \right\} \kappa(x_i, x_j). \tag{16}$$

Combining (14), (15), and (16) gives:

$$\lambda \left\| P_{\bar{\mathcal{H}}} h^* \right\|_{\bar{\mathcal{H}}}^2 - \sum_{i=1}^{m} \sum_{k \neq y_i} \alpha_{ik}^* \left\{ K_1 \bar{h}_{y_i}^*(x_i) - \bar{h}_k^*(x_i) \right\} = -\frac{1}{4\lambda} {\alpha^*}^T H \alpha^*. \tag{17}$$

Let 1_{Qm} be the vector of $\mathbb{R}^{Qm}$ whose components are all equal to 1. Given (13) and (17), at the optimum,

$$L(h^*, \xi^*, \alpha^*, \beta^*, \gamma^*, \delta^*) = -\frac{1}{4\lambda} {\alpha^*}^T H \alpha^* + K_2 1_{Qm}^T \alpha^* + J(\xi^*),$$

with

$$J(\xi^*) = \left\| M \xi^* \right\|_p^p - \sum_{i=1}^{m} \sum_{k \neq y_i} \alpha_{ik}^* \left\{ K_3 \xi_i^* + (1 - K_3) \xi_{ik}^* \right\} - (2 - p) \sum_{i=1}^{m} \left\{ K_3 \beta_i^* \xi_i^* + (1 - K_3) \sum_{k \neq y_i} \beta_{ik}^* \xi_{ik}^* \right\}.$$

Thus, to obtain the expression of the dual objective function, it remains to express $J(\xi^*)$ as a function of α^*. To that end, we distinguish the cases $p = 1$ and $p = 2$.

• $p = 1$

$$J(\xi^*) = \sum_{i=1}^{m} \sum_{k \neq y_i} m_{ik,ik} \left\{ K_3 \xi_i^* + (1 - K_3) \xi_{ik}^* \right\} - \sum_{i=1}^{m} \sum_{k \neq y_i} \alpha_{ik}^* \left\{ K_3 \xi_i^* + (1 - K_3) \xi_{ik}^* \right\}$$

$$- \sum_{i=1}^{m} \left\{ K_3 \beta_i^* \xi_i^* + (1 - K_3) \sum_{k \neq y_i} \beta_{ik}^* \xi_{ik}^* \right\}$$

$$= K_3 \sum_{i=1}^{m} \left\{ \sum_{k \neq y_i} m_{ik,ik} - \sum_{k \neq y_i} \alpha_{ik}^* - \beta_i^* \right\} \xi_i^* + (1 - K_3) \sum_{i=1}^{m} \sum_{k \neq y_i} \left\{ m_{ik,ik} - \alpha_{ik}^* - \beta_{ik}^* \right\} \xi_{ik}^*.$$

Reporting (8) and (11) in the right-hand side of this equation gives:

$$J\left(\xi^*\right) = 0.$$

• $p = 2$

$$J\left(\xi^*\right) = {\xi^*}^T M^T M \xi^* - \sum_{i=1}^{m} \sum_{k \neq y_i} \alpha_{ik}^* \left\{K_3 \xi_i^* + \left(1 - K_3\right) \xi_{ik}^*\right\}.$$

Here, we distinguish once more the cases $K_3 = 0$ and $K_3 = 1$.

• $(p = 2) \wedge (K_3 = 0)$

Let $N = M^T M$. By definition of M, the rows and columns of N whose indices are those of dummy slack variables are equal to the null vector. Let $\bar{N} \in \mathcal{M}_{(Q-1)m,(Q-1)m}\left(\mathbb{R}\right)$ be the submatrix of N obtained by suppressing these rows and columns. Once more by definition of M, $\bar{N}$ is regular. Let $N^{(-1)} \in \mathcal{M}_{Qm,Qm}\left(\mathbb{R}\right)$ be the matrix deduced from $\bar{N}^{-1}$ by "adding" to $\bar{N}^{-1}$ the aforementioned rows and columns of zeros. Given (10), by construction,

$$\xi^* = \frac{1}{2} N^{(-1)} \alpha^*. \tag{18}$$

Thus,

$$J\left(\xi^*\right) = {\xi^*}^T M^T M \xi^* - {\alpha^*}^T \xi^* = \frac{1}{2} {\alpha^*}^T \xi^* - {\alpha^*}^T \xi^* = -\frac{1}{4} {\alpha^*}^T N^{(-1)} \alpha^*.$$

• $(p = 2) \wedge (K_3 = 1)$

Let $\tilde{\alpha} = \left(\sum_{k \neq y_i} \alpha_{ik}\right)_{1 \leqslant i \leqslant m}$ and $\tilde{\xi} = \left(\xi_i\right)_{1 \leqslant i \leqslant m}$. The vectors $\tilde{\alpha}^*$ and $\tilde{\xi}^*$ are deduced from $\tilde{\alpha}$ and $\tilde{\xi}$ by replacing α_{ik} with α_{ik}^* and ξ_i with ξ_i^*. Then,

$$M \xi^* = \tilde{M} \tilde{\xi}^*,$$

where $\tilde{M} = \left(\tilde{m}_{ik,j}\right)_{1 \leqslant i,j \leqslant m, 1 \leqslant k \leqslant Q} \in \mathcal{M}_{Qm,m}\left(\mathbb{R}\right)$ is the matrix deduced from M as follows:

$$\forall \left(i, j, k\right) \in [\![1, m]\!] \times [\![1, m]\!] \times [\![1, Q]\!], \;\; \tilde{m}_{ik,j} = \sum_{l=1}^{Q} m_{ik,jl}.$$

Furthermore, (12) can be rewritten as follows:

$$2 \tilde{M}^T \tilde{M} \tilde{\xi}^* = \tilde{\alpha}^*.$$

By definition of M, the matrix $\tilde{N}$ equal to $\tilde{M}^T \tilde{M}$ is regular, so that:

$$J\left(\xi^*\right) = {\tilde{\xi}^*}^T \tilde{N} \tilde{\xi}^* - {\tilde{\alpha}^*}^T \tilde{\xi}^* = \frac{1}{2} {\tilde{\alpha}^*}^T \tilde{\xi}^* - {\tilde{\alpha}^*}^T \tilde{\xi}^* = -\frac{1}{4} {\tilde{\alpha}^*}^T \tilde{N}^{-1} \tilde{\alpha}^*.$$

Putting things together, we get the following expression for the objective function of the Wolfe dual of Problem 1:

$$J_{\text{M-SVM,d}}(\alpha) =$$
$$-\frac{1}{4}\left\{\alpha^T\left(\frac{1}{\lambda}H + (1-K_3)(p-1)N^{(-1)}\right)\alpha + K_3(p-1)\tilde{\alpha}^T\tilde{N}^{-1}\tilde{\alpha}\right\} + K_2 1_{Qm}^T\alpha. \quad (19)$$

In the case when $p = 1$, the inequality constraints are deduced from (8) and (11), as a function of the value of K_3. In the case when $p = 2$, we get simply

$$\forall i \in [\![1,m]\!],\ \forall k \in [\![1,Q]\!] \setminus \{y_i\},\ \ \alpha_{ik} \geqslant 0.$$

The equality constraints are deduced from (7). Note that we can take benefit from the fact that

$$\sum_{k=1}^{Q}\sum_{i=1}^{m}\sum_{l=1}^{Q}\left\{K_1\delta_{y_i,k} + (1-K_1)\frac{1}{Q} - \delta_{k,l}\right\}\alpha_{il} = 0$$

and

$$\begin{cases}\sum_{k=1}^{Q}\sum_{i=1}^{m}\sum_{l=1}^{Q}\left\{K_1\delta_{y_i,k} + (1-K_1)\frac{1}{Q} - \delta_{k,l}\right\}\alpha_{il} = 0 \\ \forall k \in [\![1,Q-1]\!],\ \ \sum_{i=1}^{m}\sum_{l=1}^{Q}\left\{K_1\delta_{y_i,k} + (1-K_1)\frac{1}{Q} - \delta_{k,l}\right\}\alpha_{il} = 0\end{cases}$$
$$\Longrightarrow \sum_{i=1}^{m}\sum_{l=1}^{Q}\left\{K_1\delta_{y_i,Q} + (1-K_1)\frac{1}{Q} - \delta_{Q,l}\right\}\alpha_{il} = 0$$

to reduce their number to $Q-1$. Thus, the Wolfe dual of Problem 1 is:

Problem 3 (Learning problem of a soft margin M-SVM, dual formulation)

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

$$s.t.\begin{cases}\forall i \in [\![1,m]\!],\ \forall k \in [\![1,Q]\!] \setminus \{y_i\},\ \ 0 \leqslant (1-K_3)(2-p)\alpha_{ik} \leqslant (2-p)m_{ik,ik} \\ \forall i \in [\![1,m]\!],\ \ 0 \leqslant K_3(2-p)\sum_{k\neq y_i}\alpha_{ik} \leqslant (2-p)\sum_{k\neq y_i}m_{ik,ik} \\ \forall i \in [\![1,m]\!],\ \forall k \in [\![1,Q]\!] \setminus \{y_i\},\ \ (p-1)\alpha_{ik} \geqslant 0 \\ \forall k \in [\![1,Q-1]\!],\ \ \sum_{i=1}^{m}\sum_{l=1}^{Q}\left\{K_1\delta_{y_i,k} + (1-K_1)\frac{1}{Q} - \delta_{k,l}\right\}\alpha_{il} = 0\end{cases}$$

where $J_{\textit{M-SVM,d}}(\alpha)$ is given by (19).

With slight modifications, the derivation above can be adapted to express the Wolfe dual of the learning problem of a hard margin machine. We then get:

Problem 4 (Learning problem of a hard margin M-SVM, dual formulation)

$$\max_{\alpha}\left\{-\frac{1}{4\lambda}\alpha^T H\alpha + K_2 1_{Qm}^T\alpha\right\}$$

$$s.t. \begin{cases} \forall i \in [\![1, m]\!],\ \forall k \in [\![1, Q]\!] \setminus \{y_i\},\ \ \alpha_{ik} \geqslant 0 \\ \forall k \in [\![1, Q-1]\!],\ \ \sum_{i=1}^{m} \sum_{l=1}^{Q} \left\{ K_1 \delta_{y_i,k} + (1 - K_1)\, \frac{1}{Q} - \delta_{k,l} \right\} \alpha_{il} = 0 \end{cases}.$$

The value of $b^* = (b_k^*)_{1 \leqslant k \leqslant Q}$ is deduced from the Karush-Kuhn-Tucker (KKT) complementary conditions. In the case when $K_1 = 0$, the values of the components of this vector are obtained individually, and it suffices to check that they do satisfy the sum-to-0 constraint $1_Q^T b^* = 0$. In the case when $K_1 = 1$, the KKT complementary conditions only provide us with values for the differences between the values of the components. This means that these latter values are only known up to an additive constant. This additional degree of freedom can be used to enforce the constraint $1_Q^T b^* = 0$. This completes the discussion on the reason why the constraint $\sum_{k=1}^{Q} h_k = 0$ appears in Problem 1 only for $K_1 = 0$.

2.4 Characterization of the four main M-SVMs

In chronological order, the first M-SVM is the one of Weston and Watkins (1998). As was pointed out in Section 2.2, it was independently introduced by other researchers under various forms (see for instance Vapnik, 1998; Bredensteiner and Bennett, 1999). If we reformulate its learning problem as an instance of Problem 2, then the corresponding loss function ℓ_{WW} is given by:

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

where $(\cdot)_+$ denotes the troncate function $\max(0, \cdot)$. The second M-SVM is due to Crammer and Singer (2001). It is based on the class of functions $\bar{\mathcal{H}}$, i.e., it satisfies the additional restriction $(b_k)_{1 \leqslant k \leqslant Q} = 0_Q$. The expression of its loss function ℓ_{CS} is:

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

The machine of Lee et al. (2004) corresponds to the loss function ℓ_{LLW} given by:

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

At last, the most recent model is the M-SVM^2 (Guermeur and Monfrini, 2011). Contrary to the three former models, its definition cannot be based on a specification of Problem 2 (see Section 2.2). It springs from a specification of Problem 1. Precisely, it is the M-SVM corresponding to $p = 2$ and $(K_t)_{1 \leqslant t \leqslant 3} = \left(0, \frac{1}{Q-1}, 0\right)^T$, with the matrix M being

instantiated by the matrix $M^{(2)}$ of general term:

$$m^{(2)}_{ik,jl} = (1 - \delta_{y_i,k}) \left(1 - \delta_{y_j,l}\right) \left(\delta_{k,l} + \frac{\sqrt{Q} - 1}{Q - 1}\right) \delta_{i,j}.$$

Let $N^{(2)} = {M^{(2)}}^T M^{(2)}$. Its general term is:

$$n^{(2)}_{ik,jl} = (1 - \delta_{y_i,k}) \left(1 - \delta_{y_j,l}\right) (\delta_{k,l} + 1) \, \delta_{i,j}. \tag{20}$$

Let $I_{Qm}(d_m)$ designate the diagonal matrix of $\mathcal{M}_{Qm,Qm}(d_m)$ given by:

$$I_{Qm}(d_m) = \left(\delta_{i,j}\delta_{k,l}\left(1 - \delta_{y_i,k}\right)\right)_{1 \leqslant i,j \leqslant m, 1 \leqslant k,l \leqslant Q}.$$

In order to characterize the four main M-SVMs as instances of our generic model of M-SVM, we express the primal formulation of their learning problems as a specification of Problem 1. The corresponding values of the hyperparameters are reported in Table 1.

M-SVM	M	p	K_1	K_2	K_3
WW-M-SVM	$I_{Qm}(d_m)$	1	1	1	0
CS-M-SVM	$\frac{1}{Q-1} I_{Qm}(d_m)$	1	1	1	1
LLW-M-SVM	$I_{Qm}(d_m)$	1	0	$\frac{1}{Q-1}$	0
M-SVM^2	$M^{(2)}$	2	0	$\frac{1}{Q-1}$	0

Table 1: Specifications of the four main M-SVMs. The first three machines are the ones of Weston and Watkins (WW), Crammer and Singer (CS), and Lee, Lin, and Wahba (LLW).

As mentioned in introduction, those machines exhibit distinct properties. The M-SVM of Crammer and Singer can be programmed more efficiently than the model of Weston and Watkins (Crammer and Singer, 2001; Aiolli and Sperduti, 2002). This springs from the fact that the class of functions that it uses is $\bar{\mathcal{H}}$ (instead of $\mathcal{H}$). As a consequence, the Wolfe dual of its learning problem involves no equality constraint, which implies that it can be decomposed into multiple small optimization problems. This statement must be specified. The standard formulation of the aforementioned dual problem actually contains the following equality constraints:

$$\forall i \in [\![1, m]\!], \ \sum_{k=1}^{Q} \alpha_{ik} = \sum_{k \neq y_i} m_{ik,ik}.$$

They are derived from (11), by setting for all i in $[\![1, m]\!]$, $\alpha_{iy_i} = \beta_i$. However, as in the case of all the other M-SVMs, there is no reason why the multipliers ensuring the nonnegativity

of the slack variables should appear in the dual problem. Thus, (11) precisely generates inequality constraints, as stated in Section 2.3. By getting rid of the (true) equality constraints of Problem 3, it is possible to devise iterative training algorithms such that at each step, the optimization is performed with respect to the dual variables associated with one single training example. The speed-up is obtained at the expense of the use of a model of lower capacity, since the margin Natarajan dimension (Guermeur, 2007b) of $\bar{\mathcal{H}}$ is inferior to the margin Natarajan dimension of $\mathcal{H}$. The M-SVM of Lee, Lin and Wahba was the first multi-class machine implementing asymptotically the Bayes decision rule. Its loss function is Fisher consistent (Zhang, 2004; Liu, 2007; Tewari and Bartlett, 2007). The main property of the M-SVM^2 is discussed in the next section. MSVMpack (Lauer and Guermeur, 2011) provides a unifying implementation of all four machines.

3 Quadratic loss multi-class support vector machines

The 2-norm SVM is the instance of Vapnik's model of SVM obtained by setting the data fit term of the objective function of the primal formulation of the learning problem equal to the square of the ℓ_2 norm of the vector of slack variables. Its main advantage is that the dual formulation of its learning problem can be expressed as the dual formulation of the learning problem of a hard margin machine using a different kernel. Thus, its leave-one-out cross-validation error can be upper bounded thanks to the radius-margin bound (Vapnik, 1998). 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 main M-SVMs for which $p = 1$ the empirical term with $\|\xi\|_2^2$, does not preserve this property. Section 2.4.1.4 of Guermeur (2007a) gives detailed explanations about that point. The strategy that we advocate to exhibit interesting multi-class extensions of the 2-norm SVM consists in studying the class of M-SVMs which motivated the introduction of our generic model, i.e., the class of quadratic loss M-SVMs (for which $p = 2$). In this section, we focus on three subclasses made up of the quadratic loss extensions of the three main M-SVMs for which $p = 1$. We establish whether or not these subclasses include a machine sharing the main property of the 2-norm SVM. The corresponding proofs are based on an alternate definition of the Hessian matrix H. It must be borne in mind that since this matrix appears in the formulas only through the quadratic form $\alpha^T H \alpha$ and $\alpha \in \mathbb{R}_+^{Qm}\,(d_m)$, its rows and columns corresponding to dummy variables can be set arbitrarily. Formula (16) corresponds to the simplest (most compact) expression. In the sequel, we use instead

the sparsest option:

$$\forall\,(i,j,k,l) \in [\![1,m]\!] \times [\![1,m]\!] \times [\![1,Q]\!] \times [\![1,Q]\!],$$

$$h_{ik,jl} = (1-\delta_{y_i,k})\left(1-\delta_{y_j,l}\right)\left\{K_1\left(\delta_{y_i,y_j}-\delta_{y_i,l}-\delta_{y_j,k}\right)-(1-K_1)\,\frac{1}{Q}+\delta_{k,l}\right\}\kappa(x_i,x_j),$$

which is more appropriate from a computational point of view. At last, the vectors and matrices considered are those introduced in Section 2.

3.1 The M-SVM^2 as a quadratic loss extension of the LLW-M-SVM

The M-SVM^2 was precisely designed to meet the requirement discussed above. The corresponding property can be formulated as follows.

Proposition 2 *The dual formulation of the learning problem of the M-SVM^2 is identical to the dual formulation of the learning problem of a hard margin LLW-M-SVM (using a different kernel).*

To keep the article self-contained, we give the sketch of the proof of this proposition (details are given by Guermeur and Monfrini, 2011).

Proof The specification of Problem 4 corresponding to the hard margin LLW-M-SVM is:

Problem 5 (Learning problem of a hard margin LLW-M-SVM, dual formulation)

$$\max_{\alpha}\left\{-\frac{1}{4\lambda}\alpha^T H\alpha+\frac{1}{Q-1}1_{Qm}^T\alpha\right\}$$

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

with the general term of the Hessian matrix H being

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

The specification of Problem 3 corresponding to the M-SVM^2 is:

Problem 6 (Learning problem of an M-SVM^2, dual formulation)

$$\max_{\alpha}\left\{-\frac{1}{4\lambda}\alpha^T H\alpha-\frac{1}{4}\alpha^T N^{(-1)}\alpha+\frac{1}{Q-1}1_{Qm}^T\alpha\right\}$$

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

with the matrix H being the one of Problem 5 and the general term of the matrix N being

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

$\bar{N}$ is the block diagonal matrix $I_m \otimes (\delta_{k,l} + 1)_{1 \leqslant k,l \leqslant Q-1}$, where I_m designates the identity matrix of size m and $\otimes$ denotes the Kronecker product. We first check that this matrix is actually symmetric positive definite, since its spectrum is made up of two positive eigenvalues: 1 and Q. $\bar{N}^{-1} = I_m \otimes \left((\delta_{k,l} + 1)_{1 \leqslant k,l \leqslant Q-1}\right)^{-1} = I_m \otimes \left(\delta_{k,l} - \frac{1}{Q}\right)_{1 \leqslant k,l \leqslant Q-1}$ and finally, $N^{(-1)}$ is the matrix of general term

$$n^{(-1)}_{ik,jl} = (1 - \delta_{y_i,k}) \left(1 - \delta_{y_j,l}\right) \left(\delta_{k,l} - \frac{1}{Q}\right) \delta_{i,j}. \tag{21}$$

It appears that $n^{(-1)}_{ik,jl}$ is equal to $h_{ik,jl}$ with $\kappa(x_i, x_j)$ replaced with $\delta_{i,j}$. Thus, if we define the kernel κ' as follows:

$$\forall (i,j) \in [\![1, m]\!]^2, \quad \kappa'(x_i, x_j) = \lambda \delta_{i,j}, \tag{22}$$

then we get

$$\frac{1}{4\lambda} \alpha^T H \alpha + \frac{1}{4} \alpha^T N^{(-1)} \alpha = \frac{1}{4\lambda} \alpha^T H'' \alpha,$$

where H'' is the matrix deduced from H by replacing the kernel κ with the kernel $\kappa'' = \kappa + \kappa'$. This implies that Problems 5 and 6 are identical up to a change of kernel, which concludes the proof. ■

It is noteworthy that the change of kernel considered in the proof above is the same as the one of the bi-class case. The introduction of the M-SVM^2 is useful indeed, since an extended radius-margin bound is available for the hard margin LLW-M-SVM: Theorem 2 in Guermeur and Monfrini (2011).

3.2 Quadratic loss extensions of the WW-M-SVM

In this section, we establish that the class of quadratic loss M-SVMs provides us also with an extension of the WW-M-SVM sharing the main property of the 2-norm SVM.

Proposition 3 *There exists a quadratic loss extension of the WW-M-SVM such that the dual formulation of its learning problem is identical to the dual formulation of the learning problem of a hard margin WW-M-SVM (using a different kernel).*

Proof The specification of Problem 4 corresponding to the hard margin WW-M-SVM is:

Problem 7 (Learning problem of a hard margin WW-M-SVM, dual formulation)

$$\max_{\alpha} \left\{ -\frac{1}{4\lambda} \alpha^T H \alpha + 1_{Qm}^T \alpha \right\}$$

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

with the general term of the Hessian matrix H being

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

Exhibiting a quadratic loss extension of the WW-M-SVM such that the dual formulation of its learning problem is Problem 7 up to a change of kernel amounts to exhibiting a matrix M satisfying the hypotheses of Definition 4 such that the general term of the corresponding matrix $N^{(-1)}$ is:

$$n_{ik,jl}^{(-1)} = \frac{1}{\lambda} \left(1 - \delta_{y_i,k}\right) \left(1 - \delta_{y_j,l}\right) \left(\delta_{y_i,y_j} - \delta_{y_i,l} - \delta_{y_j,k} + \delta_{k,l}\right) \kappa'(x_i, x_j). \tag{23}$$

If we make the assumption that κ' is once more given by (22), then (23) simplifies into

$$n_{ik,jl}^{(-1)} = \left(1 - \delta_{y_i,k}\right) \left(1 - \delta_{y_j,l}\right) \left(1 + \delta_{k,l}\right) \delta_{i,j}, \tag{24}$$

which, according to (20), means that $N^{(-1)}$ is equal to the matrix N associated with the M-SVM^2. This is possible if and only if the matrix N that we are looking for can be equal to the matrix $N^{(-1)}$ associated with the M-SVM^2, i.e., can be the matrix of general term:

$$n_{ik,jl} = \left(1 - \delta_{y_i,k}\right) \left(1 - \delta_{y_j,l}\right) \left(\delta_{k,l} - \frac{1}{Q}\right) \delta_{i,j}. \tag{25}$$

Thus, to complete the proof, it suffices to exhibit a matrix M satisfying the hypotheses of Definition 4 such that the general term of the matrix $M^T M$ is given by (25). A possible solution is the matrix of general term:

$$m_{ik,jl} = \left(1 - \delta_{y_i,k}\right) \left(1 - \delta_{y_j,l}\right) \left(\delta_{k,l} - \frac{\sqrt{Q} - 1}{\sqrt{Q}\,(Q-1)}\right) \delta_{i,j}. \tag{26}$$

To sum up, a quadratic loss extension of the WW-M-SVM such that the dual formulation of its learning problem is identical to the dual formulation of the learning problem of a hard margin WW-M-SVM is the machine parameterized as follows: $p = 2$ and $(K_t)_{1 \leqslant t \leqslant 3} = (1, 1, 0)^T$, with M being the matrix whose general term is given by (26). This concludes the proof. ■

3.3 Quadratic loss extensions of the CS-M-SVM

In the case of the CS-M-SVM, the result available is negative.

Proposition 4 *There exists no quadratic loss extension of the CS-M-SVM such that the dual formulation of its learning problem is identical to the dual formulation of the learning problem of a hard margin M-SVM.*

Proof Given the fact that the CS-M-SVM operates on $\bar{\mathcal{H}}$ instead of $\mathcal{H}$, the specification of Problem 3 corresponding to a quadratic loss extension of this machine is given by:

Problem 8 (Learning problem of a quadratic loss CS-M-SVM, dual formulation)

$$\max_{\alpha} \left\{ -\frac{1}{4\lambda} \alpha^T H \alpha - \frac{1}{4} \tilde{\alpha}^T \tilde{N}^{-1} \tilde{\alpha} + 1_{Qm}^T \alpha \right\}$$
$$s.t. \ \forall i \in [\![1, m]\!], \ \forall k \in [\![1, Q]\!] \setminus \{y_i\}, \ \ \alpha_{ik} \geqslant 0$$

with the matrix H being the one of Problem 7.

Thus, proving Proposition 4 is equivalent to establishing that Problem 8 cannot be reformulated as an instance of Problem 4 (without the equality constraints, to take into account once more the fact that the optimization is performed over $\bar{\mathcal{H}}$ instead of $\mathcal{H}$). This boils down to establishing that the objective function of Problem 8 cannot be reformulated as an instance of the objective function of Problem 4. A sufficient condition is to establish that one cannot exhibit a matrix $O = (o_{ij})_{1 \leqslant i,j \leqslant m} \in \mathcal{M}_{m,m}(\mathbb{R})$, a value of K_1 in $\{0, 1\}$, and a kernel κ' such that the two quadratic forms:

$$\sum_{i=1}^{m} \sum_{j=1}^{m} o_{ij} \sum_{k \neq y_i} \alpha_{ik} \sum_{l \neq y_j} \alpha_{jl}$$

and

$$\frac{1}{\lambda} \sum_{i=1}^{m} \sum_{j=1}^{m} \sum_{k \neq y_i} \sum_{l \neq y_j} \left\{ K_1 \left(\delta_{y_i, y_j} - \delta_{y_i, l} - \delta_{y_j, k} \right) - (1 - K_1) \frac{1}{Q} + \delta_{k,l} \right\} \kappa'(x_i, x_j) \alpha_{ik} \alpha_{jl}$$

are identical. This is the case indeed, since irrespective of the value of K_1, the coefficient of $\alpha_{ik}\alpha_{jl}$ in the first quadratic form, o_{ij}, depends only on (i, j), whereas the same coefficient in the second quadratic form also depends on (k, l). ■

3.4 Discussion

Even though the primal formulation of the learning problem of the 2-norm SVM does not incorporate explicitly the constraints of nonnegativity of the slack variables, these constraints are satisfied by the optimal solution, for which we get:

$$\xi^* = \frac{1}{2}\alpha^*.$$

The primal formulation of the learning problem of the quadratic loss M-SVMs does not incorporate these constraints either. In that case however, this makes a significant difference since some of these variables can be negative. In the case when $K_3 = 0$ (case for which we could exhibit interesting quadratic loss M-SVMs), at the optimum, the expression of vector ξ is given by (18). Thus, in the case of the M-SVM^2, for which the expression of the general term of the matrix $N^{(-1)}$ is given by (21), we get:

$$\forall i \in [\![1, m]\!],\ \sum_{k=1}^{Q} \xi_{ik}^* = \frac{1}{2Q} \sum_{k=1}^{Q} \alpha_{ik}^*.$$

In the case of the quadratic loss extension of the WW-M-SVM introduced in Section 3.2, for which the general term of the corresponding matrix $N^{(-1)}$ is given by (24), we get:

$$\forall i \in [\![1, m]\!],\ \sum_{k=1}^{Q} \xi_{ik}^* = \frac{Q}{2} \sum_{k=1}^{Q} \alpha_{ik}^*.$$

These equations establish that although the nonnegativity of the slack variables is not ensured, a weaker result remains available in both cases: for each training example, the optimal values of the slack variables are nonnegative on average.

The relaxation of the constraints of nonnegativity of the slack variables obviously alters the meaning of the constraints of good classification, although the global connection between a small value of the norm of ξ and a small training error is preserved. We conjecture that for any of the three main M-SVMs for which $p = 1$, no choice of the matrix M can give rise to a quadratic loss extension such that the dual formulation of its learning problem is the one of a hard margin machine and its slack variables are all nonnegative.

4 Conclusions and ongoing research

In this article, a generic model of multi-class support vector machine has been introduced. To the best of our knowledge, it provides the first unifying definition of all the machines

of this kind published so far. Using it in place of the standard unifying definition, which does not encompass the class of quadratic loss M-SVMs, opens new perspectives. It can be applied to the design of new machines exhibiting specific properties. In that respect, with the machine introduced in Section 3.2 at hand, deriving an extended radius-margin bound dedicated to the hard margin WW-M-SVM has become a problem of high interest. In our opinion, the main advantage of the generic model is to make it possible to analyse globally the statistical properties of the M-SVMs. Here, the first example that comes to mind is consistency. We already know that some of the M-SVMs asymptotically implement the Bayes decision rule whereas some others do not. Our current aim is to establish infinite-sample consistency conditions as a function of the values of the hyperparameters M, p, and $(K_t)_{1\leqslant t\leqslant 3}$.

Acknowledgments

The author would like to thank S. Canu for sharing his knowledge on optimization in RKHSs.

References

F. Aiolli and A. Sperduti. An efficient SMO-like algorithm for multiclass SVM. In *Neural Networks for Signal Processing 2002*, pages 297–306, 2002.

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

F. Bonnans. *Optimisation Continue: Cours et problèmes corrigés*. Dunod, Paris, 2006. (in French).

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

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

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

Y. Guermeur. Combining discriminant models with new multi-class SVMs. *Pattern Analysis and Applications*, 5(2):168–179, 2002.

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

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

Y. Guermeur and E. Monfrini. A quadratic loss multi-class SVM for which a radius-margin bound applies. *Informatica*, 22(1):73–96, 2011.

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

F. Lauer and Y. Guermeur. MSVMpack: a multi-class support vector machine package. *Journal of Machine Learning Research*, 12:2293–2296, 2011.

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.

Y. Liu. Fisher consistency of multicategory support vector machines. In *Eleventh International Conference on Artificial Intelligence and Statistics*, pages 289–296, 2007.

C.A. Micchelli and M. Pontil. On learning vector-valued functions. *Neural Computation*, 17(1):177–204, 2005.

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

B. Schölkopf, C. Burges, and V. Vapnik. Extracting support data for a given task. In *KDD'95*, pages 252–257, 1995.

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

J.A.K. Suykens and J. Vandewalle. Least squares support vector machine classifiers. *Neural Processing Letters*, 9(3):293–300, 1999.

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

A.N. Tikhonov and V.Y. Arsenin. *Solutions of Ill-Posed Problems*. V.H. Winston & Sons, Washington, D.C., 1977.

V. Vapnik. *Estimation of Dependences Based on Empirical Data*. Springer-Verlag, New York, 1982.

V.N. Vapnik. *The Nature of Statistical Learning Theory*. Springer-Verlag, New York, 1995.

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

K. Veropoulos, C. Campbell, and N. Cristianini. Controlling the sensitivity of support vector machines. In *IJCAI'99*, pages 55–60, 1999.

G. Wahba. Multivariate function and operator estimation, based on smoothing splines and reproducing kernels. In M. Casdagli and S. Eubank, editors, *Nonlinear Modeling and Forecasting, SFI Studies in the Sciences of Complexity*, volume XII, pages 95–112. Addison-Wesley, 1992.

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.

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

H. Zou, J. Zhu, and T. Hastie. The margin vector, admissible loss and multi-class margin-based classifiers. Technical report, School of Statistics, University of Minnesota, 2006.