Sample Complexity of Classifiers Taking Values in $\mathbb{R}^Q$, Application to Multi-Class SVMs

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

Running head: VC Bound and Rademacher Complexity for M-SVMs

Abstract

Bounds on the risk play a crucial role in statistical learning theory. They usually involve as capacity measure of the model studied the VC dimension or one of its extensions. In classification, such "VC dimensions" exist for models taking values in $\{0,1\}$, $[\![1,Q]\!]$, and $\mathbb{R}$. We introduce the generalizations appropriate for the missing case, the one of models with values in $\mathbb{R}^Q$. This provides us with a new guaranteed risk for M-SVMs. For those models, a sharper bound is obtained by using the Rademacher complexity.

1 Introduction

Vapnik's statistical learning theory (Vapnik, 1998) deals with three types of problems: pattern recognition, regression estimation and density estimation. However, the theory of bounds has primarily been developed for the computation of dichotomies only. Central in this theory is the notion of "capacity" of classes of functions. In the case of binary classifiers, the standard measure of this capacity is the Vapnik-Chervonenkis (VC) dimension. Extensions have also been proposed for real-valued bi-class models and multi-class models taking their values in the set of categories. Strangely enough, no generalized VC dimension was available so far for Q-category classifiers taking their values in $\mathbb{R}^Q$. This was all the more unsatisfactory as many classifiers are of this kind, such as the multi-layer perceptrons, or the multi-class support vector machines (M-SVMs). In this paper, scale-sensitive Ψ-dimensions named γ-Ψ-dimensions are introduced to bridge this gap. A generalization of Sauer's lemma (Sauer, 1972) is given, which relates the covering number appearing in the standard guaranteed risk for large margin multi-category discriminant models to one of these dimensions: the margin Natarajan dimension. This latter dimension is then upper bounded for the class of functions that the M-SVMs can implement. This provides us with a new bound on the sample complexity of these machines. A sharper bound is then derived using the Rademacher complexity.

The organization of the paper is as follows. Section 2 introduces the basic bound on the risk of large margin multi-category discriminant models. In Section 3, the γ-Ψ-dimensions are defined, and the corresponding generalization of Sauer's lemma is formulated. The upper bound on the margin Natarajan dimension of the M-SVMs is then described in Section 4. Finally, the bound based on a Rademacher average is established in Section 5.

2 Basic theory of large margin Q-category classifiers

We consider Q-category pattern recognition problems, with $3 \leq Q < \infty$. A pattern is represented by its description $x \in \mathcal{X}$ and the set of categories $\mathcal{Y}$ is identified with the set of indices of the

categories: $[\![1, Q]\!]$. The link between patterns and categories is supposed to be probabilistic. We assume that the product $\mathcal{X} \times \mathcal{Y}$ is a measurable space endowed with an unknown probability measure P. Let (X, Y) be a random pair with values in $\mathcal{X} \times \mathcal{Y}$ distributed according to P. We observe a m-sample $D_m = ((X_i, Y_i))_{1 \leq i \leq m}$ of independent copies of (X, Y). Training then consists in using D_m to select, in a given class of functions $\mathcal{G}$ on $\mathcal{X}$, a function classifying data in an optimal way. Given a function g in $\mathcal{G}$, the criterion characterizing the quality of the corresponding classification, the *risk* of g, is the expectation with respect to P of a *loss function*. We consider classes of functions from $\mathcal{X}$ into $\mathbb{R}^Q$. The function $g = (g_k)_{1 \leq k \leq Q} \in \mathcal{G}$ assigns $x \in \mathcal{X}$ to the category l if and only if $g_l(x) > \max_{k \neq l} g_k(x)$. Cases of ex æquo are treated as errors. This calls for the choice of a loss function ℓ defined on $\mathcal{Y} \times \mathbb{R}^Q$ by $\ell(y, v) = \mathbb{1}_{\{v_y \leq \max_{k \neq y} v_k\}}$. The risk of g is then given by:

$$R(g) = \mathbb{E}\left[\ell\left(Y, g\left(X\right)\right)\right] = \int_{\mathcal{X} \times \mathcal{Y}} \mathbb{1}_{\left\{g_y(x) \leq \max_{k \neq y} g_k(x)\right\}} dP(x, y).$$

This study deals with large margin classifiers, when the underlying notion of multi-class margin is the following one.

Definition 1 (Multi-class margin) *Let g be a function from $\mathcal{X}$ into $\mathbb{R}^Q$. Its* margin *on $(x, y) \in \mathcal{X} \times \mathcal{Y}$, $\mathcal{M}(g, x, y)$, is given by:*

$$\mathcal{M}(g, x, y) = \frac{1}{2}\left\{g_y(x) - \max_{k \neq y} g_k(x)\right\}.$$

Basically, the central elements to assign a pattern to a category and to derive a level of confidence in this assignment are the index of the highest output and the difference between this output and the second highest one. The class of functions of interest is thus the image of $\mathcal{G}$ by application of an operator extracting the appropriate pieces of information. Two such "margin operators" are considered here. The first one, Δ, preserves all the information provided by g up to an additive constant.

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

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

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

For the sake of simplicity, we have written Δg_k in place of $(\Delta g)_k$. In the sequel, this simplification will be performed implicitly with other operators. Obviously, $\mathcal{M}(g, x, y) = \Delta g_y(x)$, and thus

the example (x, y) is correctly classified by g if and only if $\Delta g_y(x) > 0$. The second operator, Δ^*, only preserves what is essential to characterize the classification. Since the aforementioned level of confidence (y being unknown) is provided by $\mathcal{M}(g, x, \cdot) = \max_k \Delta g_k(x)$, this operator is defined as follows.

Definition 3 (Δ^* operator) *Define Δ^* as an operator on $\mathcal{G}$ such that:*

$$\begin{aligned} \Delta^* : \quad & \mathcal{G} \longrightarrow \Delta^*\mathcal{G} \\ & g \mapsto \Delta^* g = (\Delta^* g_k)_{1 \leq k \leq Q} \\ \forall x \in \mathcal{X},\ \Delta^* g(x) &= (\operatorname{sign}(\Delta g_k(x)) \cdot \mathcal{M}(g, x, \cdot))_{1 \leq k \leq Q}\,. \end{aligned}$$

In the sequel, $\Delta^\#$ is used in place of Δ and Δ^* in the formulas that hold true for both operators. Obviously, the first of them is $R(g) = \mathbb{E}\left[\mathbb{1}_{\left\{\Delta^\# g_Y(X) \leq 0\right\}}\right]$. $\Delta^\#$ is also involved in the definition of the margin risk.

Definition 4 (Margin risk) *Let $\gamma \in \mathbb{R}_+^*$. The* risk with margin γ *of g, $R_\gamma(g)$, and its empirical estimate on D_m, $R_{\gamma,m}(g)$, are defined as:*

$$R_\gamma(g) = \int_{\mathcal{X}\times\mathcal{Y}} \mathbb{1}_{\left\{\Delta^\# g_y(x) < \gamma\right\}} dP(x, y), \quad R_{\gamma,m}(g) = \frac{1}{m}\sum_{i=1}^{m} \mathbb{1}_{\left\{\Delta^\# g_{Y_i}(X_i) < \gamma\right\}}.$$

For technical reasons, it is useful to squash the functions $\Delta^\# g_k$ as much as possible without altering the value of the margin risk. This is achieved by application of another operator.

Definition 5 (π_γ operator (Bartlett, 1998)) *For $\gamma \in \mathbb{R}_+^*$, define π_γ as an operator on $\mathcal{G}$ such that:*

$$\begin{aligned} \pi_\gamma : \quad & \mathcal{G} \longrightarrow \pi_\gamma\mathcal{G} \\ & g \mapsto \pi_\gamma g = (\pi_\gamma g_k)_{1 \leq k \leq Q} \\ \forall x \in \mathcal{X},\ \pi_\gamma g(x) &= (\operatorname{sign}(g_k(x)) \cdot \min(|g_k(x)|, \gamma))_{1 \leq k \leq Q}\,. \end{aligned}$$

Let $\Delta_\gamma^\#$ denote $\pi_\gamma \circ \Delta^\#$ and let $\Delta_\gamma^\#\mathcal{G}$ be defined as the set of functions $\Delta_\gamma^\# g$. The capacity of $\Delta_\gamma^\#\mathcal{G}$ is characterized by its covering numbers.

Definition 6 (ϵ-cover, ϵ-net and covering numbers) *Let (E, ρ) be a pseudo-metric space, $E' \subset E$ and $\epsilon \in \mathbb{R}_+^*$. An* ϵ-cover *of E' is a coverage of E' with open balls of radius ϵ the centers of which belong to E. These centers form an* ϵ-net *of E'. A* proper ϵ-net *of E' is an* ϵ-net *of E' included in E'. If E' has an ϵ-net of finite cardinality, then its* covering number $\mathcal{N}(\epsilon, E', \rho)$ *is*

the smallest cardinality of its ϵ-nets. If there is no such finite net, then the covering number is defined to be ∞. $\mathcal{N}^{(p)}(\epsilon, E', \rho)$ will designate the covering number of E' obtained by considering proper ϵ-nets only.

The covering numbers of interest use the following pseudo-metric:

Definition 7 (d_{x^n} pseudo-metric) *Let $n \in \mathbb{N}^*$. For a sequence $x^n = (x_i)_{1 \leq i \leq n} \in \mathcal{X}^n$, define the pseudo-metric d_{x^n} on $\mathcal{G}$ as:*

$$\forall (g, g') \in \mathcal{G}^2,\ d_{x^n}(g, g') = \max_{1 \leq i \leq n} \left\| g(x_i) - g'(x_i) \right\|_\infty .$$

Let $\mathcal{N}^{(p)}(\epsilon, \mathcal{G}, n) = \sup_{x^n \in \mathcal{X}^n} \mathcal{N}^{(p)}(\epsilon, \mathcal{G}, d_{x^n})$. The following theorem extends to the multi-class case Corollary 9 in Bartlett (1998).

Theorem 1 (Theorem 1 in Guermeur (2004)) *Let $\mathcal{G}$ be the class of functions that a large margin Q-category classifier on a domain $\mathcal{X}$ can implement. Let $\Gamma \in \mathbb{R}_+^*$ and $\delta \in (0,1)$. With probability at least $1-\delta$, for every value of γ in $(0, \Gamma]$, the risk of any function g in $\mathcal{G}$ is bounded from above by:*

$$R(g) \leq R_{\gamma,m}(g) + \sqrt{\frac{2}{m}\left(\ln\left(2\mathcal{N}^{(p)}\left(\gamma/4, \Delta_\gamma^\# \mathcal{G}, 2m\right)\right) + \ln\left(\frac{2\Gamma}{\gamma\delta}\right)\right)} + \frac{1}{m}.$$

Studying the sample complexity of a classifier $\mathcal{G}$ can thus amount to computing an upper bound on $\mathcal{N}^{(p)}(\gamma/4, \Delta_\gamma^\# \mathcal{G}, 2m)$. In Guermeur et al. (2005), we reached this goal by relating this number to the entropy numbers of the corresponding evaluation operator. In the present paper, we follow the traditional path of VC bounds, by making use of a generalized VC dimension.

3 Bounding covering numbers in terms of the margin Natarajan dimension

The Ψ-dimensions are the generalized VC dimensions that characterize the learnability of classes of $[\![1, Q]\!]$-valued functions.

Definition 8 (Ψ-dimensions (Ben-David et al., 1995)) *Let $\mathcal{F}$ be a class of functions on a set $\mathcal{X}$ taking their values in the finite set $[\![1, Q]\!]$. Let Ψ be a family of mappings ψ from $[\![1, Q]\!]$ into $\{-1, 1, *\}$, where $*$ is thought of as a null element. A subset $s_{\mathcal{X}^n} = \{x_i : 1 \leq i \leq n\}$ of $\mathcal{X}$*

is said to be Ψ-shattered *by* $\mathcal{F}$ *if there is a mapping* $\psi^n = \left(\psi^{(i)}\right)_{1\leq i\leq n}$ *in* Ψ^n *such that for each vector* v_y *in* $\{-1,1\}^n$*, there is a function* f_y *in* $\mathcal{F}$ *satisfying*

$$\left(\psi^{(i)} \circ f_y(x_i)\right)_{1\leq i\leq n} = v_y.$$

The Ψ-dimension *of* $\mathcal{F}$*, denoted by* Ψ*-dim*$(\mathcal{F})$*, is the maximal cardinality of a subset of* $\mathcal{X}$ Ψ*-shattered by* $\mathcal{F}$*, if this cardinality is finite. If no such maximum exists,* $\mathcal{F}$ *is said to have infinite* Ψ*-dimension.*

One of these dimensions needs to be singled out: the Natarajan dimension.

Definition 9 (Natarajan dimension (Natarajan, 1989)) *Let* $\mathcal{F}$ *be a class of functions on a set* $\mathcal{X}$ *taking their values in* $[\![1, Q]\!]$*. The* Natarajan dimension *of* $\mathcal{F}$*, N-dim*$(\mathcal{F})$*, is the* Ψ*-dimension of* $\mathcal{F}$ *in the specific case where* $\Psi = \{\psi_{k,l} : 1 \leq k \neq l \leq Q\}$*, such that* $\psi_{k,l}$ *takes the value* 1 *if its argument is equal to* k*, the value* -1 *if its argument is equal to* l*, and* $*$ *otherwise.*

The fat-shattering dimension characterizes the uniform Glivenko-Cantelli classes among the classes of real-valued functions.

Definition 10 (Fat-shattering dimension (Kearns and Schapire, 1994)) *Let* $\mathcal{G}$ *be a class of real-valued functions on a set* $\mathcal{X}$*. For* $\gamma \in \mathbb{R}_+^*$*, a subset* $s_{\mathcal{X}^n} = \{x_i : 1 \leq i \leq n\}$ *of* $\mathcal{X}$ *is said to be* γ-shattered *by* $\mathcal{G}$ *if there is a vector* $v_b = (b_i)$ *in* $\mathbb{R}^n$ *such that, for each vector* $v_y = (y_i)$ *in* $\{-1,1\}^n$*, there is a function* g_y *in* $\mathcal{G}$ *satisfying*

$$\forall i \in [\![1, n]\!],\ y_i\left(g_y(x_i) - b_i\right) \geq \gamma.$$

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

Given the results available for the Ψ-dimensions and the fat-shattering dimension, it appears natural, to study the generalization capabilities of classifiers taking values in $\mathbb{R}^Q$, to consider the use of capacity measures obtained as mixtures of the two concepts, namely scale-sensitive Ψ-dimensions. We now introduce a class of dimensions of this kind, the γ-Ψ-dimensions. In their definition, $\wedge$ denotes the conjunction of two events.

Definition 11 (γ-Ψ-dimensions) *Let* $\mathcal{G}$ *be a class of functions on a set* $\mathcal{X}$ *taking their values in* $\mathbb{R}^Q$*. Let* Ψ *be a family of mappings* ψ *from* $[\![1, Q]\!]$ *into* $\{-1, 1, *\}$*. For* $\gamma \in \mathbb{R}_+^*$*, a subset* $s_{\mathcal{X}^n} = \{x_i : 1 \leq i \leq n\}$ *of* $\mathcal{X}$ *is said to be* γ-Ψ-shattered *by* $\Delta^{\#}\mathcal{G}$ *if there is a mapping* $\psi^n = \left(\psi^{(i)}\right)_{1\leq i\leq n}$

in Ψ^n *and a vector* $v_b = (b_i)$ *in* $\mathbb{R}^n$ *such that, for each vector* $v_y = (y_i)$ *in* $\{-1,1\}^n$*, there is a function* g_y *in* $\mathcal{G}$ *satisfying*

$$\forall i \in [\![1,n]\!], \left\{ \begin{array}{lllll} \text{if } y_i = & 1, & \exists k : \psi^{(i)}(k) = & 1 & \wedge \ \Delta^{\#} g_{y,k}(x_i) - b_i \geq \gamma \\ \text{if } y_i = & -1, & \exists l : \psi^{(i)}(l) = & -1 & \wedge \ \Delta^{\#} g_{y,l}(x_i) + b_i \geq \gamma \end{array} \right. .$$

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

The margin Natarajan dimension is defined accordingly.

Definition 12 (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* $\gamma \in \mathbb{R}_+^*$*, a subset* $s_{\mathcal{X}^n} = \{x_i : 1 \leq i \leq n\}$ *of* $\mathcal{X}$ *is said to be* γ-N-shattered *by* $\Delta^{\#}\mathcal{G}$ *if there is a set* $I(s_{\mathcal{X}^n}) = \{(i_1(x_i), i_2(x_i)) : 1 \leq i \leq n\}$ *of* n *pairs of distinct indices in* $[\![1,Q]\!]$ *and a vector* $v_b = (b_i)$ *in* $\mathbb{R}^n$ *such that, for each vector* $v_y = (y_i)$ *in* $\{-1,1\}^n$*, there is a function* g_y *in* $\mathcal{G}$ *satisfying*

$$\forall i \in [\![1,n]\!], \left\{ \begin{array}{lll} \text{if } y_i = & 1, & \Delta^{\#} g_{y,i_1(x_i)}(x_i) - b_i \geq \gamma \\ \text{if } y_i = & -1, & \Delta^{\#} g_{y,i_2(x_i)}(x_i) + b_i \geq \gamma \end{array} \right. .$$

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*.*

For this scale-sensitive Ψ-dimension, the connection with the covering number of interest, i.e., the generalized Sauer lemma, is the following one.

Theorem 2 (After Theorem 4 in Guermeur (2004)) *Let* $\mathcal{G}$ *be a class of functions from* $\mathcal{X}$ *into* $[-M,M]^Q$*. For every value of* ϵ *in* $(0,M]$ *and every integer value of* n *satisfying* $n \geq$ N*-dim* $(\Delta\mathcal{G}, \epsilon/6)$*, the following bound is true:*

$$\mathcal{N}^{(p)}(\epsilon, \Delta^*\mathcal{G}, n) < 2 \left(n\, Q^2 (Q-1) \left\lfloor \frac{3M}{\epsilon} \right\rfloor^2 \right)^{\lceil d \log_2 \left(en \binom{Q}{2} \left(2 \left\lfloor \frac{3M}{\epsilon} \right\rfloor - 1 \right)/d \right) \rceil}$$

where $d =$ N*-dim* $(\Delta\mathcal{G}, \epsilon/6)$*.*

This theorem is the main novelty in the revised version of Guermeur (2004). What makes it a nontrivial Q-class extension of Lemma 3.5 in Alon et al. (1997) is the presence of both margin operators. The reason why Δ^* appears in the covering number instead of Δ is the very principle at the basis of all the variants of Sauer's lemma: each pair of functions separated with respect

to the functional pseudo-metric used (here d_{x^n}) shatters (at least) one point in $s_{\mathcal{X}^n}$. This is true for $\Delta_\gamma^*\mathcal{G}$, or more precisely its η-discretization, not for $\Delta_\gamma\mathcal{G}$ (see Section 5.3 in Guermeur (2004) for details). One can derive a variant of Theorem 2 involving N-dim $(\Delta^*\mathcal{G}, \epsilon/6)$. However, this alternative is of lesser interest, for reasons that will appear at the end of the following section.

4 Margin Natarajan dimension of the M-SVMs

We now give the sketch of the derivation of an upper bound on the margin Natarajan dimension of interest when $\mathcal{G}$ is the class of functions implemented by the M-SVMs. These large margin classifiers are built around a positive type function (kernel) (Berlinet and Thomas-Agnan, 2004). Let κ be such a kernel on $\mathcal{X}$ and $\left(H_\kappa, \langle\cdot,\cdot\rangle_{H_\kappa}\right)$ the corresponding reproducing kernel Hilbert space (RKHS). The existence and unicity of this space are ensured by the Moore-Aronszajn theorem. According to the Mercer representation theorem, there exists (at least) a mapping Φ on $\mathcal{X}$ satisfying:

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

where $\langle\cdot,\cdot\rangle$ is the dot product of the ℓ_2 space. "The" feature space traditionally designates any of the Hilbert spaces $\left(E_{\Phi(\mathcal{X})}, \langle\cdot,\cdot\rangle\right)$ spanned by the $\Phi\left(\mathcal{X}\right)$. By definition of a RKHS, $\mathcal{H} = \left(\left(H_\kappa, \langle\cdot,\cdot\rangle_{H_\kappa}\right) + \{1\}\right)^Q$ is the class of functions $h = (h_k)_{1\leq k\leq Q}$ from $\mathcal{X}$ into $\mathbb{R}^Q$ of the form:

$$h(\cdot) = \left(\sum_{i=1}^{m_k} \beta_{ik}\kappa\left(x_{ik}, \cdot\right) + b_k\right)_{1\leq k\leq Q}$$

where the x_{ik} are elements of $\mathcal{X}$ (the β_{ik} and b_k are scalars), as well as the limits of these functions as the sets $\{x_{ik} : 1 \leq i \leq m_k\}$ become dense in $\mathcal{X}$, in the norm induced by the dot product. Due to (1), $\mathcal{H}$ can also be seen as a multivariate affine model on $\Phi\left(\mathcal{X}\right)$. Functions h can then be rewritten as:

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

where the vectors w_k are elements of $E_{\Phi(\mathcal{X})}$. They are thus described by the pair $(\mathbf{w}, \mathbf{b})$ with $\mathbf{w} = (w_k)_{1\leq k\leq Q}$ and $\mathbf{b} = (b_k)_{1\leq k\leq Q}$. Let $\bar{\mathcal{H}}$ stand for the product space H_κ^Q. Its norm, $\|\cdot\|_{\bar{\mathcal{H}}}$, is given by $\left\|\bar{h}\right\|_{\bar{\mathcal{H}}} = \sqrt{\sum_{k=1}^Q \|w_k\|^2} = \|\mathbf{w}\|$. For convenience, $E_{\Phi(\mathcal{X})}^Q$ is endowed with a second norm: $\|\cdot\|_\infty$. It is defined by $\|\mathbf{w}\|_\infty = \max_{1\leq k\leq Q} \|w_k\|$. With these definitions at hand, a generic definition of the M-SVMs can be formulated as follows.

Definition 13 (M-SVM) *Let* $((x_i, y_i))_{1\leq i\leq m} \in (\mathcal{X} \times [\![1, Q]\!])^m$. *A* Q-category M-SVM *is a large margin discriminant model obtained by minimizing over the hyperplane* $\sum_{k=1}^Q h_k = 0$ *of* $\mathcal{H}$

an objective function J of the form:

$$J(h) = \sum_{i=1}^{m} \ell_{M\text{-}SVM}(y_i, h(x_i)) + \lambda \|\mathbf{w}\|^2$$

where the data fit component, used in place of the empirical (margin) risk, involves a loss function $\ell_{M\text{-}SVM}$ which is convex.

The three main models of M-SVMs are those of Weston and Watkins (1998), Crammer and Singer (2001), and Lee et al. (2004). It springs from Definition 13 that they only differ in the nature of $\ell_{\text{M-SVM}}$. The specification of this function is such that the introduction of the penalizer $\|\mathbf{w}\|^2$ tends to maximize geometrical margins the definition of which is directly connected with Definition 1. Theorem 1 involves a covering number of $\Delta_\gamma^{\#}\mathcal{G}$ whereas Theorem 2 involves the covering numbers of $\Delta^*\mathcal{G}$, when $\mathcal{G}$ is a class of functions taking their values in a bounded set of the form $[-M, M]^Q$. Furthermore, the class $\bar{\mathcal{H}}$ is easier to handle than the class $\mathcal{H}$. These observations call for the use of hypotheses regarding the volume occupied by data in $E_{\Phi(\mathcal{X})}$, the introduction of constraints on $(\mathbf{w}, \mathbf{b})$, as well as the formulation of an intermediate result relating the covering numbers of $\Delta_\gamma^*\mathcal{H}$ and $\Delta^*\bar{\mathcal{H}}$ as a function of the aforementioned constraints. This result is provided by the following lemma.

Lemma 1 (Lemmas 9 and 10 in Guermeur (2004)) *Let $\mathcal{H}$ be the class of functions that a Q-category M-SVM can implement under the constraint $\mathbf{b} \in [-\beta, \beta]^Q$. Let $\bar{\mathcal{H}}$ be the subset of $\mathcal{H}$ corresponding to the functions satisfying $\mathbf{b} = 0$. Let $(\gamma, \epsilon) \in {\mathbb{R}_+^*}^2$ and $n \in \mathbb{N}^*$. Then*

$$\mathcal{N}^{(p)}(\epsilon, \Delta_\gamma^*\mathcal{H}, n) \leq \left(2\left\lceil\frac{\beta}{\epsilon}\right\rceil + 1\right)^Q \mathcal{N}^{(p)}(\epsilon/2, \Delta^*\bar{\mathcal{H}}, n). \tag{2}$$

A final theorem then completes the construction of the guaranteed risk.

Theorem 3 (Theorem 5 in Guermeur (2004)) *Let $\bar{\mathcal{H}}$ be the class of functions that a Q-category M-SVM can implement under the hypothesis that $\Phi(\mathcal{X})$ is included in the closed ball of radius $\Lambda_{\Phi(\mathcal{X})}$ about the origin in $E_{\Phi(\mathcal{X})}$ and the constraints $\|\mathbf{w}\|_\infty \leq \Lambda_w$ and $\mathbf{b} = 0$. Then, for any positive real value ϵ, the following bound holds true:*

$$N\text{-}dim\left(\Delta\bar{\mathcal{H}}, \epsilon\right) \leq \binom{Q}{2}\left(\frac{\Lambda_w \Lambda_{\Phi(\mathcal{X})}}{\epsilon}\right)^2.$$

The proof combines the line of argument of the corresponding bi-class result, Theorem 4.6 in Bartlett and Shawe-Taylor (1999), with the application of the pigeonhole principle. This involves a generalization of Lemma 4.2 in Bartlett and Shawe-Taylor (1999) which can only be obtained for the Δ operator. This remark completes the discussion on the presence of both margin operators

in Theorem 2. Putting things together, the control term of the guaranteed risk obtained decreases with the size of the training sample as $\ln(m) \cdot m^{-1/2}$. This represents an improvement over the rate obtained by Guermeur et al. (2005), namely $m^{-1/4}$.

In short, what we have done so far is to derive a general purpose bound (Theorems 1 and 2), and dedicate it in the end to M-SVMs (Lemma 1 and Theorem 3). Obviously, improvements should result from applying Dudley's method of chaining (Dudley, 1984), as well as optimizing the choice of the pseudo-metrics and norms. In the following section, making full use of the fact that M-SVMs are based on a RKHS, we obtain a sharper bound.

5 Bound on the risk of M-SVMs based on the Rademacher complexity

The bound established in this section rests on the hypothesis and the constraints introduced in the preceding section, plus the constraint $\mathbf{b} = 0$ (we work with $\bar{\mathcal{H}}$). It is directly inspired from bi-class results exposed in Sections 3 and 4 in Boucheron et al. (2005). We start by giving the definition of the Rademacher complexity. For $n \in \mathbb{N}^*$, a Bernoulli or Rademacher sequence σ is a sequence $(\sigma_i)_{1 \leq i \leq n}$ of i.i.d. Bernoulli random variables for which the common value of the parameter p is $\frac{1}{2}$.

Definition 14 (Rademacher complexity) *Let $\mathcal{F}$ be a class of real-valued functions with domain $\mathcal{T}$. For $n \in \mathbb{N}^*$, let $T = (T_i)_{1 \leq i \leq n}$ be a sequence of n i.i.d. random variables taking values in $\mathcal{T}$ and let $\sigma = (\sigma_i)_{1 \leq i \leq n}$ be a Rademacher sequence. The* Rademacher complexity *of $\mathcal{F}$ is*

$$\mathcal{R}_n(\mathcal{F}) = \mathbb{E}_{\sigma T} \sup_{f \in \mathcal{F}} \frac{2}{n} \left| \sum_{i=1}^{n} \sigma_i f(T_i) \right|.$$

The use of the Rademacher complexity, and more generally Rademacher averages, is central in many results in the theory of empirical processes. For those averages, a vast set of properties is available, among which the *contraction principle*, which will prove useful in the sequel.

Theorem 4 (After Theorem 4.15 in Shawe-Taylor and Cristianini (2004)) *$\mathcal{F}$ and n being defined as in Definition 14, let $\phi : \mathbb{R} \longrightarrow \mathbb{R}$ be a function satisfying the Lipschitz condition with constant L_ϕ such that $\phi(0) = 0$. Then*

$$\mathcal{R}_n(\phi \circ \mathcal{F}) \leq 2 L_\phi \mathcal{R}_n(\mathcal{F}).$$

As in the case of Theorem 1, the derivation of the bound exposed in this section is based on the definition of a new margin risk and the use of a concentration inequality. Taking our inspiration from the loss function of the M-SVM of Crammer and Singer (2001), $\ell_{\text{M-SVM}}(y, h(x)) =$

$(1 - h_y(x) + \max_{k \neq y} h_k(x))_+$, where $(t)_+ = \max(0, t)$, we consider the following convexified risk:

$$\tilde{R}(h) = \mathbb{E}\left[(1 - \Delta h_Y(X))_+\right].$$

$\tilde{R}_m(h)$ designates the corresponding empirical risk, measured on a m-sample. The concentration inequality used here is the bounded differences inequality.

Theorem 5 (Bounded differences inequality (McDiarmid, 1989)) *For $n \in \mathbb{N}^*$, let $(T_i)_{1 \leq i \leq n}$ be a sequence of n independent random variables taking values in a set $\mathcal{T}$. Let f be a function from $\mathcal{T}^n$ into $\mathbb{R}$ such that there exists a sequence of nonnegative constants $(c_i)_{1 \leq i \leq n}$ satisfying:*

$$\forall i \in [\![1, n]\!], \quad \sup_{(t_i)_{1 \leq i \leq n} \in \mathcal{T}^n, t_i' \in \mathcal{T}} \left| f(t_1, \ldots, t_n) - f(t_1, \ldots, t_{i-1}, t_i', t_{i+1}, \ldots, t_n) \right| \leq c_i.$$

Then, for any positive value of τ, the random variable $f(T_1, \ldots, T_n)$ satisfies the following inequalities:

$$\mathbb{P}\left\{f(T_1, \ldots, T_n) - \mathbb{E}f(T_1, \ldots, T_n) > \tau\right\} \leq e^{-\frac{2\tau^2}{c}}$$

and

$$\mathbb{P}\left\{\mathbb{E}f(T_1, \ldots, T_n) - f(T_1, \ldots, T_n) > \tau\right\} \leq e^{-\frac{2\tau^2}{c}}$$

where $c = \sum_{i=1}^n c_i^2$.

These definitions and basic results being given, setting $K_{\bar{\mathcal{H}}} = \Lambda_w \Lambda_{\Phi(\mathcal{X})} + 1$, we demonstrate the following bound.

Theorem 6 *Let $\bar{\mathcal{H}}$ be the class of functions that a Q-category M-SVM can implement under the hypothesis that $\Phi(\mathcal{X})$ is included in the closed ball of radius $\Lambda_{\Phi(\mathcal{X})}$ about the origin in $E_{\Phi(\mathcal{X})}$ and the constraints $\|\mathbf{w}\|_\infty \leq \Lambda_w$ and $\mathbf{b} = 0$. Let $\delta \in (0, 1)$. With probability at least $1 - \delta$, the risk of any function $\bar{h}$ in $\bar{\mathcal{H}}$ is bounded from above by:*

$$R(\bar{h}) \leq \tilde{R}_m(\bar{h}) + \frac{4}{\sqrt{m}} + \frac{4Q(Q-1)\Lambda_w \Lambda_{\Phi(\mathcal{X})}}{\sqrt{m}} + K_{\bar{\mathcal{H}}} \sqrt{\frac{\ln\left(\frac{1}{\delta}\right)}{2m}}. \tag{3}$$

Proof Since $\mathbb{1}_{\{\Delta \bar{h}_Y(X) \leq 0\}} \leq \left(1 - \Delta \bar{h}_Y(X)\right)_+$, we have:

$$\forall \bar{h} \in \bar{\mathcal{H}}, \; R(\bar{h}) \leq \tilde{R}(\bar{h}).$$

Consequently,

$$\forall \bar{h} \in \bar{\mathcal{H}}, \; R(\bar{h}) \leq \tilde{R}_m(\bar{h}) + \sup_{\bar{h}' \in \bar{\mathcal{H}}} \left(\tilde{R}(\bar{h}') - \tilde{R}_m(\bar{h}')\right). \tag{4}$$

The rest of the proof consists in the computation of an upper bound on the supremum of the empirical process appearing in (4). Let Z denote the random pair (X, Y) and Z_i its copies

which constitute the m-sample D_m: $D_m = (Z_i)_{1\leq i\leq m}$. After simplifying notation this way, the bounded differences inequality can be applied to the supremum of interest by setting $n = m$, $(T_i)_{1\leq i\leq n} = D_m$ (i.e., $T_i = Z_i$), and $f(T_1, \ldots, T_n) = \sup_{\bar{h}\in\bar{\mathcal{H}}} \left(\tilde{R}(\bar{h}) - \tilde{R}_m(\bar{h})\right)$. Due to the hypotheses of Theorem 6, the functions $\bar{h}$ of $\bar{\mathcal{H}}$, and thus the functions $\Delta\bar{h}$ of $\Delta\bar{\mathcal{H}}$, take their values in $[-M_{\bar{\mathcal{H}}}, M_{\bar{\mathcal{H}}}]^Q$ with $M_{\bar{\mathcal{H}}} = \Lambda_w \Lambda_{\Phi(\mathcal{X})}$. Consequently, the loss function associated with the risk $\tilde{R}$ takes its values in the interval $[0, K_{\bar{\mathcal{H}}}]$. One can thus choose the sequence $(c_i)_{1\leq i\leq m}$ in the following way: $\forall i \in [\![1, m]\!]$, $c_i = \frac{K_{\bar{\mathcal{H}}}}{m}$. Since we are only interested in computing an upper bound on the supremum, it is the first inequality which is used. We then get the following result: with probability at least $1 - \delta$,

$$\sup_{\bar{h}\in\bar{\mathcal{H}}} \left(\tilde{R}(\bar{h}) - \tilde{R}_m(\bar{h})\right) \leq \mathbb{E}_{D_m} \sup_{\bar{h}\in\bar{\mathcal{H}}} \left(\tilde{R}(\bar{h}) - \tilde{R}_m(\bar{h})\right) + K_{\bar{\mathcal{H}}} \sqrt{\frac{\ln\left(\frac{1}{\delta}\right)}{2m}}. \tag{5}$$

The introduction of a ghost sample $D'_m = ((X'_i, Y'_i))_{1\leq i\leq m} = (Z'_i)_{1\leq i\leq m}$, exhibiting the same properties as the initial sample D_m, and independent of this sample, makes it possible to apply a symmetrization. Let $\tilde{R}_{D_m}(\bar{h})$ and $\tilde{R}_{D'_m}(\bar{h})$ be the empirical margin risks respectively associated with D_m and D'_m (we have thus $\tilde{R}_{D_m}(\bar{h}) = \tilde{R}_m(\bar{h})$). Then,

$$\mathbb{E}_{D_m} \sup_{\bar{h}\in\bar{\mathcal{H}}} \left(\tilde{R}(\bar{h}) - \tilde{R}_m(\bar{h})\right) = \mathbb{E}_{D_m} \sup_{\bar{h}\in\bar{\mathcal{H}}} \left(\mathbb{E}_{D'_m}\left[\tilde{R}_{D'_m}(\bar{h}) - \tilde{R}_{D_m}(\bar{h}) | D_m\right]\right).$$

Since the supremum is convex, applying Jensen's inequality gives:

$$\mathbb{E}_{D_m} \sup_{\bar{h}\in\bar{\mathcal{H}}} \left(\mathbb{E}_{D'_m}\left[\tilde{R}_{D'_m}(\bar{h}) - \tilde{R}_{D_m}(\bar{h}) | D_m\right]\right) \leq \mathbb{E}_{D_{2m}} \sup_{\bar{h}\in\bar{\mathcal{H}}} \left(\tilde{R}_{D'_m}(\bar{h}) - \tilde{R}_{D_m}(\bar{h})\right),$$

where D_{2m} denotes the concatenation of the samples D_m and D'_m. By substitution into (5), we get:

$$\sup_{\bar{h}\in\bar{\mathcal{H}}} \left(\tilde{R}(\bar{h}) - \tilde{R}_m(\bar{h})\right) \leq \mathbb{E}_{D_{2m}} \sup_{\bar{h}\in\bar{\mathcal{H}}} \left(\tilde{R}_{D'_m}(\bar{h}) - \tilde{R}_{D_m}(\bar{h})\right) + K_{\bar{\mathcal{H}}} \sqrt{\frac{\ln\left(\frac{1}{\delta}\right)}{2m}}.$$

$$\mathbb{E}_{D_{2m}} \sup_{\bar{h}\in\bar{\mathcal{H}}} \left(\tilde{R}_{D'_m}(\bar{h}) - \tilde{R}_{D_m}(\bar{h})\right) = \mathbb{E}_{D_{2m}} \left[\sup_{\bar{h}\in\bar{\mathcal{H}}} \frac{1}{m} \sum_{i=1}^{m} \left(\left(1 - \Delta\bar{h}_{Y'_i}(X'_i)\right)_+ - \left(1 - \Delta\bar{h}_{Y_i}(X_i)\right)_+\right)\right]$$

$$\leq \mathbb{E}_{D_{2m}} \left[\sup_{\bar{h}\in\bar{\mathcal{H}}} \frac{1}{m} \left|\sum_{i=1}^{m} \left(\left(1 - \Delta\bar{h}_{Y'_i}(X'_i)\right)_+ - \left(1 - \Delta\bar{h}_{Y_i}(X_i)\right)_+\right)\right|\right].$$

At this level, the introduction of a weighting with Rademacher variables can be seen as the application on D_{2m} of a permutation belonging to the "swapping" subgroup of $\mathfrak{S}_{2m}$, the symmetric group of degree $2m$. Since coordinate permutations preserve the product distribution P^{2m}, we get:

$$\mathbb{E}_{D_{2m}} \sup_{\bar{h}\in\bar{\mathcal{H}}} \left(\tilde{R}_{D'_m}(\bar{h}) - \tilde{R}_{D_m}(\bar{h})\right) \leq \mathbb{E}_{\sigma D_{2m}} \left[\sup_{\bar{h}\in\bar{\mathcal{H}}} \frac{1}{m} \left|\sum_{i=1}^{m} \sigma_i \left(\left(1 - \Delta\bar{h}_{Y'_i}(X'_i)\right)_+ - \left(1 - \Delta\bar{h}_{Y_i}(X_i)\right)_+\right)\right|\right].$$

$$\mathbb{E}_{D_{2m}} \sup_{\bar{h}\in\bar{\mathcal{H}}} \left(\tilde{R}_{D'_m}(\bar{h}) - \tilde{R}_{D_m}(\bar{h})\right) \leq 2\mathbb{E}_{\sigma D_m}\left[\sup_{\bar{h}\in\bar{\mathcal{H}}} \frac{1}{m}\left|\sum_{i=1}^{m}\sigma_i\left(1-\Delta\bar{h}_{Y_i}(X_i)\right)_+\right|\right]. \tag{6}$$

For all $\bar{h}\in\bar{\mathcal{H}}$, let $f_{\bar{h}}$ be the real-value function defined on $\mathcal{X}\times\mathcal{Y}$ by:

$$\forall (x,y)\in\mathcal{X}\times\mathcal{Y},\ f_{\bar{h}}(x,y) = 1-\Delta\bar{h}_y(x).$$

Let $\mathcal{F}_{\bar{\mathcal{H}}} = \left\{f_{\bar{h}} : \bar{h}\in\bar{\mathcal{H}}\right\}$. The right-hand side of (6) can be rewritten as follows

$$2\mathbb{E}_{\sigma D_m}\left[\sup_{f_{\bar{h}}\in\mathcal{F}_{\bar{\mathcal{H}}}} \frac{1}{m}\left|\sum_{i=1}^{m}\sigma_i\left(f_{\bar{h}}(X_i,Y_i)\right)_+\right|\right] = \mathcal{R}_m\left(\left(\mathcal{F}_{\bar{\mathcal{H}}}\right)_+\right).$$

Since the function $(\cdot)_+$ satisfies the Lipschitz condition with constant 1, applying the contraction principle gives:

$$\begin{aligned}
\mathbb{E}_{\sigma D_m}\left[\sup_{\bar{h}\in\bar{\mathcal{H}}} \frac{1}{m}\left|\sum_{i=1}^{m}\sigma_i\left(1-\Delta\bar{h}_{Y_i}(X_i)\right)_+\right|\right] &\leq 2\mathbb{E}_{\sigma D_m}\left[\sup_{\bar{h}\in\bar{\mathcal{H}}} \frac{1}{m}\left|\sum_{i=1}^{m}\sigma_i\left(1-\Delta\bar{h}_{Y_i}(X_i)\right)\right|\right] \\
&\leq 2\left(\mathbb{E}_\sigma\left[\frac{1}{m}\left|\sum_{i=1}^{m}\sigma_i\right|\right] + \mathbb{E}_{\sigma D_m}\left[\sup_{\bar{h}\in\bar{\mathcal{H}}}\frac{1}{m}\left|-\sum_{i=1}^{m}\sigma_i\Delta\bar{h}_{Y_i}(X_i)\right|\right]\right) \\
&\leq 2\left(\frac{1}{\sqrt{m}} + \mathbb{E}_{\sigma D_m}\left[\sup_{\bar{h}\in\bar{\mathcal{H}}}\frac{1}{m}\left|\sum_{i=1}^{m}\sigma_i\Delta\bar{h}_{Y_i}(X_i)\right|\right]\right).
\end{aligned}$$

$$\mathbb{E}_{\sigma D_m}\left[\sup_{\bar{h}\in\bar{\mathcal{H}}}\left|\sum_{i=1}^{m}\sigma_i\Delta\bar{h}_{Y_i}(X_i)\right|\right] = \mathbb{E}_{\sigma D_m}\left[\sup_{\bar{h}\in\bar{\mathcal{H}}}\left|\sum_{i=1}^{m}\sigma_i\frac{1}{2}\left(\bar{h}_{Y_i}(X_i) - \max_{k\neq Y_i}\bar{h}_k(X_i)\right)\right|\right].$$

The computations performed so far are not proper to M-SVMs but in the expression of the constant $K_{\bar{\mathcal{H}}}$. In the sequel, we make full use of the fact that the class $\bar{\mathcal{H}}$ is built around a RKHS induced by the kernel κ. For $n\in\mathbb{N}^*$, let $z^n = ((x_i,y_i))_{1\leq i\leq n}\in(\mathcal{X}\times\mathcal{Y})^n$ and let cat be a mapping from $\bar{\mathcal{H}}\times\mathcal{X}\times\mathcal{Y}$ into $[\![1,Q]\!]^2$ such that

$$\forall\left(\bar{h},x,y\right)\in\bar{\mathcal{H}}\times\mathcal{X}\times\mathcal{Y},\ \mathrm{cat}\left(\bar{h},x,y\right) = (k,l) \Longrightarrow (k=y)\wedge(l\neq y)\wedge\left(\bar{h}_l(x) = \max_{p\neq y}\bar{h}_p(x)\right).$$

By construction of the mapping cat,

$$\begin{aligned}
\forall z^m\in(\mathcal{X}\times\mathcal{Y})^m,\ \frac{1}{2}\mathbb{E}_\sigma&\left[\sup_{\bar{h}\in\bar{\mathcal{H}}}\left|\sum_{i=1}^{m}\sigma_i\left(\bar{h}_{y_i}(x_i) - \max_{k\neq y_i}\bar{h}_k(x_i)\right)\right|\right] \\
&\leq \frac{1}{2}\mathbb{E}_\sigma\left[\sup_{\bar{h}\in\bar{\mathcal{H}}}\left|\sum_{k\neq l}\sum_{i:\ \mathrm{cat}\left(\bar{h},x_i,y_i\right)=(k,l)}\sigma_i\left(\bar{h}_k(x_i) - \bar{h}_l(x_i)\right)\right|\right].
\end{aligned}$$

Since the reproducing property implies that $\bar{h}_k(x_i) - \bar{h}_l(x_i) = \langle\bar{h}_k - \bar{h}_l, \kappa(x_i,\cdot)\rangle$,

$$\frac{1}{2}\mathbb{E}_\sigma\left[\sup_{\bar{h}\in\bar{\mathcal{H}}}\left|\sum_{i=1}^{m}\sigma_i\left(\bar{h}_{y_i}(x_i) - \max_{k\neq y_i}\bar{h}_k(x_i)\right)\right|\right] \leq \frac{1}{2}\mathbb{E}_\sigma\left[\sup_{\bar{h}\in\bar{\mathcal{H}}}\sum_{k\neq l}\left|\sum_{i:\ \mathrm{cat}\left(\bar{h},x_i,y_i\right)=(k,l)}\sigma_i\langle\bar{h}_k - \bar{h}_l, \kappa(x_i,\cdot)\rangle\right|\right]. \tag{7}$$

Due to the constraint $\|\mathbf{w}\|_\infty \leq \Lambda_w$, the Cauchy-Schwarz inequality provides us with the following upper bound: $\forall \bar{h} \in \bar{\mathcal{H}},\ \forall (k,l) \in [\![1,Q]\!]^2,\ k \neq l$,

$$\frac{1}{2}\left|\sum_{i:\ \mathrm{cat}\left(\bar{h},x_i,y_i\right)=(k,l)} \sigma_i \langle \bar{h}_k - \bar{h}_l, \kappa\left(x_i,\cdot\right)\rangle\right| = \frac{1}{2}\left|\langle \bar{h}_k - \bar{h}_l, \sum_{i:\ \mathrm{cat}\left(\bar{h},x_i,y_i\right)=(k,l)} \sigma_i \kappa\left(x_i,\cdot\right)\rangle\right|$$

$$\leq \Lambda_w \left\| \sum_{i:\ \mathrm{cat}\left(\bar{h},x_i,y_i\right)=(k,l)} \sigma_i \kappa\left(x_i,\cdot\right)\right\|.$$

Thus,

$$\frac{1}{2}\mathbb{E}_\sigma\left[\sup_{\bar{h}\in\bar{\mathcal{H}}}\sum_{k\neq l}\left|\sum_{i:\ \mathrm{cat}\left(\bar{h},x_i,y_i\right)=(k,l)} \sigma_i \langle \bar{h}_k - \bar{h}_l, \kappa\left(x_i,\cdot\right)\rangle\right|\right] \leq \Lambda_w \mathbb{E}_\sigma\left[\sup_{\bar{h}\in\bar{\mathcal{H}}}\sum_{k\neq l}\left\|\sum_{i:\ \mathrm{cat}\left(\bar{h},x_i,y_i\right)=(k,l)} \sigma_i \kappa\left(x_i,\cdot\right)\right\|\right].$$

Let $\mathcal{P}_m$ be the set of all mappings p_m from $[\![1,m]\!]$ into $[\![1,Q]\!]^2$ such that for all value of i, the pair $p_m(i)$ is always made up of two different values.

$$\Lambda_w \mathbb{E}_\sigma\left[\sup_{\bar{h}\in\bar{\mathcal{H}}}\sum_{k\neq l}\left\|\sum_{i:\ \mathrm{cat}\left(\bar{h},x_i,y_i\right)=(k,l)} \sigma_i \kappa\left(x_i,\cdot\right)\right\|\right] \leq \Lambda_w \sum_{k\neq l}\mathbb{E}_\sigma\left[\sup_{p_m\in\mathcal{P}_m}\left\|\sum_{i:\ p_m(i)=(k,l)} \sigma_i \kappa\left(x_i,\cdot\right)\right\|\right]. \tag{8}$$

Consequently, to complete the derivation of the bound, it suffices to find a uniform upper bound on the expressions of the form:

$$\mathbb{E}_\sigma\left\|\sum_{i\in\mathcal{I}_m}\sigma_i\kappa\left(x_i,\cdot\right)\right\|,$$

where $\mathcal{I}_m$ is a subset of $[\![1,m]\!]$.

$$\mathbb{E}_\sigma\left\|\sum_{i\in\mathcal{I}_m}\sigma_i\kappa\left(x_i,\cdot\right)\right\| = \mathbb{E}_\sigma\left[\langle\sum_{i\in\mathcal{I}_m}\sigma_i\kappa\left(x_i,\cdot\right), \sum_{j\in\mathcal{I}_m}\sigma_j\kappa\left(x_j,\cdot\right)\rangle^{\frac{1}{2}}\right]. \tag{9}$$

By application of Jensen's inequality, the right-hand side of Equation 9 is bounded from above as follows:

$$\mathbb{E}_\sigma\left[\langle\sum_{i\in\mathcal{I}_m}\sigma_i\kappa\left(x_i,\cdot\right), \sum_{j\in\mathcal{I}_m}\sigma_j\kappa\left(x_j,\cdot\right)\rangle^{\frac{1}{2}}\right] \leq \left(\mathbb{E}_\sigma\left[\sum_{i\in\mathcal{I}_m}\sum_{j\in\mathcal{I}_m}\sigma_i\sigma_j\kappa\left(x_i,x_j\right)\right]\right)^{\frac{1}{2}} = \left(\sum_{i\in\mathcal{I}_m}\kappa\left(x_i,x_i\right)\right)^{\frac{1}{2}}.$$

Since κ is a positive type function, $\kappa\left(x_i,x_i\right) \geq 0$, and thus

$$\forall \mathcal{I}_m \subset [\![1,m]\!],\ \mathbb{E}_\sigma\left\|\sum_{i\in\mathcal{I}_m}\sigma_i\kappa\left(x_i,\cdot\right)\right\| \leq \left(\sum_{i\in\mathcal{I}_m}\kappa\left(x_i,x_i\right)\right)^{\frac{1}{2}} \leq \left(\sum_{i=1}^{m}\kappa\left(x_i,x_i\right)\right)^{\frac{1}{2}} \leq \Lambda_{\Phi(\mathcal{X})}\sqrt{m}.$$

(the proof of the partial result $\mathbb{E}_\sigma\left[\frac{1}{m}\left|\sum_{i=1}^m \sigma_i\right|\right] \leq \frac{1}{\sqrt{m}}$ we used earlier is exactly the same). To sum up:

$$\forall (k,l) \in [\![1,Q]\!]^2,\ k \neq l,\ \mathbb{E}_\sigma\left[\sup_{p_m \in \mathcal{P}_m}\left\|\sum_{i:\ p_m(i)=(k,l)} \sigma_i \kappa\left(x_i,\cdot\right)\right\|\right] \leq \Lambda_{\Phi(\mathcal{X})}\sqrt{m}.$$

By substitution in the right-hand side of (8), and then in the right-hand side of (7), we get

$$\forall z^m \in (\mathcal{X}\times\mathcal{Y})^m,\ \frac{1}{2}\mathbb{E}_\sigma\left[\sup_{\bar{h}\in\bar{\mathcal{H}}}\left|\sum_{i=1}^m \sigma_i\left(\bar{h}_{y_i}\left(x_i\right) - \max_{k\neq y_i}\bar{h}_k\left(x_i\right)\right)\right|\right] \leq Q(Q-1)\Lambda_w\Lambda_{\Phi(\mathcal{X})}\sqrt{m}$$

which implies that

$$\mathbb{E}_{\sigma D_m}\left[\sup_{\bar{h}\in\bar{\mathcal{H}}}\frac{1}{m}\left|\sum_{i=1}^m \sigma_i \Delta\bar{h}_{Y_i}\left(X_i\right)\right|\right] \leq \frac{Q(Q-1)\Lambda_w\Lambda_{\Phi(\mathcal{X})}}{\sqrt{m}}.$$

Gathering all the partial results produces the bound (3), which concludes the proof. ■

In this bound, the control term is a $O\left(m^{-1/2}\right)$. We have thus gained a $\ln(m)$ factor compared to the bound involving the margin Natarajan dimension.

6 Conclusions and ongoing research

A new class of generalized VC dimensions dedicated to large margin multi-category discriminant models has been introduced: the γ-Ψ-dimensions. They can be seen either as multivariate extensions of the fat-shattering dimension or scale-sensitive Ψ-dimensions. Their finiteness (for all positive values of the scale parameter γ) characterizes learnability for the classes of functions considered. Their introduction thus bridges an important gap in the VC theory. Furthermore, using the margin Natarajan dimension, we derived a bound on the risk of M-SVMs where the control term decreases with the size of the training sample as $\ln(m)\cdot m^{-1/2}$. Although this guaranteed risk is tighter than the one established by Guermeur et al. (2005), the sharpest rate of convergence we obtained for these machines resulted from using the Rademacher complexity. We conjecture that this rate of convergence, in $m^{-1/2}$, could also be attained using γ-Ψ-dimensions. Thanks to Dudley's method of chaining (Dudley, 1984), the VC bound could obviously be improved by a factor $\sqrt{\ln(m)}$. The difficulty thus consists in improving the generalized Sauer Lemma (Theorem 2) in order to get rid of the remaining $\sqrt{\ln(m)}$ factor. This corresponds to deriving an upper bound on the covering number of interest growing polynomially with m, to replace the current $m^{O(\ln(m))}$. This could be possible since this dependence is precisely the one of Sauer's (original) lemma. Obtaining these improvements is the subject of an ongoing research.

Acknowledgments

The author would like to thank O. Bousquet for introducing him to the use of Rademacher averages. Thanks are also due to L. Ralaivola for checking the proof of Theorem 6.

References

N. Alon, S. Ben-David, N. Cesa-Bianchi, and D. Haussler. Scale-sensitive dimensions, uniform convergence, and learnability. *Journal of the ACM*, 44(4):615–631, 1997.

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.

P.L. Bartlett and J. Shawe-Taylor. Generalization performance of support vector machines and other pattern classifiers. In B. Schölkopf, C.J.C. Burges, and A. Smola, editors, *Advances in Kernel Methods, Support Vector Learning*, chapter 4, pages 43–54. The MIT Press, Cambridge, MA, 1999.

S. Ben-David, N. Cesa-Bianchi, D. Haussler, and P.M. Long. Characterizations of learnability for classes of $\{0, \ldots, n\}$-valued functions. *Journal of Computer and System Sciences*, 50(1): 74–86, 1995.

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

S. Boucheron, O. Bousquet, and G. Lugosi. Theory of classification: A survey of some recent advances. *ESAIM: Probability and Statistics*, 9:323–375, 2005.

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

R.M. Dudley. A course on empirical processes. In P.L. Hennequin, editor, *Ecole d'Eté de Probabilités de Saint-Flour XII - 1982*, volume 1097 of *Lecture Notes in Mathematics*, pages 1–142. Springer-Verlag, 1984.

Y. Guermeur. Large margin multi-category discriminant models and scale-sensitive Ψ-dimensions. Technical Report RR-5314, INRIA, 2004. (revised in 2006).

Y. Guermeur, M. Maumy, and F. Sur. Model selection for multi-class SVMs. In *International Symposium on Applied Stochastic Models and Data Analysis*, pages 507–517, 2005.

M.J. Kearns and R.E. Schapire. Efficient distribution-free learning of probabilistic concepts. *Journal of Computer and System Sciences*, 48(3):464–497, 1994.

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.

C. McDiarmid. On the method of bounded differences. *Surveys in Combinatorics*, 141:148–188, 1989. Cambridge University Press.

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

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

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

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