Combining Multi-class SVMs with Linear Ensemble Methods that Estimate the Class Posterior Probabilities

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

January 8, 2013

Running Title: Linear ensemble methods for M-SVMs

Keywords: multi-class support vector machines, ensemble methods, multivariate linear models, class posterior probability estimates, capacity control

Mathematics Subject Classification: 68T10, 62G08

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 the 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 each model has its advantages and drawbacks. These observations call for the evaluation of combinations of multi-class support vector machines. In this article, we study the combination of multi-class support vector machines with linear ensemble methods. Their sample complexity is low, which should prevent them from overfitting, and the outputs of two of them are estimates of the class posterior probabilities.

1 Introduction

The inferential principles of most of the models developed for pattern recognition share a common property: they do not change fundamentally with the number of categories. This is not the case for the support vector machines (SVMs). Initially, Vapnik and his co-authors devised a (1-norm) machine dedicated to the computation of dichotomies [5]. Although bi-class variants of this machine exist that exhibit appealing properties, such as the 2-norm SVM [5] or the least squares SVM [25], their use has remained marginal so far. The first studies dealing with the use of SVMs for multi-category classification report results obtained with decomposition methods [24] involving Vapnik's machine. Multi-class support vector machines (M-SVMs) were only introduced three years later [31].

During the last decade, many M-SVMs and decomposition methods involving bi-class SVMs have been introduced and evaluated (see [9, 11] for a survey). Currently, the attention of the community is focused on four main models of M-SVMs: the model of Weston and Watkins [31], the one of Crammer and Singer [6], the one of Lee, Lin, and Wahba [17] and the M-SVM^2 [12]. From an analytical point of view, their learning problems can be seen as straightforward extensions of those of bi-class SVMs. However, they exhibit distinct statistical properties (see for instance [19, 26] for analyses of their consistency). In recent years, several comparative studies between M-SVMs and decomposition methods have been published [8, 15]. 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. In accordance with what was predicted by the theory, the behaviours observed are different. Strangely enough, to the best of our knowledge, nobody has tried so far to take benefit of that phenomenon by combining different M-SVMs. Filling this void is the subject of this article. More precisely, we deal with the combination of M-SVMs with two requirements in mind: the outputs must be exploitable class posterior probability estimates and the sample complexity of the *combiners* must be low. The first requirement is motivated by the will to make the post-processing of the outputs easier. Indeed, it is well know that none of these machines produces outputs from which class posterior probability estimates can be derived straightforwardly. As for the second requirement, it simply stems from the fact that overfitting is one of the main limiting factors in the field of model combination.

Taking our inspiration from the works of Breiman and Friedman dealing with multivariate regression [4], we propose to combine the post-processed outputs of M-SVMs with linear ensemble methods. These methods are based on a multivariate linear model and differ with respect to their objective function. Their sample complexity can be upper bounded thanks to the use of a γ-Ψ-dimension [10] and two of them generate class posterior probability estimates. Their use requires to post-process beforehand the outputs of the machines so that they are nonnegative and sum to one. This is obtained by applying a polytomous logistic regression.

The organization of the paper is as follows. Section 2 introduces the M-SVMs through our generic model encompassing all of them. Section 3 is devoted to the definition and statistical analysis of the linear ensemble methods. Their implementation and assessment are addressed in Section 4. At last, we draw conclusions and outline our ongoing research in Section 5. To make reading easier, proofs have been gathered in appendix.

2 Multi-class support vector machines

The theoretical framework for the M-SVMs is the one of large margin multi-category classifiers [10]. It is summarized below.

2.1 Theoretical framework

We consider the case of Q-category classification problems with $Q \in \mathbb{N} \setminus [\![0, 2]\!]$. 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 $(\mathcal{X}, \mathcal{A})$ and $(\mathcal{Y}, \mathcal{B})$ are measurable spaces and the link between descriptions and categories can be characterized by an unknown probability measure P on the measurable space $(\mathcal{X} \times \mathcal{Y}, \mathcal{A} \otimes \mathcal{B})$. (X, Y) is a random pair with values in $\mathcal{X} \times \mathcal{Y}$, distributed according to P. We are given a class $\mathcal{G}$ of functions $g = (g_k)_{1 \leqslant k \leqslant Q}$ from $\mathcal{X}$ into $\mathbb{R}^Q$. By default, it is supposed to be of infinite cardinality. To each function g in $\mathcal{G}$ corresponds a decision rule d_g from $\mathcal{X}$ into $[\![1, Q]\!] \bigcup \{*\}$ defined as follows:

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

where $*$ denotes a dummy category introduced to deal with the cases of ex æquo. In that context, the learning problem consists in minimizing over $\mathcal{G}$ the *risk* $P\left(d_g\left(X\right) \neq Y\right)$. In practice, since P is unknown, the risk cannot be used directly as objective function. The optimization process, called *training*, makes use of a *training sample*, i.e., an m-sample $D_m = \left((X_i, Y_i)\right)_{1 \leqslant i \leqslant m}$ made up of independent copies of (X, Y), to infer knowledge on P.

2.2 Class of functions and learning problem

Defining a bi-class SVM is simple once the concept of *maximum margin hyperplane* [28] has been introduced. This initial linear separator can be turned into a nonlinear separator (in the description space), the hard margin SVM [3], by substituting in the formulas the Euclidean inner product by a kernel, i.e., a real-valued positive type function [1]. The model is then linear in the reproducing kernel Hilbert space (RKHS) [1] spanned by the kernel. Finally, tolerance to misclassifications is obtained by introducing slack variables in the constraints and the objective function of the learning problem. This last model is called the soft margin SVM [5]. Unfortunately, the concept of maximum margin hyperplane does not extend nicely to the multi-class case. One can find in this difficulty the main reason why among the different models of M-SVM, none is clearly favoured, and the first unifying definition of this family of machines, introduced with minor differences by several researchers (see for instance [33, 19, 10]), is recent. Its main drawback rests in the fact that it does not cover the class of *quadratic loss M-SVMs* [12]. In [11], we introduced the first generic model of M-SVM encompassing all the machines of this kind published so far. In the sequel, the M-SVMs are considered as instances of this model. To keep the article self-contained, the rest of the section is devoted to its presentation. Given κ, a real-valued positive type function on $\mathcal{X}^2$, a Q-category M-SVM with kernel κ operates on a vector space of $\mathbb{R}^Q$-valued functions: $\mathcal{H}_{\kappa,Q}$. This class is derived from another one, $\mathbf{H}_{\kappa,Q}$, which is endowed with a structure of RKHS of $\mathbb{R}^Q$-valued functions according to the definition provided in Section 6 of [30].

Definition 1 (RKHS of $\mathbb{R}^Q$-valued functions $\mathbf{H}_{\kappa,Q}$) *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 $\mathcal{X}^2$ and let $\tilde{\kappa}$ be the real-valued positive type function on $(\mathcal{X} \times [\![1, Q]\!])^2$ deduced from κ as follows:*

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

where δ is the Kronecker symbol. 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

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

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

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

Proposition 1 (Alternative characterization of $\mathbf{H}_{\kappa,Q}$) *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 $\mathcal{X}^2$ 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:*

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

Definition 2 (Class of functions $\mathcal{H}_{\kappa,Q}$) *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 $\mathcal{X}^2$ and let $\mathbf{H}_{\kappa,Q}$ be the class of functions derived from κ according to Definition 1. 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*

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

For all x in $\mathcal{X}$, κ_x denotes the element of $\mathbf{H}_\kappa$ such that for all x' in $\mathcal{X}$, $\kappa_x\left(x'\right) = \kappa\left(x, x'\right)$. For all h in $\mathcal{H}_{\kappa,Q}$, there exist $\bar{h} = \left(\bar{h}_k\right)_{1 \leqslant k \leqslant Q} \in \mathbf{H}_\kappa^Q$ and $b = \left(b_k\right)_{1 \leqslant k \leqslant Q} \in \mathbb{R}^Q$ such that

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

Thus, the functions in $\mathcal{H}_{\kappa,Q}$ can also be seen as multivariate affine functions on $\mathbf{H}_\kappa$. For $d_m = ((x_i, y_i))_{1 \leqslant i \leqslant m}$ in $(\mathcal{X} \times [\![1, Q]\!])^m$, $\mathbb{R}^{Qm}(d_m)$ denotes the subset of $\mathbb{R}^{Qm}$ made up of the vectors $v = (v_t)_{1 \leqslant t \leqslant Qm}$ satisfying:

$$\left(v_{(i-1)Q+y_i}\right)_{1 \leqslant i \leqslant m} = 0_m. \tag{2}$$

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,k) in $[\![1, m]\!] \times [\![1, Q]\!]$. Thus, (2) 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:

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

Our generic model of M-SVM is defined as follows.

Definition 3 (Generic model of M-SVM, Definition 4 in [11]) *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 1 and 2. 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 (QP) problem of the form*

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

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

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

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.

2.3 Discussion

The reformulations of the learning problems of the four main M-SVMs evoked in introduction as instances of Problem 1 can be found in [11]. Looking at this problem, it appears clearly that even though the concept of maximum margin hyperplane is no longer at the origin of the M-SVMs, the meaning of the values taken by their outputs remains geometrical. Indeed, denoting $H_{k,l}$ the hyperplane separating categories k and l in the RKHS $\mathbf{H}_\kappa$, for all x in $\mathcal{X}$, the distance between κ_x and $H_{k,l}$ is given by:

$$d\left(\kappa_x, H_{k,l}\right) = \frac{|h_k(x) - h_l(x)|}{\left\|\bar{h}_k - \bar{h}_l\right\|_{\mathbf{H}_\kappa}}.$$

In contrast, the values of the outputs do not provide directly class posterior probability estimates.

3 Linear ensemble methods

For all n in $\mathbb{N}^*$, let U_n be the unit $(n-1)$-simplex, i.e., $U_n = \left\{u = (u_p)_{1 \leqslant p \leqslant n} \in \mathbb{R}_+^n : \sum_{p=1}^n u_p = 1\right\}$. In this section, we make the hypothesis that N classifiers (functions from $\mathcal{X}$ into $\mathbb{R}^Q$) are available to perform the classification task of interest. Their outputs are supposed to be nonnegative and sum to one, so that they belong to U_Q. We first give a general description of the multivariate linear model (MLM) at the basis of the combiners considered. The linear ensemble methods (LEMs) are obtained by minimizing over this class of functions different objective functions derived from convex loss functions.

3.1 Multivariate linear model

Let $g^{(j)} = \left(g_k^{(j)}\right)_{1 \leqslant k \leqslant Q}$ be the j^{th} classifier. Let $\tilde{g}$ denote the function from $\mathcal{X}$ into U_Q^N obtained by appending the component functions of the classifiers $g^{(j)}$, i.e., $\tilde{g} = \left(g^{(j)}\right)_{1 \leqslant j \leqslant N}$. Precisely, $g_k^{(j)}$ is its component function of index $(j-1)Q + k$.

Definition 4 (Multivariate linear model) *We consider the multivariate linear model parameterized by the matrix $B \in \mathcal{M}_{Q,NQ}(\mathbb{R})$ such that*

$$\forall x \in \mathcal{X}, \;\; g_B(x) = B\tilde{g}(x)$$

$$s.t. \;\; \forall v \in U_Q^N, \;\; Bv \in U_Q.$$

This model generalizes the convex combination $g_{\Theta_c} = \sum_{j=1}^N \Theta_{c,j} g^{(j)}$ with $\Theta_c = (\Theta_{c,j})_{1 \leqslant j \leqslant N} \in U_N$. The transposes of the rows of B are denoted β_k, so that the model can be rewritten as:

$$\forall x \in \mathcal{X}, \; \forall k \in [\![1, Q]\!], \;\; (g_B)_k(x) = g_{\beta_k}(x) = \beta_k^T \tilde{g}(x). \tag{3}$$

We denote $\beta = (\beta_k)_{1 \leqslant k \leqslant Q} \in \mathbb{R}^{NQ^2}$ and use alternatively g_β to designate g_B. For the sake of interpretability, the general term of B (or β) is written with three indices, i.e., β_{kjl} (β_{kjl} is the component of vector β_k of index $(j-1)Q+l$). This provides a simple reformulation of (3):

$$\forall x \in \mathcal{X},\ \forall k \in [\![1, Q]\!],\ \ g_{\beta_k}(x) = \sum_{j=1}^{N} \sum_{l=1}^{Q} \beta_{kjl} g_l^{(j)}(x).$$

The convex combination is the degenerate case obtained by setting $\beta_{kjl} = \Theta_{c,j}\delta_{k,l}$. We define:

$$\forall\, (k,j) \in [\![1, Q]\!] \times [\![1, N]\!],\ \ \beta'_{kj} = \min_{1 \leqslant l \leqslant Q} \beta_{kjl}.$$

For all n in $\mathbb{N}^*$, let 1_n be the vector of $\mathbb{R}^n$ whose components are all equal to 1. The constraint $\forall v \in U_Q^N,\ \ Bv \in U_Q$ defines a convex polytope in $\mathbb{R}^{NQ^2}$.

Proposition 2 *The expression of the system of constraints of the MLM as a function of the components of matrix B (vector β) is:*

$$\begin{cases} \forall k \in [\![1, Q]\!],\ \ \sum_{j=1}^{N} \beta'_{kj} \geqslant 0 \\ \forall j \in [\![1, N]\!],\ \forall l \in [\![1, Q-1]\!],\ \ \sum_{k=1}^{Q} (\beta_{kjl} - \beta_{kjQ}) = 0 \\ 1_{NQ^2}^T \beta = Q \end{cases}.$$

3.2 Generic definition of a linear ensemble method

We consider LEMs corresponding to choosing the matrix B as a sample-based minimizer of a convex risk functional subject to the constraints of Proposition 2.

Definition 5 (Linear ensemble method) *Given a convex loss function ℓ_{LEM}, a* linear ensemble method *trained on d_m is an instance of the MLM whose matrix of parameters, B^*, is obtained by solving the following convex programming problem:*

Problem 2

$$\min_B \sum_{i=1}^{m} \ell_{LEM}\left(\tilde{g}(x_i), y_i, B\right)$$
$$s.t.\ \ \forall v \in U_Q^N,\ \ Bv \in U_Q.$$

Remark 1 *Problem 2 is underdetermined. This is due to the fact that the predictors $g_l^{(j)}(x)$ are linearly dependent. Let $\left(U_Q^N\right)^\perp = \left\{ w \in \mathbb{R}^{NQ} : \forall v \in U_Q^N,\ \ w^T v = 0 \right\}$. If β^* is an optimal solution of Problem 2 and $\gamma \in \left\{ \left(U_Q^N\right)^\perp \right\}^Q$, then $\beta^* + \gamma$ is also an optimal solution of Problem 2.*

Since we are only looking for one optimal solution of Problem 2, the linear dependency of the predictors can be turned into an advantage by making use of Remark 1 to apply a restriction on the feasible region keeping the quality of the optima unchanged while simplifying computation. The restriction systematically considered in the sequel is characterized by Proposition 3.

Proposition 3 *Irrespective of the nature of ℓ_{LEM}, there is an optimal solution of Problem 2 which belongs to the nonnegative hyperorthant, i.e., to the convex polytope $V_{N,Q}$ given by:*

$$\begin{cases} \beta \in \mathbb{R}_+^{NQ^2} \\ \forall j \in [\![1, N]\!],\ \forall l \in [\![1, Q-1]\!],\ \sum_{k=1}^{Q} (\beta_{kjl} - \beta_{kjQ}) = 0 \\ 1_{NQ^2}^T \beta = Q \end{cases} .$$

An additional benefit of this restriction is that the set of constraints is directly expressed in *standard form*. Above all, it makes it possible to characterize the LEMs as implementing a two-level weighting of the predictors. This weighting is defined by Proposition 4.

Proposition 4 (Alternative characterization of $V_{N,Q}$) *A vector β in $\mathbb{R}^{NQ^2}$ belongs to $V_{N,Q}$ if and only if there exists a vector $\Theta = (\Theta_j)_{1 \leqslant j \leqslant N}$ in U_N and a vector θ in $[0,1]^{NQ^2}$ satisfying*

$$\forall (j,l) \in [\![1, N]\!] \times [\![1, Q]\!],\ (\theta_{kjl})_{1 \leqslant k \leqslant Q} \in U_Q$$

such that

$$\forall (j,l) \in [\![1, N]\!] \times [\![1, Q]\!],\ (\beta_{kjl})_{1 \leqslant k \leqslant Q} = \Theta_j (\theta_{kjl})_{1 \leqslant k \leqslant Q} .$$

This proposition highlights the difference between an LEM (with β in $V_{N,Q}$) and a convex combination. With a convex combination, each predictor $g_l^{(j)}(x)$ gives its "vote" to the corresponding category ($\theta_{kjl} = \delta_{k,l}$), and this vote is weighted by the weight of the corresponding classifier, $\Theta_{c,j}$. With an LEM, each predictor can split its vote between the different categories ($(\theta_{kjl})_{1 \leqslant k \leqslant Q} \in U_Q$). The introduction of this degree of freedom can affect the weighting of the classifiers, i.e., for a given loss function, one can have $\Theta^* \neq \Theta_c^*$.

3.3 Estimation of the class posterior probabilities

In this subsection, we focus on two natural choices for the loss function ℓ_{LEM} that give rise to class posterior probability estimates. For all k in $[\![1, Q]\!]$, let t_k denote the *one of Q coding* of category k, i.e., $t_k = (\delta_{k,l})_{1 \leqslant l \leqslant Q}$. The quadratic loss ℓ_{Quad} is defined as:

$$\forall (x, y, \beta) \in \mathcal{X} \times \mathcal{Y} \times V_{N,Q},\ \ \ell_{\text{Quad}} (\tilde{g}(x), y, \beta) = \frac{1}{2} \left\| t_y - g_\beta (x) \right\|_2^2 .$$

Let $\tilde{G}$ be the matrix of $\mathcal{M}_{m,NQ}(\mathbb{R})$ such that its row of index i is the transpose of the vector $\tilde{g}(x_i)$. For all k in $[\![1, Q]\!]$, let $\mathbf{y}_k = (\delta_{y_i,k})_{1 \leqslant i \leqslant m} \in \{0,1\}^m$ and let $\mathbf{y} = (\mathbf{y}_k)_{1 \leqslant k \leqslant Q} \in \{0,1\}^{Qm}$. The objective function corresponding to the quadratic loss is given by:

$$J_{\text{Quad}}(\beta) = \frac{1}{2}\beta^T \left\{ I_Q \otimes \left(\tilde{G}^T \tilde{G} \right) \right\} \beta - \left\{ \mathbf{y}^T \left(I_Q \otimes \tilde{G} \right) \right\} \beta$$

where $\otimes$ denotes the Kronecker product. The minimization of $J_{\text{Quad}}(\beta)$ subject to $\beta \in V_{N,Q}$ is a convex QP problem. The expression of the cross-entropy loss ℓ_{CE} is:

$$\forall (x, y, \beta) \in \mathcal{X} \times \mathcal{Y} \times V_{N,Q}, \ \ \ell_{\text{CE}}(\tilde{g}(x), y, \beta) = -\sum_{k=1}^{Q} \delta_{y,k} \ln \left(g_{\beta_k}(x) \right).$$

The expression of the corresponding objective function is

$$J_{\text{CE}}(\beta) = -\sum_{i=1}^{m} \sum_{k=1}^{Q} \delta_{y_i,k} \ln \left(\beta_k^T \tilde{g}(x_i) \right).$$

The minimization of $J_{\text{CE}}(\beta)$ subject to $\beta \in V_{N,Q}$ is a convex programming problem.

To specify the way those two LEMs estimate the class posterior probabilities, we need to introduce additional notations. For all x in $\mathcal{X}$, let P_x be the probability measure on $\mathcal{Y}$ given by:

$$\forall k \in [\![1, Q]\!], \ \ P_x(k) = P(k \mid x).$$

For all x in $\mathcal{X}$ and all β in $V_{N,Q}$, let $P_{x,\beta}$ be the probability measure on $\mathcal{Y}$ given by:

$$\forall k \in [\![1, Q]\!], \ \ P_{x,\beta}(k) = g_{\beta_k}(x).$$

Furthermore, let D_{KL} denote the Kullback-Leibler divergence. The aforementioned specification is provided by the following proposition.

Proposition 5 *Irrespective of the nature of ℓ_{LEM}, let $\beta^*(m) = (\beta_k^*(m))_{1 \leqslant k \leqslant Q}$ in $V_{N,Q}$ be an optimal solution of Problem 2 with D_m as training sample. Then if the loss function is the quadratic one,*

$$\mathbb{E}_X \left\{ \sum_{k=1}^{Q} \left[P(k \mid X) - g_{\beta_k^*(m)}(X) \right]^2 \right\} \xrightarrow[m \longrightarrow +\infty]{P} \inf_{\beta \in V_{N,Q}} \mathbb{E}_X \left\{ \sum_{k=1}^{Q} \left[P(k \mid X) - g_{\beta_k}(X) \right]^2 \right\}, \tag{4}$$

whereas with the cross-entropy loss we get

$$\mathbb{E}_X \left[D_{KL}\left(P_X \parallel P_{X,\beta^*(m)} \right) \right] \xrightarrow[m \longrightarrow +\infty]{P} \inf_{\beta \in V_{N,Q}} \mathbb{E}_X \left[D_{KL}\left(P_X \parallel P_{X,\beta} \right) \right]. \tag{5}$$

The proof of Proposition 5 is inspired from the proofs of similar results obtained for neural networks (see for instance [22]). There are however two fundamental differences regarding the asymptotic behaviour. On the one hand, the classes of functions at the basis of the neural networks considered are universal approximators [13], unlike the MLM. This means that in the first case, the *approximation error* could be null, whereas in the second case, it should be positive. On the other hand, an advantage of the LEMs over the aforementioned neural networks, for which the training algorithm may get stuck in local (suboptimal) minima, is that since Problem 2 is a convex programming problem, the training procedure systematically produces an optimal solution. In other words, the *estimation error* should asymptotically be null. This analysis calls for a justification of the choice of an LEM. As was pointed out in the introduction, it rests on our concern to devise combiners of low capacity, with the aim to avoid overfitting. We conjecture that in many practical cases of pattern recognition, the capacity of a combiner should be superior to that of a simple convex combination and inferior to that of a neural network (or an M-SVM).

The quality of the class posterior probability estimates is primarily governed by three factors: the number of classifiers combined, the nature of their outputs, and the correlation of their errors. Obviously, this quality will be all the better as the predictors $g_k^{(j)}(x)$ are themselves good estimates of those probabilities. This should be taken into account when processing the outputs of the M-SVMs so as to obtain vectors in U_Q (see Section 4).

3.4 Model selection with the ℓ_1 norm

The main advantage of the choice of the ℓ_1 norm in place of the ℓ_2 one is well known in machine learning: it leads to sparse solutions. In the framework of our study, instantiating ℓ_{LEM} with

$$\forall (x, y, \beta) \in \mathcal{X} \times \mathcal{Y} \times V_{N,Q}, \;\; \ell_{\ell_1}\left(\tilde{g}(x), y, \beta\right) = \left\|t_y - g_\beta\left(x\right)\right\|_1$$

produces the following linear programming (LP) problem:

Problem 3

$$\min_{\beta} \left\{ - \left[\sum_{i=1}^{m} \left(t_{y_i} \otimes \tilde{g}\left(x_i\right)\right) \right]^T \beta \right\}$$
$$s.t. \;\; \beta \in V_{N,Q}.$$

The sparsity of the optimal solutions of Problem 3 can be characterized exactly. This calls for the characterization of the extreme points of $V_{N,Q}$.

Lemma 1 (Extreme point of $V_{N,Q}$) *A vector β in $V_{N,Q}$ is an extreme point of $V_{N,Q}$ if and only if*

$$\exists j_0 \in [\![1, N]\!]: \begin{cases} \forall l \in [\![1, Q]\!], \ \ \exists k_0(l) \in [\![1, Q]\!]: (\beta_{kj_0l})_{1 \leqslant k \leqslant Q} = t_{k_0(l)} \\ \forall j \in [\![1, N]\!] \setminus \{j_0\}, \ \forall (k, l) \in [\![1, Q]\!]^2, \ \ \beta_{kjl} = 0 \end{cases}. \tag{6}$$

Remark 2 *A vector β_0 in $V_{N,Q}$ is an extreme point of $V_{N,Q}$ if and only if*

$$\|\beta_0\|_2 = \max_{\beta \in V_{N,Q}} \|\beta\|_2 = \sqrt{Q}.$$

The following proposition is a direct consequence of Lemma 1.

Proposition 6 *There exists an optimal solution $\beta^* = (\beta_k^*)_{1 \leqslant k \leqslant Q}$ of Problem 3 that can be characterized as follows: there exists j_0 in $[\![1, N]\!]$ and a map k_0 from $[\![1, Q]\!]$ to itself such that*

$$\forall k \in [\![1, Q]\!], \ \ g_{\beta_k^*} = \sum_{l: k_0(l) = k} g_l^{(j_0)}. \tag{7}$$

Corollary 1 *Let us consider the LEM whose learning problem is specified by Problem 3. Keeping the notations of Proposition 6, it appears that except in pathological cases, the map k_0 should be the identity, with the consequence that $g_{\beta^*} = g^{(j_0)}$: the LEM is in fact a model selection method.*

3.5 Sample complexity of the linear ensemble methods

To compute the sample complexity of the LEMs, we derive an upper bound on the capacity of the MLM. To that end, we consider an extended definition of this model, corresponding to changing its domain for U_Q^N. Indeed, the bound only depends on the predictor vector through the domain in which it takes its value (irrespective of the nature of its components). The extension of Definition 4 is thus:

$$\forall \beta = (\beta_k)_{1 \leqslant k \leqslant Q} \in V_{N,Q}, \ \ g_\beta : \begin{array}{l} U_Q^N \longrightarrow U_Q \\ v \mapsto g_\beta(v) = \left(\beta_k^T v\right)_{1 \leqslant k \leqslant Q}. \end{array}$$

In [10], we enriched the Vapnik-Chervonenkis (VC) theory of large margin multi-category classifiers by proving that for the classes of functions at the basis of these classifiers, the appropriate generalizations of the standard capacity measure of the binary models, the VC dimension [29], are the γ-Ψ-dimensions. Their use is based on the application of *margin operators.* The operator needed to characterize the capacity of the MLM is the Δ one.

Definition 6 (Δ operator) *Let $\mathcal{G}$ be a class of functions on a set $\mathcal{X}$ taking their values in $\mathbb{R}^Q$. Δ is defined as an operator on $\mathcal{G}$ such that:*

$$\begin{aligned}\Delta: \ & \mathcal{G} \longrightarrow \Delta\mathcal{G} \\ & g \mapsto \Delta g = ((\Delta g)_k)_{1 \leqslant k \leqslant Q} \end{aligned}$$

$$\forall x \in \mathcal{X}, \ \ \Delta g(x) = \frac{1}{2}\left(g_k(x) - \max_{l \neq k} g_l(x)\right)_{1 \leqslant k \leqslant Q}.$$

The γ-Ψ-dimension used is a scale-sensitive extension of the Natarajan dimension [20].

Definition 7 (Natarajan dimension with margin γ) *Let $\mathcal{G}$ be a class of functions on a set $\mathcal{X}$ taking their values in $\mathbb{R}^Q$. For γ in $\mathbb{R}_+^*$, a subset $s_{\mathcal{X}^n} = \{x_i : 1 \leqslant i \leqslant n\}$ of $\mathcal{X}$ is said to be* γ-N-shattered *by $\Delta\mathcal{G}$ if there is a set $I(s_{\mathcal{X}^n}) = \{(i_1(x_i), i_2(x_i)) : 1 \leqslant i \leqslant n\}$ of n pairs of distinct indices in $[\![1, Q]\!]$ and a vector $\mathbf{c} = (c_i)_{1 \leqslant i \leqslant n}$ in $\mathbb{R}^n$ such that, for each vector $\mathbf{y} = (y_i)_{1 \leqslant i \leqslant n}$ in $\{-1, 1\}^n$, there is a function $g_{\mathbf{y}}$ in $\mathcal{G}$ satisfying*

$$\forall i \in [\![1, n]\!], \ \begin{cases} \textit{if } y_i = \ \ 1, & (\Delta g_{\mathbf{y}})_{i_1(x_i)}(x_i) - c_i \geqslant \gamma \\ \textit{if } y_i = -1, & (\Delta g_{\mathbf{y}})_{i_2(x_i)}(x_i) + c_i \geqslant \gamma \end{cases}.$$

The Natarajan dimension with margin γ *of the class $\Delta\mathcal{G}$, N-dim$(\Delta\mathcal{G}, \gamma)$, is the maximal cardinality of a subset of $\mathcal{X}$ γ-N-shattered by $\Delta\mathcal{G}$, if this cardinality is finite. If no such maximum exists, $\Delta\mathcal{G}$ is said to have infinite Natarajan dimension with margin γ.*

Let $\mathcal{G}_\beta = \{g_\beta : \beta \in V_{N,Q}\}$. An upper bound on N-dim$(\Delta\mathcal{G}_\beta, \gamma)$ is provided by Theorem 1.

Theorem 1 (Upper bound on the capacity of the MLM)

$$\forall \gamma \in \left(0, \frac{1}{2}\right], \ \ \textit{N-dim}\left(\Delta\mathcal{G}_\beta, \gamma\right) \leqslant \binom{Q}{2}\frac{NQ}{4\gamma^2}. \tag{8}$$

For $\gamma > \frac{1}{2}$, N-dim$\left(\Delta\mathcal{G}_\beta, \gamma\right) = 0$. Theorem 1 thus deals with the nontrivial case. In conjunction with Theorem 4.1 in [10], it can be used to optimize the split of the training set into data used to train the M-SVMs, their post-processing, and the selected LEM. It is noteworthy that for the different models of interest, a guaranteed risk can be obtained that involves an alternative measure of capacity: the Rademacher average [7].

4 Implementation and assessment of the linear ensemble methods

Since the M-SVMs do not take their values in U_Q, their outputs must be post-processed prior to being combined. This post-processing is all the more important as it performs a

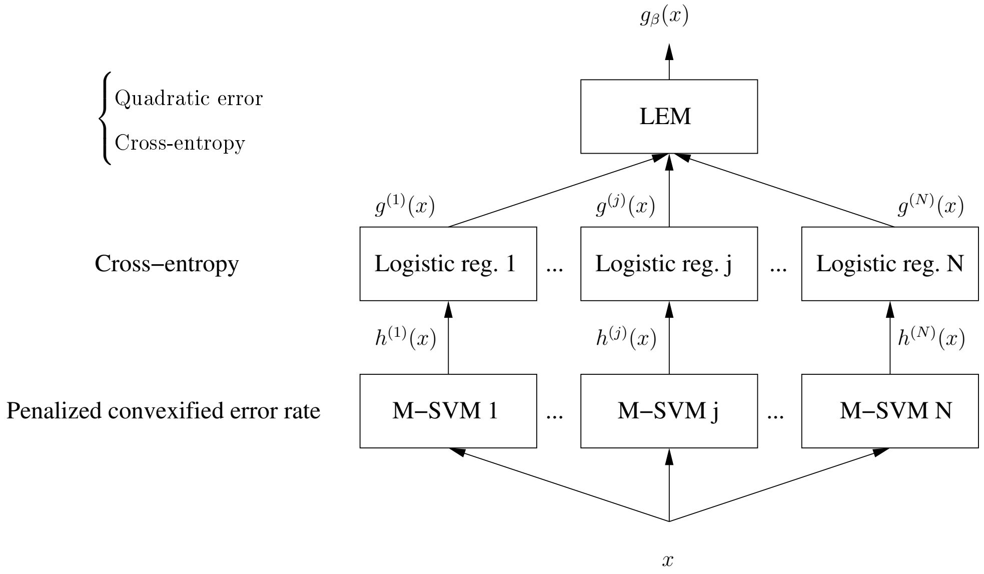

Figure 1: Flowchart of the computation of the outputs of an LEM. For each level, the set of possible training criteria is mentioned on the left.

transition between models based on different notions of risk. The objective function of an M-SVM is more directly related to the ultimate criterion, the recognition rate, than the objective function of an LEM (using the quadratic loss or the cross-entropy loss) with the problematic consequence that there is no guarantee that the combination should improve this rate. This calls for the choice of a post-processing maximizing the correlation between the behaviours assessed by means of the different measures of prediction accuracy involved. The solution we propose is a variant of the polytomous logistic regression model [14]. Thus, the functional dependency between an input $x \in \mathcal{X}$ and the corresponding output of an LEM, $g_\beta(x) \in U_Q$, is represented by Figure 1.

The predictors of the LEMs take the form

$$\forall j \in [\![1, N]\!],\ \forall k \in [\![1, Q]\!],\ \forall x \in \mathcal{X},\ \ g_k^{(j)}(x) = \frac{\exp\left(a_{jk} h_k^{(j)}(x) + b_{jk}\right)}{\sum_{l=1}^{Q} \exp\left(a_{jl} h_l^{(j)}(x) + b_{jl}\right)}.$$

This formulation is used to emphasize the fact that the classifiers $g^{(j)}$ actually take their values in U_Q. It is noteworthy that this model can also be seen as a multivariate extension of Platt's model [21] consisting in fitting a sigmoid after a bi-class SVM. A maximum likelihood estimation of the N models is performed, thanks to an extension of the training algorithm introduced in [18]. The outputs $g_k^{(j)}(x)$ are initial estimates of the class posterior probabilities, which is a useful feature given the specifications of the LEMs. The learning problems of the two LEMs introduced in Section 3.3 are solved by means of a variant of the gradient projection method [23].

An evaluation of the LEMs can be found in [2]. This comparative study of the accuracy of several combiners focuses on two criteria: the recognition rate and the quality of the class posterior probability estimates. LEMs are also used as *structure-to-structure classifiers* in our method of protein secondary structure prediction [27]. In this hybrid architecture inspired from [16], discriminant models organized in cascade generate estimates of the class posterior probabilities from which the emission probabilities of a generative model are derived. Thus, this application in structural biology provides another evaluation of the LEMs with respect to the aforementioned criteria.

5 Conclusions and ongoing research

In this article, a class of linear ensemble methods devoted to the combination, after an appropriate post-processing, of Q-category classifiers taking their values in $\mathbb{R}^Q$, has been introduced. Their specifications should make them well suited for the combination of M-SVMs. Indeed, the corresponding loss function can be chosen so that the outputs are class posterior probability estimates and their low sample complexity should prevent them from overfitting.

We are currently performing a large scale comparative study of the performance of the LEMs. The focus is laid on the quality of the probability estimates. The application on problems involving large numbers of categories and base classifiers, i.e., large numbers of parameters, calls for the design of dedicated training algorithms. From a theoretical point of view, it should be instructive to carry on this study by comparing the combinations of M-SVMs we consider with the extension of the M-SVM of Lee and co-authors to a Bayesian model described in [32].

Acknowledgements The author would like to thank the anonymous reviewers for their comments.

A Proofs of the main results of the article

A.1 Proof of Proposition 2

The proof of Proposition 2 is made up of three steps.

1. $\forall v \in U_Q^N, \;\; Bv \in \mathbb{R}_+^Q$

   $\forall k \in [\![1, Q]\!]$, $\forall v \in U_Q^N$, $\beta_k^T v = \sum_{j=1}^N \sum_{l=1}^Q \beta_{kjl} v_{(j-1)Q+l}$. Given the vector β_k, the minimum of this inner product is obtained for a vector v satisfying:

   $$\forall j \in [\![1, N]\!], \quad \sum_{l : \beta_{kjl} = \beta'_{kj}} v_{(j-1)Q+l} = 1.$$

   As a consequence, $\min_{v \in U_Q^N} \beta_k^T v = \sum_{j=1}^N \beta'_{kj}$, from which it springs that $\forall k \in [\![1, Q]\!], \;\; \sum_{j=1}^N \beta'_{kj} \geqslant 0$ is a necessary and sufficient condition of nonnegativity of the outputs.

2. $\exists K : \forall v \in U_Q^N, \;\; 1_Q^T Bv = K$

   To derive the corresponding constraints, it suffices to notice that given any two vectors $v^{(0)}$ and w in U_Q^N, one can generate a finite sequence $\left(v^{(n)}\right)_{1 \leqslant n \leqslant n^*}$ of vectors in U_Q^N such that $v^{(n^*)} = w$ and $v^{(n+1)}$ is deduced from $v^{(n)}$ by applying an elementary step of the form:

   (a) choose $(j, l_1, l_2) \in [\![1, N]\!] \times [\![1, Q]\!] \times [\![1, Q]\!]$ such that $l_1 \neq l_2$, $v_{(j-1)Q+l_1}^{(n)} < 1$ and $v_{(j-1)Q+l_2}^{(n)} > 0$;

   (b) choose $\delta \in \mathbb{R}_+^*$ satisfying $v_{(j-1)Q+l_1}^{(n)} + \delta \leqslant 1$ and $v_{(j-1)Q+l_2}^{(n)} - \delta \geqslant 0$;

   (c) set $v^{(n+1)}$ equal to $v^{(n)}$ except for its components of indices $(j-1)Q + l_1$ and $(j-1)Q + l_2$ which are respectively set to $v_{(j-1)Q+l_1}^{(n)} + \delta$ and to $v_{(j-1)Q+l_2}^{(n)} - \delta$.

   Given the generative algorithm detailed above, keeping the sum $1_Q^T Bv$ constant over the whole set U_Q^N boils down to ensuring that this sum does not vary when an elementary step is applied, so that

   $$\exists K \in \mathbb{R} : \forall v \in U_Q^N, \;\; 1_Q^T Bv = K \Longleftrightarrow$$

   $$\forall \, (j, l_1, l_2) \in [\![1, N]\!] \times [\![1, Q]\!] \times [\![1, Q]\!], \;\; \sum_{k=1}^Q (\beta_{kjl_1} - \beta_{kjl_2}) = 0 \Longleftrightarrow$$

   $$\forall j \in [\![1, N]\!], \; \forall l \in [\![1, Q-1]\!], \;\; \sum_{k=1}^Q (\beta_{kjl} - \beta_{kjQ}) = 0.$$

3. $K = 1$

   Once the conditions of the second step of the proof are satisfied, $1_Q^T Bv$ does not depend on v anymore. Thus, the constraint corresponding to $K = 1$ can be derived from an arbitrary choice of v in U_Q^N. Setting $v = \frac{1}{Q}1_{NQ}$ gives:

$$K = 1 \iff \frac{1}{Q}\sum_{k=1}^{Q}\sum_{j=1}^{N}\sum_{l=1}^{Q}\beta_{kjl} = 1 \iff 1_{NQ^2}^T\beta = Q.$$

The conjunction of this constraint and the ones obtained at the second step provides us with a stronger result that will prove useful in the sequel:

$$\forall l \in [\![1, Q]\!],\ \sum_{k=1}^{Q}\sum_{j=1}^{N}\beta_{kjl} = 1. \tag{9}$$

A.2 Proof of Proposition 3

With Remark 1 in mind, to prove Proposition 3, it suffices to establish that for all vector β satisfying the constraints of Proposition 2, one can exhibit a vector γ in $\left\{\left(U_Q^N\right)^{\perp}\right\}^Q$ such that $\beta + \gamma \in \mathbb{R}_+^{NQ^2}$. For β satisfying the constraints of Proposition 2 and $k \in [\![1, Q]\!]$, let I_{k-} and I_{k+} be the subsets of $[\![1, N]\!]$ such that $I_{k-} = \left\{j \in [\![1, N]\!] : \beta'_{kj} < 0\right\}$ and $I_{k+} = [\![1, N]\!] \setminus I_{k-}$. The vector $\gamma^* = (\gamma_k^*)_{1\leqslant k\leqslant Q}$ defined as follows:

$$\forall k \in [\![1, Q]\!],\ \forall j_0 \in [\![1, N]\!],\ \forall l \in [\![1, Q]\!],\ \begin{cases}\text{if } j_0 \in I_{k-},\ \ \gamma_{kj_0l}^* = -\beta'_{kj_0}\\ \text{if } j_0 \in I_{k+},\ \ \gamma_{kj_0l}^* = \beta'_{kj_0}\frac{\sum_{j\in I_{k-}}\beta'_{kj}}{\sum_{j\in I_{k+}}\beta'_{kj}}\end{cases}$$

with $\gamma_{kj_0l}^*$ being the component of index $(j_0 - 1)\,Q + l$ of $\gamma_k^* \in \mathbb{R}^{NQ}$, meets the aforementioned requirements.

A.3 Proof of Proposition 4

A possible construction of vectors Θ and θ corresponding to a given vector β in $V_{N,Q}$ is:

$$\forall j \in [\![1, N]\!],\ \Theta_j = \sum_{k=1}^{Q}\beta_{kjQ}$$

and

$$\forall\,(j, l) \in [\![1, N]\!]\times[\![1, Q]\!],\ \begin{cases}\text{if } \Theta_j > 0,\ \ (\theta_{kjl})_{1\leqslant k\leqslant Q} = \Theta_j^{-1}\,(\beta_{kjl})_{1\leqslant k\leqslant Q}\\ \text{if } \Theta_j = 0,\ \ (\theta_{kjl})_{1\leqslant k\leqslant Q} = \frac{1}{Q}1_Q\end{cases}.$$

Indeed, we deduce from (9) that

$$\sum_{j=1}^{N}\Theta_j = \sum_{j=1}^{N}\sum_{k=1}^{Q}\beta_{kjQ} = 1,$$

which implies that $\Theta \in U_N$. Proving that the vectors $(\theta_{kjl})_{1\leqslant k\leqslant Q}$ belong to U_Q is also straightforward. Conversely, if two vectors Θ and θ satisfy the hypotheses of Proposition 4, then obviously the vector β of general term $\beta_{kjl} = \Theta_j\theta_{kjl}$ belongs to $V_{N,Q}$.

A.4 Proof of Proposition 5

Given the hypotheses made in Section 2.1,

$$2\mathbb{E}_{(X,Y)}\left[\ell_{\text{Quad}}\left(\tilde{g}\left(X\right), Y, \beta\right)\right] = \int_{\mathcal{X}\times\mathcal{Y}} \left\|t_y - g_\beta\left(x\right)\right\|_2^2 dP\left(x, y\right),$$

and Fubini's theorem for nonnegative measurable functions can be applied to the measure P. This gives:

$$2\mathbb{E}_{(X,Y)}\left[\ell_{\text{Quad}}\left(\tilde{g}\left(X\right), Y, \beta\right)\right] = \mathbb{E}_X\left\{\sum_{k=1}^{Q}\left\|t_k - g_\beta\left(X\right)\right\|_2^2 P\left(k \mid X\right)\right\}$$

$$= \mathbb{E}_X\left\{\sum_{k=1}^{Q}\sum_{l=1}^{Q}\left[\delta_{k,l}^2 - 2\delta_{k,l}g_{\beta_l}\left(X\right) + g_{\beta_l}\left(X\right)^2\right] P\left(k \mid X\right)\right\}$$

$$= 1 - 2\mathbb{E}_X\left\{\sum_{k=1}^{Q} g_{\beta_k}\left(X\right) P\left(k \mid X\right)\right\} + \mathbb{E}_X\left\{\sum_{k=1}^{Q} g_{\beta_k}\left(X\right)^2\right\}$$

$$= 1 - \mathbb{E}_X\left\{\sum_{k=1}^{Q} P\left(k \mid X\right)^2\right\} + \mathbb{E}_X\left\{\sum_{k=1}^{Q}\left[P\left(k \mid X\right) - g_{\beta_k}\left(X\right)\right]^2\right\}.$$

Thus,

$$\underset{\beta\in V_{N,Q}}{\text{argmin}}\,\mathbb{E}_{(X,Y)}\left[\ell_{\text{Quad}}\left(\tilde{g}\left(X\right), Y, \beta\right)\right] = \underset{\beta\in V_{N,Q}}{\text{argmin}}\,\mathbb{E}_X\left\{\sum_{k=1}^{Q}\left[P\left(k \mid X\right) - g_{\beta_k}\left(X\right)\right]^2\right\}. \tag{10}$$

Given (10), (4) simply expresses the consistency of the principle of empirical risk minimization for the class of functions $\left\{\ell_{\text{Quad}}\left(\cdot, \cdot, \beta\right) : \beta \in V_{N,Q}\right\}$ (see Chapters 3 and 5 in [29]). Thus, to finish the proof of (4), it suffices to establish this consistency. Without going into particulars, this can be done by proving the finiteness of the capacity of the MLM. A result of this kind is provided by Theorem 1.

$$\mathbb{E}_{(X,Y)}\left[\ell_{\text{CE}}\left(\tilde{g}\left(X\right), Y, \beta\right)\right] = -\int_{\mathcal{X}\times\mathcal{Y}}\sum_{k=1}^{Q}\delta_{y,k}\ln\left(g_{\beta_k}\left(x\right)\right) dP\left(x, y\right).$$

Applying Fubini's theorem for nonnegative measurable functions to the measure P gives:

$$\mathbb{E}_{(X,Y)}\left[\ell_{\text{CE}}\left(\tilde{g}\left(X\right), Y, \beta\right)\right] = -\mathbb{E}_X\left[\sum_{k=1}^{Q} P\left(k \mid X\right)\ln\left(g_{\beta_k}\left(X\right)\right)\right]$$

$$= -\mathbb{E}_X\left[\sum_{k=1}^{Q} P\left(k \mid X\right)\ln\left(\frac{g_{\beta_k}\left(X\right)}{P\left(k \mid X\right)}\right)\right] - \mathbb{E}_X\left[\sum_{k=1}^{Q} P\left(k \mid X\right)\ln\left(P\left(k \mid X\right)\right)\right]$$

$$= \mathbb{E}_X\left[D_{\text{KL}}\left(P_X \parallel P_{X,\beta}\right)\right] - \mathbb{E}_X\left[\sum_{k=1}^{Q} P\left(k \mid X\right)\ln\left(P\left(k \mid X\right)\right)\right].$$

Thus,

$$\underset{\beta \in V_{N,Q}}{\operatorname{argmin}} \mathbb{E}_{(X,Y)} \left[\ell_{\mathrm{CE}} \left(\tilde{g}\left(X\right), Y, \beta \right) \right] = \underset{\beta \in V_{N,Q}}{\operatorname{argmin}} \mathbb{E}_X \left[D_{\mathrm{KL}} \left(P_X \parallel P_{X,\beta} \right) \right].$$

With this equation at hand, the proof of (5) can be completed by using the line of argument used to complete the proof of (4).

A.5 Proof of Lemma 1

The simplest way to establish Lemma 1 consists in making use of the alternative representation of the vectors of $V_{N,Q}$ introduced in Proposition 4. Then, the proof is made up of two proofs by contradiction. Since they present no difficulty, we only give the sketch of the reasoning. The first proof by contradiction consists in establishing that if β is an extreme point of $V_{N,Q}$, then necessarily Θ is an extreme point of U_N, which implies that

$$\exists j_0 \in [\![1, N]\!] : \begin{cases} \Theta_{j_0} = 1 \\ \forall j \in [\![1, N]\!] \setminus \{j_0\}, \ \forall (k, l) \in [\![1, Q]\!]^2, \ \ \beta_{kjl} = 0 \end{cases}.$$

The second proof by contradiction establishes that for all value of l in $[\![1, Q]\!]$, the vector $\left(\theta_{kj_0l}\right)_{1 \leqslant k \leqslant Q}$ must be an extreme point of U_Q. This means that:

$$\forall l \in [\![1, Q]\!], \ \ \exists k_0\left(l\right) \in [\![1, Q]\!] : \left(\theta_{kj_0l}\right)_{1 \leqslant k \leqslant Q} = t_{k_0(l)}.$$

Combining the two partial results and using the fact that for all (k, l) in $[\![1, Q]\!]^2$, $\beta_{kj_0l} = \theta_{kj_0l}$, we get (6), which concludes the proof.

A.6 Proof of Proposition 6

Since Problem 3 is an LP problem in standard form, there exists an extreme point of the feasible region which is an optimal solution. Thus, (7) simply corresponds to a restriction of (6) to the extreme points of $V_{N,Q}$ that are also optimal solutions of Problem 3.

A.7 Proof of Theorem 1

The proof of Theorem 1 follows the sketch of the proof of Theorem 4.1 in [10]. In the same way as this proof, it is based on two lemmas.

Lemma 2 *Let $\gamma \in \left(0, \frac{1}{2}\right]$ and $n \in \mathbb{N}^*$. If a subset $s_n = \{v_i : 1 \leqslant i \leqslant n\}$ of U_Q^N is N-shattered with margin γ by $\Delta\mathcal{G}_\beta$, then there exists a subset s_p of s_n of cardinality p equal* *to* $\left\lceil \frac{n}{\binom{Q}{2}} \right\rceil$ *such that for every split of* s_p *into two subsets* $s_{p,1}$ *and* $s_{p,2}$*, the following bound holds true:*

$$\left\| \sum_{v_i \in s_{p,1}} v_i - \sum_{v_i \in s_{p,2}} v_i \right\|_2 \geqslant \frac{2 \left\lceil \frac{n}{\binom{Q}{2}} \right\rceil}{\sqrt{Q}} \gamma.$$

Proof Suppose that $s_n = \{v_i : 1 \leqslant i \leqslant n\}$ is a subset of U_Q^N N-shattered with margin γ by $\Delta\mathcal{G}_\beta$. Let $(I(s_n), \mathbf{c})$ witness this shattering. Without loss of generality, we can assume that $I(s_n)$ satisfies: $\forall i \in [\![1, n]\!]$, $i_1(v_i) < i_2(v_i)$. According to the pigeonhole principle, there is at least one couple of indices (k_1, k_2) with $1 \leqslant k_1 < k_2 \leqslant Q$ such that there are at least $p = \left\lceil \frac{n}{\binom{Q}{2}} \right\rceil$ points in s_n for which the couple $(i_1(v_i), i_2(v_i))$ is (k_1, k_2). For the sake of simplicity, the points in s_n are reordered so that the p first of them exhibit this property. The corresponding subset of s_n is denoted s_p. This means that for all vector $\mathbf{y} = (y_i)$ in $\{-1, 1\}^n$, there is a function $g_{\beta(\mathbf{y})}$ in $\mathcal{G}_\beta$ characterized by the vector $\beta(\mathbf{y}) = (\beta_k(\mathbf{y}))_{1 \leqslant k \leqslant Q} \in V_{N,Q}$ such that:

$$\forall i \in [\![1, p]\!], \begin{cases} \text{if } y_i = \ \ 1, & \frac{1}{2}\left(\beta_{k_1}(\mathbf{y})^T v_i - \max_{k \neq k_1} \beta_k(\mathbf{y})^T v_i\right) - c_i \geqslant \gamma \\ \text{if } y_i = -1, & \frac{1}{2}\left(\beta_{k_2}(\mathbf{y})^T v_i - \max_{k \neq k_2} \beta_k(\mathbf{y})^T v_i\right) + c_i \geqslant \gamma \end{cases}$$

which implies that

$$\forall i \in [\![1, p]\!], \begin{cases} \text{if } y_i = \ \ 1, & \frac{1}{2}\left(\beta_{k_1}(\mathbf{y})^T v_i - \beta_{k_2}(\mathbf{y})^T v_i\right) - c_i \geqslant \gamma \\ \text{if } y_i = -1, & \frac{1}{2}\left(\beta_{k_2}(\mathbf{y})^T v_i - \beta_{k_1}(\mathbf{y})^T v_i\right) + c_i \geqslant \gamma \end{cases}. \tag{11}$$

Consider now any split of s_p into two subsets $s_{p,1}$ and $s_{p,2}$. Consider any vector $\mathbf{y}$ in $\{-1, 1\}^n$ such that $y_i = 1$ if $v_i \in s_{p,1}$ and $y_i = -1$ if $v_i \in s_{p,2}$. It results from (11) that:

$$\frac{1}{2}\left(\beta_{k_1}(\mathbf{y}) - \beta_{k_2}(\mathbf{y})\right)^T \left(\sum_{v_i \in s_{p,1}} v_i - \sum_{v_i \in s_{p,2}} v_i \right) - \sum_{v_i \in s_{p,1}} c_i + \sum_{v_i \in s_{p,2}} c_i \geqslant p\gamma. \tag{12}$$

Conversely, consider any vector $\mathbf{y}$ such that $y_i = -1$ if $v_i \in s_{p,1}$ and $y_i = 1$ if $v_i \in s_{p,2}$. We have:

$$\frac{1}{2}\left(\beta_{k_2}(\mathbf{y}) - \beta_{k_1}(\mathbf{y})\right)^T \left(\sum_{v_i \in s_{p,1}} v_i - \sum_{v_i \in s_{p,2}} v_i \right) + \sum_{v_i \in s_{p,1}} c_i - \sum_{v_i \in s_{p,2}} c_i \geqslant p\gamma. \tag{13}$$

Combining (12), (13), and the Cauchy-Schwarz inequality, it appears that (whatever the sign of $\sum_{v_i \in s_{p,1}} c_i - \sum_{v_i \in s_{p,2}} c_i$ is) there is a function g_β in $\mathcal{G}_\beta$ such that

$$\frac{1}{2} \|\beta_{k_1} - \beta_{k_2}\|_2 \left\| \sum_{v_i \in s_{p,1}} v_i - \sum_{v_i \in s_{p,2}} v_i \right\|_2 \geqslant p\gamma. \tag{14}$$

$$\forall \beta \in V_{N,Q},\ \forall (k,l) : 1 \leqslant k < l \leqslant Q,\ \ \|\beta_k - \beta_l\|_2^2 = \|\beta_k\|_2^2 + \|\beta_l\|_2^2 - 2\beta_k^T \beta_l \leqslant \|\beta_k\|_2^2 + \|\beta_l\|_2^2 .$$

Thus,

$$\forall \beta \in V_{N,Q},\ \ \max_{1 \leqslant k < l \leqslant Q} \|\beta_k - \beta_l\|_2^2 \leqslant \max_{1 \leqslant k < l \leqslant Q} \left\{ \|\beta_k\|_2^2 + \|\beta_l\|_2^2 \right\} \leqslant \|\beta\|_2^2 .$$

Since we know that $\max_{\beta \in V_{N,Q}} \|\beta\|_2 = \sqrt{Q}$ (see Remark 2), we get

$$\forall \beta \in V_{N,Q},\ \ \max_{1 \leqslant k < l \leqslant Q} \|\beta_k - \beta_l\|_2 \leqslant \sqrt{Q}.$$

A substitution of this upper bound in (14) then concludes the proof. ■

Lemma 3 *For all $n \in \mathbb{N}^*$, all subset $s_n = \{v_i : 1 \leqslant i \leqslant n\}$ of U_Q^N can be split into two subsets $s_{n,1}$ and $s_{n,2}$ satisfying*

$$\left\| \sum_{v_i \in s_{n,1}} v_i - \sum_{v_i \in s_{n,2}} v_i \right\|_2 \leqslant \sqrt{Nn}. \tag{15}$$

Proof Let $s_n = \{v_i : 1 \leqslant i \leqslant n\} \subset U_Q^N$. Let $\sigma = (\sigma_i)_{1 \leqslant i \leqslant n}$ be a Rademacher sequence, i.e., a sequence of i.i.d. random variables taking the values -1 and 1 with probability $\frac{1}{2}$. We have:

$$\mathbb{E}_\sigma \left\| \sum_{i=1}^n \sigma_i v_i \right\|_2^2 = \mathbb{E}_\sigma \left[\sum_{i=1}^n \sum_{j=1}^n \sigma_i \sigma_j v_i^T v_j \right] = \sum_{i=1}^n \sum_{j=1}^n v_i^T v_j \mathbb{E}_\sigma [\sigma_i \sigma_j] = \sum_{i=1}^n \|v_i\|_2^2 \leqslant n \max_{v \in U_Q^N} \|v\|_2^2 .$$

Obviously, the vectors of U_Q^N whose ℓ_2 norm is maximum are its vertices (extreme points). The corresponding value of the norm is $\sqrt{N}$. Thus,

$$\mathbb{E}_\sigma \left\| \sum_{i=1}^n \sigma_i v_i \right\|_2^2 \leqslant Nn.$$

This implies that there exists a binary vector $\mathbf{y} = (y_i)_{1 \leqslant i \leqslant n} \in \{-1, 1\}^n$ such that

$$\left\| \sum_{i=1}^n y_i v_i \right\|_2 \leqslant \sqrt{Nn}.$$

Setting $s_{n,1} = \{v_i \in s_n : y_i = 1\}$ and $s_{n,2} = s_n \setminus s_{n,1}$ then concludes the proof. ■

With Lemmas 2 and 3 at hand, the proof of Theorem 1 is elementary.

Proof Let $s_q = \{v_i : 1 \leqslant i \leqslant q\}$ be a subset of U_Q^N N-shattered with margin γ by $\Delta \mathcal{G}_\beta$. According to Lemma 2, there is at least a subset s_n of s_q of cardinality n equal to $\left\lceil \frac{q}{\binom{Q}{2}} \right\rceil$ satisfying

$$\left\| \sum_{v_i \in s_{n,1}} v_i - \sum_{v_i \in s_{n,2}} v_i \right\|_2 \geqslant \frac{2n}{\sqrt{Q}} \gamma$$

for all its splits into two subsets $s_{n,1}$ and $s_{n,2}$. Since, according to Lemma 3, there is at least one of these splits for which (15) holds true,

$$\frac{2n}{\sqrt{Q}}\gamma \leqslant \sqrt{Nn}$$

which implies that

$$q \leqslant \binom{Q}{2}\frac{NQ}{4\gamma^2},$$

which is precisely (8). ■

References

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

[2] R. Bonidal, F. Thomarat, and Y. Guermeur. Estimating the class posterior probabilities in biological sequence segmentation. In *SMTDA'12*, 2012.

[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] L. Breiman and J.H. Friedman. Predicting multivariate responses in multiple linear regression. *Journal of the Royal Statistical Society. Series B*, 59(1):3–54, 1997.

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

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

[7] Y. Guermeur. Guaranteed risk for large margin multi-category classifiers, application to M-SVMs. (submitted).

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

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

[10] 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.

[11] Y. Guermeur. A generic model of multi-class support vector machine. *International Journal of Intelligent Information and Database Systems*, 6(6):555–577, 2012.

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

[13] K. Hornik, M. Stinchcombe, and H. White. Multilayer feedforward networks are universal approximators. *Neural Networks*, 2(5):359–366, 1989.

[14] D.W. Hosmer and S. Lemeshow. *Applied Logistic Regression.* Wiley, London, 1989.

[15] 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.

[16] A. Krogh and S.K. Riis. Hidden neural networks. *Neural Computation*, 11(2):541–563, 1999.

[17] 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.

[18] H.-T. Lin, C.-J. Lin, and R.C. Weng. A note on Platt's probabilistic outputs for support vector machines. *Machine Learning*, 68(3):267–276, 2007.

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

[20] B.K. Natarajan. On learning sets and functions. *Machine Learning*, 4(1):67–97, 1989.

[21] J.C. Platt. Probabilities for SV machines. In A.J. Smola, P.L. Bartlett, B. Schölkopf, and D. Schuurmans, editors, *Advances in Large Margin Classifiers*, chapter 5, pages 61–73. The MIT Press, Cambridge, MA, 2000.

[22] M.D. Richard and R.P. Lippmann. Neural network classifiers estimate Bayesian *a posteriori* probabilities. *Neural Computation*, 3(4):461–483, 1991.

[23] J.B. Rosen. The gradient projection method for nonlinear programming. Part I. Linear constraints. *Journal of the Society for Industrial and Applied Mathematics*, 8(1):181–217, 1960.

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

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

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

[27] F. Thomarat, F. Lauer, and Y. Guermeur. Cascading discriminant and generative models for protein secondary structure prediction. In *PRIB'12*, pages 166–177, 2012.

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

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

[30] 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.

[31] 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.

[32] Z. Zhang and M.I. Jordan. Bayesian multicategory support vector machines. In *UAI'06*, pages 552–559, 2006.

[33] 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.