Model Selection for Multi-class SVMs

Yann Guermeur[1], Myriam Maumy[2], and Frédéric Sur[1]

[1] LORIA-CNRS
Campus Scientifique, BP 239,
54506 Vandœuvre-lès-Nancy Cedex, France
(e-mail: Yann.Guermeur@loria.fr, Frederic.Sur@loria.fr)
[2] IRMA-ULP
7 rue René Descartes
67084 Strasbourg Cedex, France
(e-mail: mmaumy@math.u-strasbg.fr)

Abstract. In the framework of statistical learning, fitting a model to a given problem is usually done in two steps. First, model selection is performed, to set the values of the hyperparameters. Second, training results in the selection, for this set of values, of a function performing satisfactorily on the problem. Choosing the values of the hyperparameters remains a difficult task, which has only been addressed so far in the case of bi-class SVMs. We derive here a solution dedicated to M-SVMs. It is based on a new bound on the risk of large margin classifiers.
Keywords: Multi-class SVMs, hyperparameters, soft margin parameter.

1 Introduction

When support vector machines (SVMs) [Vapnik, 1998] were introduced in the early nineties, they were seen by some as off-the-shelf tools. This idealistic picture soon proved too optimistic. Not only does their training raise technical difficulties, but the tuning of the kernel parameters and the soft margin parameter C also remains a difficult task. In literature, this question is addressed for (two-class) pattern recognition and function estimation SVMs. The methods proposed often rest on estimates of the true risk of the machine [Chapelle *et al.*, 2002]. The case of multi-class discriminant analysis was only considered in the framework of decomposition schemes [Passerini *et al.*, 2004]. The case of multi-class SVMs (M-SVMs) calls for specific solutions. Indeed, the implementation of the structural risk minimization (SRM) inductive principle [Vapnik, 1982] utterly rests on the availability of tight error bounds and the standard uniform convergence results do not carry over nicely to the case of multi-category large margin classifiers. In this paper, we derive a new bound on the generalization performance of M-SVMs in terms of constraints on the hyperplanes. This bound, interesting in its own right, makes central use of a result relating covering problems and the degree of compactness of operators. It serves as an objective function to tune the value of the soft margin parameter. This way, the value of C and the dual variables α can be determined simultaneously, at a cost of the same

order of magnitude as the one of a standard training. The organization of the paper is as follows. Section 2 is devoted to the description of the bound on which the study is based. In Section 3, the measure of capacity involved is bounded in terms of the entropy numbers of a linear operator. The resulting objective function is used in Section 4, to derive the algorithm tuning C and the parameters α. A first assessment of this algorithm on a toy problem is described in Section 5. Due to lack of space, proofs are omitted.

2 Bound on the risk of large margin classifiers

We consider the case of a Q-category pattern recognition problem, with $Q \geq 3$ to exclude the degenerate case of dichotomies. Let $\mathcal{X}$ be the space of description and $\mathcal{C} = \{C_1, \ldots, C_k, \ldots, C_Q\}$ the set of categories. We make the assumption that there is a joint probability measure μ, fixed but unknown, on $(\mathcal{X} \times \mathcal{C}, \mathcal{B})$, where $\mathcal{B}$ is a σ-algebra on $\mathcal{X} \times \mathcal{C}$. This measure utterly characterizes the problem of interest. Our goal is to find, in a given set $\mathcal{H}$ of functions from $\mathcal{X}$ into $\mathbb{R}^Q$, a function with the lowest "error rate" on this problem. The "error rate" of a function h in $\mathcal{H}$ with component functions h_k, $(1 \leq k \leq Q)$, is the *expected risk* of the corresponding discrimination function, obtained by assigning each pattern x to the category C_k in $\mathcal{C}$ satisfying: $h_k(x) = \max_l h_l(x)$. The patterns for which this assignation is ambiguous are assigned to a dummy category, so that they contribute to the computation of the different risks considered below. Hereafter, $C(x)$ will denote indifferently the category of the (labelled) pattern x, or the index of this category. To simplify notations, when no confusion is possible, the labels of the categories will be identified with their indices, i.e. k could be used in place of C_k. First of all, we define the functional that is to be minimized, the expected risk.

Definition 1 (Expected risk). The *expected risk* of a function f from $\mathcal{X}$ into $\mathcal{C}$ is the probability that $f(x) \neq C(x)$ for a labelled example $(x, C(x))$ chosen randomly according to μ, i.e.:

$$R(f) = \mu\{(x,k) : f(x) \neq k\} = \int_{\mathcal{X}\times\mathcal{C}} \mathbb{1}_{\{f(x)\neq k\}}(x,k) d\mu(x,k) \qquad (1)$$

where $\mathbb{1}_{\{f(x)\neq k\}}$ is the indicator function of the set $\{(x,k) \in \mathcal{X} \times \mathcal{C} : f(x) \neq k\}$.

In the framework of large margin multi-category pattern recognition, the class of functions of interest is not $\mathcal{H}$ itself, but rather its image by an adequately chosen operator. Basically, this is due to the fact that the two central elements to assign a pattern to a category and to derive a level of confidence in this assignation are respectively the index of the highest output and the difference between this output and the second highest one. The operator used here was introduced in previous works.

Definition 2 (Δ operator). Define Δ as an operator on $\mathcal{H}$ such that:

$$\begin{aligned}&\Delta : \mathcal{H} \longrightarrow \Delta\mathcal{H}\\ &\quad h = (h_k)_{1\leq k\leq Q} \mapsto \Delta h = \left(\Delta h_k : x \mapsto \tfrac{1}{2}\left\{h_k(x) - \max_{l\neq k} h_l(x)\right\}\right)_{1\leq k\leq Q}.\end{aligned}$$

Let s_m be a m-sample of examples independently drawn from μ. The empirical margin risk is defined as follows:

Definition 3 (Empirical margin risk). The empirical risk with margin $\gamma > 0$ of h on a set s_m is

$$R_{\gamma,s_m}(h) = \frac{1}{m} \cdot \#\left\{(x_i, C(x_i)) \in s_m :\ \Delta h_{C(x_i)}(x_i) < \gamma\right\}, \tag{2}$$

where # returns the cardinality of the set to which it is applied.

For technical reasons, it is useful to bound the values taken by the functions Δh_k in $[-\gamma, \gamma]$, the smallest interval such that this change has no incidence on the empirical margin risk. This is achieved by application of the π_γ operator.

Definition 4 (π_γ operator [Bartlett, 1998]). Let $\mathcal{G}$ be a set of functions from $\mathcal{X}$ into $\mathbb{R}^Q$. For $\gamma > 0$, let $\pi_\gamma : g = (g_k)_{1\leq k\leq Q} \mapsto \pi_\gamma(g) = (\pi_\gamma(g_k))_{1\leq k\leq Q}$ be the piecewise-linear squashing operator defined as:

$$\forall x \in \mathcal{X},\ \pi_\gamma(g_k)(x) = \begin{cases}\gamma.sign\left(g_k(x)\right) & \text{if } |g_k(x)| \geq \gamma\\ g_k(x) & \text{otherwise}\end{cases}. \tag{3}$$

Let Δ_γ denote $\pi_\gamma \circ \Delta$ and $\Delta_\gamma\mathcal{H}$ be defined as the set of functions $\Delta_\gamma h$. Our guaranteed risk is made up of two terms, the empirical margin risk given above and a "confidence interval" involving a covering number of $\Delta_\gamma\mathcal{H}$.

Definition 5 (ϵ-cover, ϵ-net and covering numbers). Let (E, ρ) be a pseudo-metric space and E' be a subset of E. An *ϵ-cover* of E' is a coverage of E' with balls of radius ϵ the centers of which belong to E. These centers form an *ϵ-net* of E'[1]. If E' has an ϵ-cover of finite cardinality, then its *covering number* $\mathcal{N}(\epsilon, E', \rho)$ is the smallest cardinality of its ϵ-covers. If there is no such finite cover, then the covering number is defined to be ∞.

The covering number of interest uses the following pseudo-metric:

Definition 6. Let $\mathcal{G}$ be a set of functions from $\mathcal{X}$ into $\mathbb{R}^Q$. For a set s of points in $\mathcal{X}$ of finite cardinality, define the pseudo-metric d_s on $\mathcal{G}$ as:

$$\forall (g, g') \in \mathcal{G}^2,\ d_s(g, g') = \max_{x\in s} \|g(x) - g'(x)\|_\infty. \tag{4}$$

[1] Hereafter, we will only consider a restricted case in which the ϵ-nets of E' will be supposed to be subsets of E' itself.

Let $\mathcal{N}_{\infty,\infty}(\epsilon, \Delta_\gamma \mathcal{H}, m) = \sup_{s_m \in \mathcal{X}^m} \mathcal{N}(\epsilon, \Delta_\gamma \mathcal{H}, d_{s_m})$. These definitions being given, we can formulate the following theorem, which extends to the multi-class case Corollary 9 in [Bartlett, 1998].

Theorem 1 (Theorem 1 in [Guermeur, 2004]). *Let s_m be a m-sample of examples independently drawn from μ. With probability at least $1-\delta$, for every value of γ in $(0,1]$, the risk of any function h in the class $\mathcal{H}$ of functions computed by a Q-class large margin classifier is bounded from above by:*

$$R(h) \leq R_{\gamma,s_m}(h) + \sqrt{\frac{2}{m}\left(\ln(2\mathcal{N}_{\infty,\infty}(\gamma/4, \Delta_\gamma \mathcal{H}, 2m)) + \ln\left(\frac{2}{\gamma\delta}\right)\right)} + \frac{1}{m}. \quad (5)$$

The practical interest of such a bound utterly rests on the possibility to derive a tight bound on the covering number appearing in the "confidence interval". To that end, a preliminary simplification is useful.

Proposition 1. $\forall (\gamma, \epsilon) : 0 < \epsilon \leq \gamma \leq 1,\ \ \mathcal{N}_{\infty,\infty}(\epsilon, \Delta_\gamma \mathcal{H}, m) \leq \mathcal{N}_{\infty,\infty}(\epsilon, \mathcal{H}, m)$.

Theorem 1 and Proposition 1 imply that deriving a guaranteed risk for $\mathcal{H}$ can boil down to deriving a bound on $\mathcal{N}_{\infty,\infty}(\epsilon, \mathcal{H}, m)$. In [Guermeur, 2004], to bound the covering number appearing in (5), we investigated a standard pathway, consisting in relating this capacity measure to a generalized VC dimension [Vapnik, 1998] through an extension of Sauer's lemma [Sauer, 1972]. It appeared then that in the multivariate case, establishing the connection between the separation of functions (with respect to the selected pseudo-metric) and their shattering capacity is no longer trivial. Taking our inspiration from [Carl and Stephani, 1990,Williamson *et al.*, 2000], we assess here a more direct approach: relating the covering numbers of $\mathcal{H}$ to the entropy numbers of a linear operator.

3 Bound on the covering numbers of M-SVMs

SVMs [Cortes and Vapnik, 1995] are learning systems introduced by Vapnik and co-workers as a nonlinear extension of the maximal margin hyperplane [Vapnik, 1982]. Originally, they were designed to compute dichotomies. In this context, the principle on which they are based can be outlined very simply. First, the examples are mapped into a high-dimensional Hilbert space thanks to a nonlinear transform. Second, the maximal margin hyperplane is computed in that space, to separate the two categories. Initially, the extension to perform multi-class discriminant analysis utterly rested on decomposition schemes. The M-SVMs are globally more recent (see [Guermeur, 2004] for references). The family $\mathcal{H}$ of functions $h = (h_k)_{1 \leq k \leq Q}$ computed by these machines can be defined by:

$$\forall k \in \{1, \ldots, Q\},\ h_k(x) = \langle w_k, \Phi(x) \rangle + b_k, \quad (6)$$

where Φ is some mapping from $\mathcal{X}$ into a Reproducing Kernel Hilbert Space (RKHS) [Aronszajn, 1950] $\left(E_{\Phi(\mathcal{X})}, \langle ., . \rangle\right)$, derived from a symmetric positive kernel κ. The vectors w_k belong to $E_{\Phi(\mathcal{X})}$, whereas the b_k are real numbers. As in the case of all kernel machines, Φ does not appear explicitly in the computations. Thanks to the "kernel trick", which rests on the equation:

$$\forall (x, x') \in \mathcal{X}^2,\ \kappa(x, x') = \langle \Phi(x), \Phi(x') \rangle, \tag{7}$$

all what is needed to perform training or testing are the values taken by the kernel κ. To ensure the finiteness of the capacity measures, we make the additional assumption that $\Phi(\mathcal{X})$ is included in the closed ball of radius $\Lambda_{\Phi(\mathcal{X})}$ in $E_{\Phi(\mathcal{X})}$, that is: $\forall x \in \mathcal{X},\ \ \|\Phi(x)\| = \sqrt{\kappa(x, x)} \leq \Lambda_{\Phi(\mathcal{X})}$. To upperbound $\mathcal{N}_{\infty,\infty}(\epsilon, \mathcal{H}, m)$ when $\mathcal{H}$ is a M-SVM, we use a result regarding linear operators on Banach spaces. This implies that the covering numbers of $\mathcal{H}$ could be bounded in terms of the covering numbers of its linear counterpart.

Proposition 2. *Let $\mathcal{H}$ be the class of functions implemented by a Q-category SVM under the constraint that $b = (b_k) \in [-\beta, \beta]^Q$. Let $\tilde{\mathcal{H}}$ be the subset of $\mathcal{H}$ made up of the functions for which $b = 0$. Then, for all $\epsilon > 0$,*

$$\mathcal{N}_{\infty,\infty}(\epsilon, \mathcal{H}, m) \leq \left(2 \left\lceil \frac{\beta}{\epsilon} \right\rceil + 1\right)^Q \mathcal{N}_{\infty,\infty}(\epsilon/2, \tilde{\mathcal{H}}, m). \tag{8}$$

A function $\tilde{h}$ in $\tilde{\mathcal{H}}$ is characterized by the vector $\mathbf{w} = (w_k)_{1 \leq k \leq Q}$ in $E^Q_{\Phi(\mathcal{X})}$. This space is endowed with a Hilbertian structure. Its dot product is given by: $\forall\, (\mathbf{w}, \mathbf{w}') \in \left(E^Q_{\Phi(\mathcal{X})}\right)^2,\ \langle \mathbf{w}, \mathbf{w}' \rangle = \sum_{k=1}^{Q} \langle w_k, w'_k \rangle$. Its norm is the one derived from $\langle ., . \rangle$. Since the additional hypothesis $\|\mathbf{w}\| \leq 1$ will also be used, we introduce another proposition.

Proposition 3. *Let $\tilde{\mathcal{H}}$ be defined as above, under the additional constraint that $\|\mathbf{w}\| \leq \Lambda_w$. Let $\mathcal{U}$ be its restriction to the functions satisfying $\|\mathbf{w}\| \leq 1$.*

$$\forall \epsilon > 0,\ \ \mathcal{N}_{\infty,\infty}(\Lambda_w \epsilon, \tilde{\mathcal{H}}, m) \leq \mathcal{N}_{\infty,\infty}(\epsilon, \mathcal{U}, m). \tag{9}$$

Definition 7 (entropy numbers). Let (E, ρ) be a pseudo-metric space. Let E' be a subset of E. The nth *entropy number* of E', $\epsilon_n(E')$, is defined as the smallest real ϵ such that there exists an ϵ-cover of E' of cardinality at most n. Let E and F be two Banach spaces. $\mathfrak{L}(E, F)$ denotes the Banach space of all (bounded linear) operators from E into F equipped with the usual norm. Let U_E be the closed unit ball of E. The nth entropy number of $S \in \mathfrak{L}(E, F)$ is defined as

$$\epsilon_n(S) = \epsilon_n\left(S(U_E)\right). \tag{10}$$

By ℓ_p^n we denote the vector space of n-tuples equipped with the norm $\|.\|_p$.

Definition 8 (Evaluation operator). Let s_m be any element of $\mathcal{X}^m$. We define S_{s_m} as the linear operator given by:

$$\begin{array}{rccc} S_{s_m} : & E_{\Phi(\mathcal{X})}^Q & \longrightarrow & \ell_\infty^{Qm} \\ & \mathbf{w} & \mapsto & S_{s_m}(\mathbf{w}) = (\langle w_k, \Phi(x_i)\rangle)_{1\leq k\leq Q,\ 1\leq i\leq m} \end{array}$$

The connection between $\mathcal{N}_{\infty,\infty}(\epsilon,\mathcal{U},m)$ and the entropy numbers of S_{s_m} is given by the following proposition.

Proposition 4. *If for all $s_m \in \mathcal{X}^m$, $\epsilon_n(S_{s_m}) \leq \epsilon$, then $\mathcal{N}_{\infty,\infty}(\epsilon,\mathcal{U},m) \leq n$.*

To bound $\epsilon_n(S_{s_m})$, we use a result due to Maurey and Carl.

Lemma 1 (Lemma 6.4.1 in [Carl and Stephani, 1990]). *Let H be a Hilbert space, m a positive integer and $S \in \mathfrak{L}(H, \ell_\infty^m)$. Then, for $1 \leq n \leq m$,*

$$\epsilon_{2^{n-1}}(S) \leq c\|S\| \left(\frac{1}{n}\log\left(1+\frac{m}{n}\right)\right)^{1/2}, \tag{11}$$

where c is a universal constant and by log *we denote the logarithm to base* 2.

Lemma 1 still holds without the hypothesis $n \leq m$. Gathering the results from Propositions 1 to 4 together with this lemma (applied on S_{s_m}) produces a handy bound on the covering number of interest.

Theorem 2. *Let $\mathcal{H}$ be the class of functions computed by a Q-category M-SVM under the hypothesis that $\Phi(\mathcal{X})$ is included in the closed ball of radius $\Lambda_{\Phi(\mathcal{X})}$ in $E_{\Phi(\mathcal{X})}$ and the constraints that $\|\mathbf{w}\| \leq \Lambda_w$ and $b \in [-\beta,\beta]^Q$. For every value of γ in $(0,1]$,*

$$\mathcal{N}_{\infty,\infty}(\gamma/4, \Delta_\gamma\mathcal{H}, 2m) \leq \left(2\left\lceil\frac{4\beta}{\gamma}\right\rceil+1\right)^Q \cdot 2^{\frac{8c\Lambda_w\Lambda_{\Phi(\mathcal{X})}}{\gamma}\sqrt{\frac{2Qm}{\ln(2)}}-1}. \tag{12}$$

4 Tuning the soft margin parameter

To tune C thanks to the guaranteed risk derived above, we propose a simple line search. Although it is compatible with any of the training algorithms published, for the sake of simplicity, we focus here on the case of the most common machine, introduced in [Weston and Watkins, 1998]. Training it amounts to solving the following quadratic programming (QP) problem:

Problem 1 (Primal).

$$\min_{(\mathbf{w},b)} \left\{ \frac{1}{2}\sum_{k=1}^{Q}\|w_k\|^2 + C\sum_{i=1}^{m}\sum_{k=1}^{Q}\xi_{ik} \right\}$$

$$\text{s.t.} \begin{cases} h_{C(x_i)}(x_i) - h_k(x_i) \geq 1-\xi_{ik}, & (1\leq i\leq m),\ (1\leq k\neq C(x_i)\leq Q) \\ \xi_{ik} \geq 0, & (1\leq i\leq m),\ (1\leq k\neq C(x_i)\leq Q) \end{cases}.$$

In the objective function, the sum of slack variables is used in place of the empirical margin risk, whereas the penalty term $\frac{1}{2}\sum_{k=1}^{Q}\|w_k\|^2$ is added to perform capacity control. To the best of our knowledge, Theorem 2 offers the first justification for this choice. By setting the soft margin parameter C, one specifies a compromise between training accuracy and complexity. If the objective function itself cannot be used in that purpose, since it is only distantly related to a guaranteed risk, performing n-fold cross-validation is a sensible possibility. However, it implies training the machine n times for each value of C considered, which can be prohibitive in terms of `cpu` time requirements. Furthermore, this no longer corresponds to the implementation of the SRM principle. In that respect, our solution should prove more satisfactory. To detail it, we first introduce the formulation in which Problem 1 is solved, its Wolfe dual. Let α_{ik} be the Lagrange multiplier associated with the constraint $\langle w_{C(x_i)} - w_k, \Phi(x_i)\rangle + b_{C(x_i)} - b_k - 1 + \xi_{ik} \geq 0$. Let

$$J(\alpha) = \frac{1}{2}\left\{\sum_{i\simeq j}\sum_{k=1}^{Q}\sum_{l=1}^{Q}\alpha_{ik}\alpha_{jl}\kappa(x_i,x_j) - 2\sum_{i=1}^{m}\sum_{j=1}^{m}\sum_{k=1}^{Q}\alpha_{ik}\alpha_{jC(x_i)}\kappa(x_i,x_j) + \sum_{i=1}^{m}\sum_{j=1}^{m}\sum_{k=1}^{Q}\alpha_{ik}\alpha_{jk}\kappa(x_i,x_j)\right\} - \sum_{i=1}^{m}\sum_{k=1}^{Q}\alpha_{ik},$$

with $i \simeq j$ meaning that x_i and x_j belong to the same category.

Problem 2 (Dual).

$$\min_{\alpha} J(\alpha)$$

$$\text{s.t.} \begin{cases} \sum_{x_i\in C_k}\sum_{l=1}^{Q}\alpha_{il} - \sum_{i=1}^{m}\alpha_{ik} = 0 & (1 \leq k \leq Q-1) \\ 0 \leq \alpha_{ik} \leq C & (1 \leq i \leq m),\ (1 \leq k \neq C(x_i) \leq Q) \end{cases}.$$

Based on this dual formulation, our algorithm can be expressed as follows:

```
/* Initialization */
  C_0 := C^(0), α^(0) := 0_Qm;
/* Main loop */
  For i := 1 to nb_iter do
        train_SVM(C_{i-1}, s_m, α^(i-1)) ⟶ α^(i);
        C_i := C_{i-1} + ε;
  done
/* Termination */
  i_0 := Argmin_{1≤i≤nb_iter} { compute_bound(C_{i-1}, s_m, α^(i)) };
  C := C_{i_0};
```

In words, this algorithm consists in training the M-SVM a given number of times (calls of the function `train_SVM`) for increasing values of C, and

checking each time the value of the guaranteed risk (calls of the function `compute_bound`). Eventually, the value retained is the one corresponding to the "argmin", C_{i_0}. The benefit in terms of `cpu` time springs from the fact that the initial feasible solution used for the i+1-th training is the optimal solution of the i-th training, $\alpha^{(i)}$. Note that this is possible since we are working with increasing values of C. As a consequence, each training procedure converges more quickly than if the starting feasible solution was simply the null vector. Obviously, this exploration of the regularization path could also benefit from the implementation of a multi-class extension of the algorithm proposed in [Hastie *et al.*, 2004].

5 Experimental results

The bound provided by the conjunction of Theorem 1 and Theorem 2 can be applied to any M-SVM, whatever the kernel is. This is not a trivial property indeed, since it means that the feature space can be infinite dimensional, as in the case of a Gaussian kernel. In this section, for the sake of simplicity, we restrict to the case of a linear machine, i.e. a machine where the kernel is the Euclidean dot product. In that case, we can make use of a simpler result than Lemma 1 to bound from above the covering numbers of interest.

Proposition 5 (Proposition 1.3.1 in [Carl and Stephani, 1990]). *Let E and F be Banach spaces and $S \in \mathfrak{L}(E,F)$. If S is of rank r, then for $n \geq 1$,*

$$\epsilon_n(S) \leq 4\|S\|n^{-1/r}. \tag{13}$$

The bound resulting from this proposition is the following.

Theorem 3. *Let $\mathcal{H}$ be the class of functions computed by a Q-category M-SVM under the hypothesis that $\Phi(\mathcal{X})$ is included in the closed ball of radius $\Lambda_{\Phi(\mathcal{X})}$ in $E_{\Phi(\mathcal{X})}$ and the constraints that $\|\mathbf{w}\| \leq \Lambda_w$ and $b \in [-\beta,\beta]^Q$. Suppose further that the dimensionality of $E_{\Phi(\mathcal{X})}$ is finite and equal to d. For every value of γ in $(0,1]$,*

$$\mathcal{N}_{\infty,\infty}(\gamma/4, \Delta_\gamma\mathcal{H}, 2m) \leq \left(2\left\lceil\frac{4\beta}{\gamma}\right\rceil + 1\right)^Q \cdot \left(\frac{32\Lambda_w\Lambda_{\Phi(\mathcal{X})}}{\gamma}\right)^{Qd}. \tag{14}$$

The derivation of this bound rests on the fact that under the hypothesis $\dim\left(E_{\Phi(\mathcal{X})}\right) = d$, the rank of S_{s_m} (or $S_{s_{2m}}$) is bounded from above by the dimensionality of its domain, Qd. Otherwise, the sole bound on the rank available would be Qm (resp. $2Qm$), which would not meet our purpose (the guaranteed risk would not tend to the margin risk as m tends to infinity).

The algorithm of Section 4 is evaluated on a toy problem: the discrimination between three categories corresponding to isotropic Gaussian distributions in the plane with respective means and variances $\left((2.5\cdot\sqrt{3},-2.5),1\right)$, $((0,5),4)$

and $\big((-2.5 \cdot \sqrt{3}, -2.5), 16\big)$. The priors on the categories are equal. The training set is made up of 3000 points, 1000 for each category. This problem is illustrated on Figure 1. The optimal separating surfaces, implementing

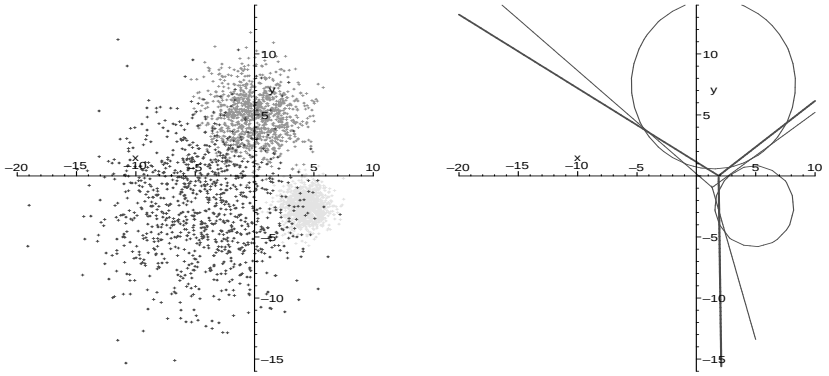

Fig. 1. Separating 3 Gaussian-distributed categories in $\mathbb{R}^2$. **Left:** training set. **Right:** Bayes' classifier (circles), optimal linear classifier and boundaries computed by the linear M-SVM for the (estimated) optimal value of C (thick lines).

Bayes' classifier, are two circles. The smaller one, at the bottom right of the right subfigure, corresponds to the boundary of the first category, the other one corresponding to the boundary of the second category. For this classifier, a Monte-Carlo method provides us with an estimate of the expected risk equal to 5.27%. With the same method, the estimates of the risks of the optimal linear separator and the M-SVM specified by the algorithm of Section 4 are respectively 5.85% and 6.30%. Thus, the estimation error is slightly inferior to the approximation error. Obviously, the significance of these initial results is limited, since they were obtained with a linear model, for which overfitting seldom happens. Additional experiments are currently being performed with a polynomial kernel in place of the Euclidean dot product.

6 Conclusions and future work

In this paper, a bound on the covering numbers of M-SVMs in terms of constraints on the parameters of their hyperplanes has been established. When plugged into the guaranteed risk derived in [Guermeur, 2004], it provides us with an objective function which can be used to implement the SRM inductive principle, and especially to tune the hyperparameters. An experimental validation on real-world data is underway, in protein secondary structure prediction, with the aim to improve the accuracy of the classifier introduced in [Guermeur *et al.*, 2004].

Acknowledgements

This study was inspired by an old discussion with A. Smola and a past collaboration with A. Elisseeff. The work of YG and FS is supported by the ACI "Masses de Données".

References

[Aronszajn, 1950]N. Aronszajn. Theory of reproducing kernels. *Trans. Amer. Math. Soc.*, 68:337–404, 1950.

[Bartlett, 1998]P.L. Bartlett. The sample complexity of pattern classification with neural networks: the size of the weights is more important than the size of the network. *IEEE Transactions on Information Theory*, 44(2):525–536, 1998.

[Carl and Stephani, 1990]B. Carl and I. Stephani. *Entropy, compactness, and the approximation of operators.* Cambridge University Press, Cambridge, UK, 1990.

[Chapelle *et al.*, 2002]O. Chapelle, V.N. Vapnik, O. Bousquet, and S. Mukherjee. Choosing Multiple Parameters for Support Vector Machines. *Machine Learning*, 46(1):131–159, 2002.

[Cortes and Vapnik, 1995]C. Cortes and V.N. Vapnik. Support-Vector Networks. *Machine Learning*, 20:273–297, 1995.

[Guermeur *et al.*, 2004]Y. Guermeur, A. Lifchitz, and R. Vert. A kernel for protein secondary structure prediction. In B. Schölkopf, K. Tsuda, and J.-P. Vert, editors, *Kernel Methods in Computational Biology*, pages 193–206. The MIT Press, 2004.

[Guermeur, 2004]Y. Guermeur. Large margin multi-category discriminant models and scale-sensitive Ψ-dimensions. Technical Report RR-5314, INRIA, 2004.

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

[Passerini *et al.*, 2004]A. Passerini, M. Pontil, and P. Frasconi. New results on error correcting output codes of kernel machines. *IEEE Transactions on Neural Networks*, 15(1):45–54, 2004.

[Sauer, 1972]N. Sauer. On the density of families of sets. *Journal of Combinatorial Theory (A)*, 13:145–147, 1972.

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

[Vapnik, 1998]V.N. Vapnik. *Statistical learning theory.* John Wiley & Sons, Inc., N.Y., 1998.

[Weston and Watkins, 1998]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.

[Williamson *et al.*, 2000]R.C. Williamson, A.J. Smola, and B. Schölkopf. Entropy numbers of linear function classes. In *COLT'00*, pages 309–319, 2000.