Scale-sensitive Ψ-dimensions: the Capacity Measures for Classifiers Taking Values in $\mathbb{R}^Q$

Yann Guermeur

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

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 $\{0,1\}$, $\{1,\ldots,Q\}$ 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 which appears superior to the existing one.
Keywords: Large margin classifiers, Generalized VC dimensions, M-SVMs.

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 measure of this capacity is the famous Vapnik-Chervonenkis (VC) dimension. Extensions have also been proposed for real-valued bi-class models and multi-class models taking theirs 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 exhibit this property, such as the multi-layer perceptrons, or the multi-class support vector machines (M-SVMs). In this paper, the scale-sensitive Ψ-dimensions are introduced to fill this gap. A generalization of Sauer's lemma [Sauer, 1972] is given, which relates the covering numbers 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 bounded from above for the architecture shared by all the M-SVMs proposed so far. This provides us with a sharper bound on their sample 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 scale-sensitive Ψ-dimensions are defined, and the generalized Sauer lemma is formulated. The upper bound on the margin Natarajan dimension of the M-SVMs is then described in Section 4. For lack of space, proofs are omitted. They can be found in [Guermeur, 2004].

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, $\{1, \ldots, Q\}$. The link between patterns and categories is supposed to be probabilistic. $\mathcal{X}$ and $\mathcal{Y}$ are probability spaces, and $\mathcal{X} \times \mathcal{Y}$ is endowed with a probability measure P, fixed but unknown. Let (X, Y) be a random pair distributed according to P. Training consists in using a m-sample $s_m = ((X_i, Y_i))_{1 \leq i \leq m}$ of independent copies of (X, Y) to select, in a given class of functions $\mathcal{G}$, a function classifying data in an optimal way. The criterion to be optimized, the *risk*, is the expectation with respect to P of a given loss function. The way the functions in $\mathcal{G}$ perform classification must be specified. We consider classes of functions from $\mathcal{X}$ into $\mathbb{R}^Q$. $g = (g_k)_{1 \leq k \leq Q} \in \mathcal{G}$ 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 $\mathcal{G} \times \mathcal{X} \times \mathcal{Y}$ by $\ell\left(y, g(x)\right) = \mathbb{1}_{\{g_y(x) \leq \max_{k \neq y} g_k(x)\}}$. The risk of g is then given by:

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

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 $(x, y) \in \mathcal{X} \times \mathcal{Y}$, $\mathcal{M}_{xy}(g, x, y)$, is given by:

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

Basically, the central elements to assign a pattern to a category and to derive a level of confidence in this assignation 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 appropriate operator. Two such "margin operators" are considered here, Δ and Δ^*.

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

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

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

$\forall\, (g, x) \in \mathcal{G} \times \mathcal{X}$, let $\mathcal{M}_x(g, x) = \max_k \Delta g_k(x)$.

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

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

In the sequel, $\Delta^\#$ is used in place of Δ and Δ^* in the formulas that hold true for both operators. The empirical margin risk is defined as follows.

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

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

For technical reasons, it is useful to squash the functions $\Delta^\# g_k$ as much as possible without altering the value of the empirical 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:

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

Let $\Delta_\gamma^\#$ denote $\pi_\gamma \circ \Delta^\#$ and $\Delta_\gamma^\#\mathcal{G}$ 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 *$\epsilon$-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 cover, 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 (functional pseudo-metric). Let $\mathcal{G}$ be a class of functions from $\mathcal{X}$ into $\mathbb{R}^Q$. For a set $s_{\mathcal{X}^n} \subset \mathcal{X}$ of cardinality n, define the pseudo-metric $d_{\ell_\infty,\ell_\infty(s_{\mathcal{X}^n})}$ on $\mathcal{G}$ as:

$$\forall (g,g') \in \mathcal{G}^2,\ d_{\ell_\infty,\ell_\infty(s_{\mathcal{X}^n})}(g,g') = \max_{x\in s_{\mathcal{X}^n}} \left\|g(x) - g'(x)\right\|_\infty.$$

Let $\mathcal{N}^{(p)}_{\infty,\infty}(\epsilon, \Delta^{\#}_{\gamma}\mathcal{G}, n) = \sup_{s_{\mathcal{X}^n} \subset \mathcal{X}} \mathcal{N}^{(p)}(\epsilon, \Delta^{\#}_{\gamma}\mathcal{G}, d_{\ell_\infty, \ell_\infty(s_{\mathcal{X}^n})})$. The following theorem 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 a probability distribution on $\mathcal{X} \times \mathcal{Y}$. With probability at least $1 - \delta$, for every value of γ in $(0, 1]$, the risk of any function g in a class $\mathcal{G}$ is bounded from above by:*

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

Studying the sample complexity of a classifier $\mathcal{G}$ can thus amount to computing an upper bound on $\mathcal{N}^{(p)}_{\infty,\infty}(\gamma/4, \Delta^{\#}_{\gamma}\mathcal{G}, 2m)$. In [Guermeur *et al.*, 2005], we reached this goal by relating these numbers 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 $\{1, \ldots, Q\}$-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 $\{1, \ldots, Q\}$. Let Ψ be a set of mappings ψ from $\{1, \ldots, Q\}$ 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 $\psi^n = \left(\psi^{(1)}, \ldots, \psi^{(i)}, \ldots, \psi^{(n)}\right)$ in Ψ^n such that for each vector v_y of $\{-1, 1\}^n$, 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 it is finite, or infinity otherwise.

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

Definition 9 (Natarajan dimension [Ben-David *et al.*, 1995]). Let $\mathcal{F}$ be a class of functions on a set $\mathcal{X}$ taking their values in $\{1, \ldots, Q\}$. The *Natarajan dimension* of $\mathcal{F}$, N-dim$(\mathcal{F})$, is the Ψ-dimension of $\mathcal{F}$ in the specific case where Ψ is the set of $Q(Q-1)$ mappings $\psi_{k,l}$, $(1 \leq k \neq l \leq Q)$, 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 [Alon *et al.*, 1997]). Let $\mathcal{G}$ be a class of functions from $\mathcal{X}$ into $\mathbb{R}$. For $\gamma \in \mathbb{R}_+^*$, $s_{\mathcal{X}^n} = \{x_i : 1 \leq i \leq n\} \subset \mathcal{X}$ 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

$$\forall i \in \{1, \ldots, n\}\,,\ y_i \left(g_y(x_i) - b_i\right) \geq \gamma.$$

The *fat-shattering dimension* of $\mathcal{G}$, $P_\gamma\text{-dim}\,(\mathcal{G})$, is the maximal cardinality of a subset of $\mathcal{X}$ γ-shattered by $\mathcal{G}$, if it is finite, or infinity otherwise.

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.

Definition 11 (Ψ-dimension with margin γ). 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 $\{1, \ldots, Q\}$ 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 $\psi^n = \left(\psi^{(1)}, \ldots, \psi^{(i)}, \ldots, \psi^{(n)}\right)$ in Ψ^n and a vector $v_b = (b_i)$ in $\mathbb{R}^n$ such that, for each vector $v_y = (y_i)$ of $\{-1, 1\}^n$, there is a function g_y in $\mathcal{G}$ satisfying

$$\forall i \in \{1, \ldots, n\}\,,\ \begin{cases} \text{if } y_i = \ \ 1, \exists k : \psi^{(i)}(k) = \ \ 1 \wedge\ \Delta^\# g_{y,k}(x_i) - b_i \geq \gamma \\ \text{if } y_i = -1, \exists l : \psi^{(i)}(l) = -1 \wedge\ \Delta^\# g_{y,l}(x_i) + b_i \geq \gamma \end{cases}.$$

The γ-Ψ-*dimension* of $\Delta^\# \mathcal{G}$, $\Psi\text{-dim}(\Delta^\# \mathcal{G}, \gamma)$, is the maximal cardinality of a subset of $\mathcal{X}$ γ-Ψ-shattered by $\Delta^\# \mathcal{G}$, if it is finite, or infinity otherwise.

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

$$\forall i \in \{1, \ldots, n\}\,,\ \begin{cases} \text{if } y_i = \ \ 1, \Delta^\# g_{y,i_1(x_i)}(x_i) - b_i \geq \gamma \\ \text{if } y_i = -1, \Delta^\# g_{y,i_2(x_i)}(x_i) + b_i \geq \gamma \end{cases}.$$

The *Natarajan dimension with margin* γ of the class $\Delta^\# \mathcal{G}$, $\text{N-dim}(\Delta^\# \mathcal{G}, \gamma)$, is the maximal cardinality of a subset of $\mathcal{X}$ γ-N-shattered by $\Delta^\# \mathcal{G}$, if it is finite, or infinity otherwise.

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

Theorem 2 (Theorem 4 in [Guermeur, 2004]). *Let $\mathcal{G}$ be a class of functions from a domain $\mathcal{X}$ into $\mathbb{R}^Q$. For every value of γ in $(0,1]$ and every $m \in \mathbb{N}^*$ satisfying $2m \geq$ N-dim$(\Delta_\gamma \mathcal{G}, \gamma/24)$, the following bound is true:*

$$\mathcal{N}_{\infty,\infty}^{(p)}(\gamma/4, \Delta_\gamma^* \mathcal{G}, 2m) < 2\left(288\ m\ Q^2(Q-1)\right)^{\lceil d \log_2(23emQ(Q-1)/d)\rceil} \tag{2}$$

where $d =$ N-dim$(\Delta_\gamma \mathcal{G}, \gamma/24)$.

This theorem is the central result of the paper (and the 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: two functions separated with respect to the functional pseudo-metric used (here $d_{\ell_\infty, \ell_\infty(s_{\mathcal{X}^n})}$) shatter (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$\left(\Delta_\gamma^* \mathcal{G}, \gamma/24\right)$. This alternative is however of lesser interest, for reasons that will appear below.

4 Margin Natarajan dimension of the M-SVMs

We now compute an upper bound on the margin Natarajan dimension of interest when $\mathcal{G}$ is the class of functions computed by the M-SVMs. These large margin classifiers are built around a Mercer kernel. Let κ be such a kernel on $\mathcal{X}$ and $\left(H_\kappa, \langle ., . \rangle_{H_\kappa}\right)$ the corresponding reproducing kernel Hilbert space (RKHS) [Aronszajn, 1950]. Let Φ be any of the mappings on $\mathcal{X}$ satisfying:

$$\forall\, (x, x') \in \mathcal{X}^2,\ \kappa\,(x, x') = \langle \Phi\,(x)\,, \Phi\,(x') \rangle, \tag{3}$$

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

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

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 when the sets $\{x_{ik} : 1 \leq i \leq l_k\}$ become dense in $\mathcal{X}$ in the norm induced by the dot product. Due to (3), $\mathcal{H}$ can also be seen as a multivariate affine model on $\Phi\,(\mathcal{X})$. Functions h can then be rewritten as:

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

where 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 $\|.\|_{\bar{\mathcal{H}}}$ is given by $\left\|\bar{h}\right\|_{\bar{\mathcal{H}}} = \sqrt{\sum_{k=1}^{Q} \|w_k\|^2} = \|\mathbf{w}\|$.

Definition 13 (M-SVM). A M-SVM is a large margin multi-category discriminant model obtained by minimizing over the hyperplane $\sum_{k=1}^{Q} h_k = 0$ of $\mathcal{H}$ an objective function of the form:

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

where the empirical term, used in place of the empirical risk, involves a loss function $\ell_{\text{M-SVM}}$ which is convex.

The M-SVMs 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 a notion of margin directly connected with the one of Definition 1. The formulation of the generalized Sauer lemma provided here (Theorem 2) is the one obtained under the weakest hypotheses. Proceeding as in the bi-class case, we express below a bound on the margin Natarajan dimension of the M-SVMs as a function of the volume occupied by data in $E_{\Phi(\mathcal{X})}$ and constraints on $(\mathbf{w}, \mathbf{b})$, thus restricting the study to functions with a well-defined range. In that case, a variant of Theorem 2 can be derived from Lemma 7 in [Guermeur, 2004] which does not involve π_γ but relates the covering numbers of $\Delta^*\mathcal{G}$ to the margin Natarajan dimension of $\Delta\mathcal{G}$. Its use for M-SVMs is advantageous since N-dim $\left(\Delta\bar{\mathcal{H}}, \epsilon\right)$ is easier to bound than N-dim $(\Delta_\gamma \mathcal{H}, \epsilon)$ (nonlinearity is difficult to handle). This change of generalized Sauer lemma calls for the use of an intermediate formula relating the covering numbers of $\Delta_\gamma^* \mathcal{H}$ and $\Delta^* \bar{\mathcal{H}}$. It 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 hypothesis $\mathbf{b} \in [-\beta, \beta]^Q$. Let $(\gamma, \epsilon) \in \mathbb{R}^2$ satisfy $0 < \epsilon \leq \gamma \leq 1$. Then*

$$\mathcal{N}_{\infty,\infty}^{(p)}(\epsilon, \Delta_\gamma^* \mathcal{H}, m) \leq \left(2 \left\lceil \frac{\beta}{\epsilon} \right\rceil + 1\right)^Q \mathcal{N}_{\infty,\infty}^{(p)}(\epsilon/2, \Delta^* \bar{\mathcal{H}}, m). \tag{4}$$

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 $1/2 \max_{1 \leq k < l \leq Q} \|w_k - w_l\| \leq \Lambda_w$ and $\mathbf{b} = 0$. Then, for any positive real value ϵ, the following bound holds true:*

$$N\text{-}dim\left(\Delta\bar{\mathcal{H}}, \epsilon\right) \leq C_Q^2 \left(\frac{\Lambda_w \Lambda_{\Phi(\mathcal{X})}}{\epsilon}\right)^2. \tag{5}$$

The proof follows the line of argument of the corresponding bi-class result, Theorem 4.6 in [Bartlett and Shawe-Taylor, 1999]. This involves a generalization of Lemma 4.2 which can only be performed for the Δ operator. The discussion on the presence of both Δ and Δ^* in Theorem 2 is thus completed. Putting things together, the control term of the guaranteed risk decreases with the size of the training sample as $\ln(m) \cdot m^{-1/2}$. This represents an improvement over the rate obtained in [Guermeur *et al.*, 2005], $m^{-1/4}$.

5 Conclusions and future work

A new class of generalized VC dimensions dedicated to large margin multi-category discriminant models has been introduced. 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) is also a necessary and sufficient condition for learnability. A generalized Sauer lemma has been provided for one of these capacity measures, the margin Natarajan dimension. This latter dimension has been bounded from above in the case where the classifier is a multi-class SVM. This study provides us with new arguments to support the thesis that the theory of multi-category pattern recognition cannot be developed by extending in a straightforward way bi-class results. We are currently making use of the specificities identified here to extend new concentration inequalities to the multi-class case with the goal to obtain improved convergence rates.

References

Alon *et al.*, 1997.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.

Aronszajn, 1950.N. Aronszajn. Theory of reproducing kernels. *Transactions of the American Mathematical Society*, 68(3):337–404, 1950.

Bartlett and Shawe-Taylor, 1999.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*, pages 43–54. The MIT Press, Cambridge, 1999.

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.

Ben-David *et al.*, 1995.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*, 50:74–86, 1995.

Guermeur *et al.*, 2005.Y. Guermeur, M. Maumy, and F. Sur. Model selection for multi-class SVMs. In *ASMDA'05*, pages 507–516, 2005.

Guermeur, 2004.Y. Guermeur. Large margin multi-category discriminant models and scale-sensitive Ψ-dimensions. Technical Report RR-5314, INRIA, http://hal.inria.fr/inria-00070686, 2004. (revised in 2006).

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

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