Journal of Machine Learning Research 8 (2007) 2551-2594 Submitted 1/07; Revised 8/07; Published 11/07

VC Theory of Large Margin Multi-Category Classifiers

Yann Guermeur YANN.GUERMEUR@LORIA.FR
LORIA-CNRS
Campus Scientifique, BP 239
54506 Vandœuvre-lès-Nancy cedex, France

Editors: Isabelle Guyon and Amir Saffari

Abstract

In the context of discriminant analysis, Vapnik's statistical learning theory has mainly been developed in three directions: the computation of dichotomies with binary-valued functions, the computation of dichotomies with real-valued functions, and the computation of polytomies with functions taking their values in finite sets, typically the set of categories itself. The case of classes of vector-valued functions used to compute polytomies has seldom been considered independently, which is unsatisfactory, for three main reasons. First, this case encompasses the other ones. Second, it cannot be treated appropriately through a naïve extension of the results devoted to the computation of dichotomies. Third, most of the classification problems met in practice involve multiple categories.

In this paper, a VC theory of large margin multi-category classifiers is introduced. Central in this theory are generalized VC dimensions called the γ-Ψ-dimensions. First, a uniform convergence bound on the risk of the classifiers of interest is derived. The capacity measure involved in this bound is a covering number. This covering number can be upper bounded in terms of the γ-Ψ-dimensions thanks to generalizations of Sauer's lemma, as is illustrated in the specific case of the scale-sensitive Natarajan dimension. A bound on this latter dimension is then computed for the class of functions on which multi-class SVMs are based. This makes it possible to apply the structural risk minimization inductive principle to those machines.

Keywords: multi-class discriminant analysis, large margin classifiers, uniform strong laws of large numbers, generalized VC dimensions, multi-class SVMs, structural risk minimization inductive principle, model selection

1. Introduction

One of the central domains of Vapnik's statistical learning theory (Vapnik, 1998) is the theory of bounds, which is at the origin of the structural risk minimization (SRM) inductive principle (Vapnik, 1982; Shawe-Taylor et al., 1998) and, as such, has not only a theoretical interest, but also a practical one. This theory has been developed for pattern recognition, regression estimation and density estimation. The first results in the field of discrimination, exposed in Vapnik and Chervonenkis (1971), were dealing with the computation of dichotomies with binary-valued functions. Later on, several studies were devoted to the case of multi-class $[\![1, Q]\!]$-valued classifiers (Ben-David et al., 1995), and large margin classifiers computing dichotomies (Alon et al., 1997; Bartlett, 1998; Bartlett and Shawe-Taylor, 1999) (see also Bartlett et al., 1996, for the case of regression). However, the case of large margin classifiers computing polytomies (models taking their values in $\mathbb{R}^Q$) has seldom been tackled independently, although it cannot be considered as a trivial extension of the three former ones (Guermeur et al., 1999).

In this paper, we unify two complementary and well established theories, the theory of large margin (bi-class) classifiers and the theory of multi-class $[\![1,Q]\!]$-valued classifiers, to lay the bases of a simple theory of large margin multi-class classifiers. Central in the process is the specification of a new class of generalized Vapnik-Chervonenkis (VC) dimensions, the γ-Ψ-dimensions. They can be seen either as scale-sensitive extensions of the Ψ-dimensions (Ben-David et al., 1995), or multivariate extensions of the fat-shattering dimension (Kearns and Schapire, 1994). An application to the class of functions on which multi-class SVMs (M-SVMs) are based is provided. This makes it possible to justify a posteriori the choice of their training criteria, which appear as implementations of the SRM inductive principle. This also gives birth to a model selection procedure of low computational cost. The main stages of our study are summarized in Figure 1.

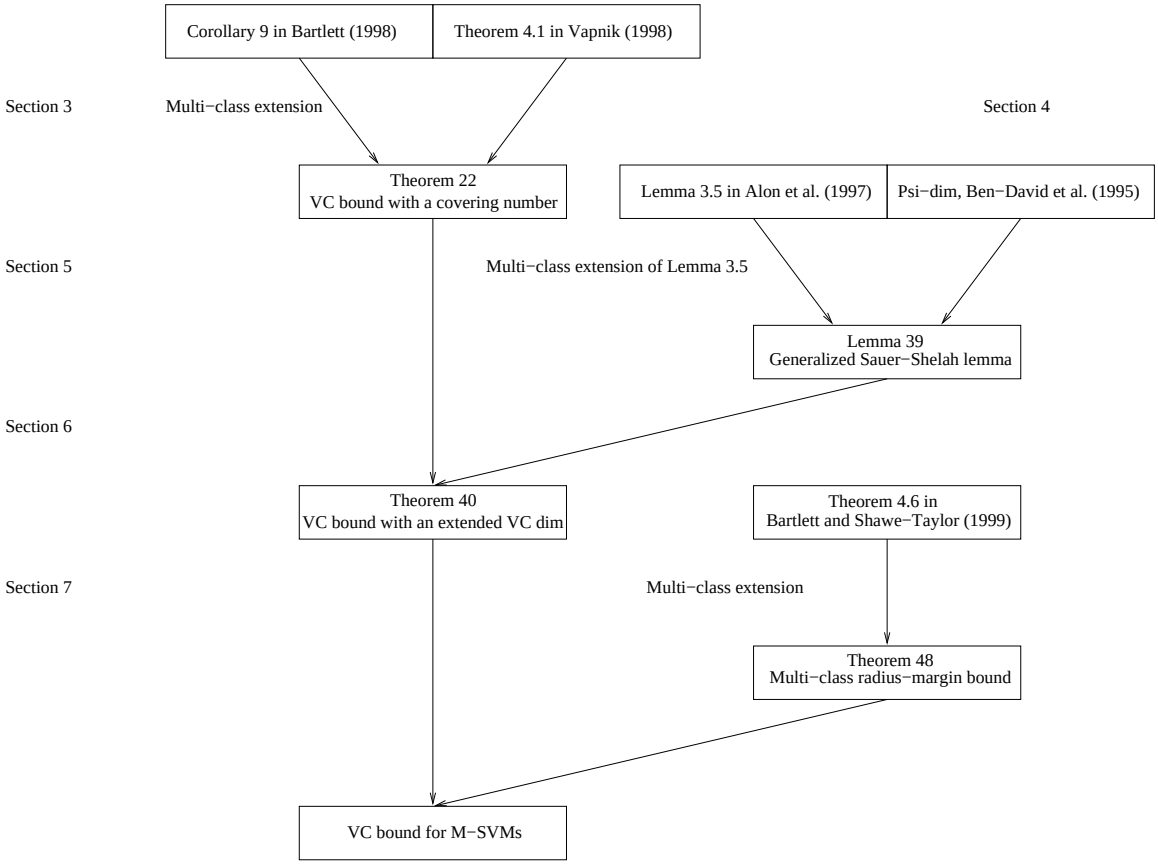

Figure 1: Organigram of the results of the paper.

Although the theorems of this theoretical contribution take the form of guaranteed risks, our aim is not to derive a tight bound on the risk, but rather to highlight the instructive features of the unification, and some specificities of the multi-class case. In that sense, our work is similar in spirit to the one exposed in Tewari and Bartlett (2007). We are all interested in the way a convergence can happen in the multi-class case. They consider the problem from the point of view of the training algorithm, whereas we focus on the capacity of the class of functions. In short, the new theory can be derived by extending concepts and results from only three famous papers: Ben-David et al. (1995), Alon et al. (1997) and Bartlett (1998), plus a fourth reference, Bartlett and Shawe-Taylor (1999), to

treat specifically the case of M-SVMs. This derivation appears rather straightforward once one has understood that different descriptors of the behaviour of the class of functions of interest are to be taken into account at the different steps of the reasoning, and this calls for the application of two different "margin operators" to this class. This phenomenon, a specificity of the multi-class case, is most noticeable at the level of the generalized Sauer-Shelah lemma, where the transition between the two operators is performed.

The organization of the paper is as follows. Section 2 introduces the notion of multi-class margin and margin risk for multi-class discriminant models, as well as the capacity measure that will appear in the confidence interval of the basic guaranteed risk, a covering number. Section 3 is then devoted to the formulation of this risk and its discussion. The γ-Ψ-dimensions are introduced in Section 4. The extension of Sauer's lemma relating the covering number of interest to one of the γ-Ψ-dimensions, the margin Natarajan dimension, is established in Section 5. Our master theorem, a combination of the basic convergence result and the aforementioned lemma, is then exposed in Section 6. Section 7 is devoted to the computation of a bound on the margin Natarajan dimension of the architecture shared by all the M-SVMs. In Section 8, the synthesis of the results derived in the preceding sections is performed, underlining the specificities of the multi-class case. This section also highlights the usefulness of our uniform convergence result for model selection. At last, we draw conclusions and outline our ongoing research in Section 9.

2. Margin Risk for Multi-Category Discriminant Models

In this section, the theoretical framework of the study is introduced. It is based on a notion of margin generalizing to an arbitrary (but finite) number of categories the standard (bi-class) one.

2.1 Formalization of the Learning Problem

We consider the case of a Q-category pattern recognition problem, with $3 \leq Q < \infty$ (so that the degenerate case of dichotomies is a priori excluded). A pattern is represented by its description $x \in \mathcal{X}$ and the set of categories $\mathcal{Y}$ is identified with the set of indexes of the categories, $[\![1,Q]\!]$. The link between patterns and categories is supposed to be of probabilistic nature. We make the assumption that $\mathcal{X}$, $\mathcal{Y}$ and the product space $\mathcal{X} \times \mathcal{Y}$ are probability spaces, and $\mathcal{X} \times \mathcal{Y}$ is endowed with a probability measure P, fixed but unknown. The measure P completely characterizes the problem of interest. In the PAC framework, this standard setting is known as *probabilistic concept learning* (Kearns and Schapire, 1994). Hereafter, $\mathcal{Z}$ will designate the product space $\mathcal{X} \times \mathcal{Y}$, and $z = (x,y)$ its elements. Our goal is to find, in a given set $\mathcal{G}$ of functions $g = (g_k)_{1 \leq k \leq Q}$ from $\mathcal{X}$ into $\mathbb{R}^Q$, a function classifying data in an optimal way. Let (X,Y) be a random pair distributed according to P. The function selection procedure, or training, makes use of a m-sample $D_m = ((X_i,Y_i))_{1 \leq i \leq m}$ of independent copies of (X,Y). It consists in trying to optimize over $\mathcal{G}$ a criterion, called the *(expected) risk*, which is the expectation with respect to P of a given *loss function*. At this point, the properties of the functions in $\mathcal{G}$ and the way they perform classification must be specified. They are supposed to satisfy some measurability conditions that will appear implicitly in the sequel (see Dudley, 1984, Chap. 10 for a detailed study of the question in a similar context), plus the constraint $\sum_{k=1}^{Q} g_k = 0$ (the purpose of this constraint will appear later). g assigns $x \in \mathcal{X}$ to the category l if and only if $g_l(x) > \max_{k \neq l} g_k(x)$. In case of ex æquo, x is assigned to a dummy category denoted by $*$. Let f be the decision function (from $\mathcal{X}$ into $\mathcal{Y} \bigcup \{*\}$) associated with g. The criterion to be

optimized is the probability of error $P(f(X) \neq Y)$. This calls for the choice of the following loss function.

Definition 1 (Multi-Class loss) *Let ℓ, the* multi-class loss function, *be defined on $\mathcal{Y} \times \mathbb{R}^Q$ by:*

$$\forall (y,v) \in \mathcal{Y} \times \mathbb{R}^Q,\ \ell(y,v) = \mathbb{1}_{\left\{v_y \leq \max_{k \neq y} v_k\right\}}$$

where $\mathbb{1}$ is the indicator function, which takes the value 1 if its argument is true, and 0 otherwise.

ℓ is simply the 0-1 loss in the multi-class setting. The expected risk of a function g is consequently defined as follows.

Definition 2 (Expected risk) *The* expected risk *of a function $g \in \mathcal{G}$, $R(g)$, is given by:*

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

The empirical risk is simply the estimate of the risk computed on the training sample.

Definition 3 (Empirical risk) *The empirical risk of $g \in \mathcal{G}$ measured on a m-sample, $R_m(g)$, is the random variable given by:*

$$R_m(g) = \frac{1}{m} \sum_{i=1}^{m} \mathbb{1}_{\left\{g_{Y_i}(X_i) \leq \max_{k \neq Y_i} g_k(X_i)\right\}}.$$

When needed, the m-sample used will be specified, by writing for instance $R_{D_m}(g)$ in place of $R_m(g)$. Let $n \in \mathbb{N}^* = \mathbb{N} \setminus \{0\}$ and let $z^n = ((x_i, y_i))_{1 \leq i \leq n} \in \mathcal{Z}^n$. In the sequel, $R_{z^n}(g)$ will designate the frequency of errors $\frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\left\{g_{y_i}(x_i) \leq \max_{k \neq y_i} g_k(x_i)\right\}}$.

2.2 Multi-Class Margin and Multi-Class Margin Risk

For the classes of vector-valued functions we are interested in, the two elements which are the most important 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. This calls for the use of a measure different from the standard indicator function ℓ to assess the quality of a discrimination. This measure can be built around a notion of multi-class margin which has been studied independently by different groups of authors (see for instance Elisseeff et al., 1999; Allwein et al., 2000). To define it, we first define an auxiliary function.

Definition 4 (Function M) *Let M be the function from $\mathbb{R}^Q \times [\![1, Q]\!]$ to $\mathbb{R}$ defined as:*

$$\forall (v,k) \in \mathbb{R}^Q \times [\![1, Q]\!],\ M(v,k) = \frac{1}{2}\left(v_k - \max_{l \neq k} v_l\right).$$

Let $M(v,.) = \max_{1 \leq k \leq Q} M(v,k)$.

Definition 5 (Multi-Class margin) *Let g be a function of a class $\mathcal{G}$. Its* margin *on $(x,y) \in \mathcal{X} \times \mathcal{Y}$ is defined to be $M(g(x), y)$.*

To take this margin into account, the following operators are introduced:

Definition 6 (Δ operator) *Define Δ as an operator on $\mathcal{G}$ such that:*

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

$$\forall x \in \mathcal{X},\ \Delta g(x) = \left(M\left(g(x),k\right)\right)_{1\leq k\leq Q}.$$

For the sake of simplicity, we write Δg_k in place of $(\Delta g)_k$. In the sequel, similar simplifications will be performed implicitly with other operators.

Definition 7 (Δ^* operator) *Define Δ^* as an operator on $\mathcal{G}$ such that:*

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

$$\forall x \in \mathcal{X},\ \Delta^* g(x) = \left(\max\left(\Delta g_k(x), -M\left(g(x),.\right)\right)\right)_{1\leq k\leq Q}.$$

Remark 8 *If $M(g(x),.) > 0$, $\Delta g(x)$ has a unique (strictly) positive component, otherwise it has none. Let us consider the first case, and let $k^* = \operatorname{argmax}_{1\leq k\leq Q} \Delta g_k(x) = \operatorname{argmax}_{1\leq k\leq Q} g_k(x)$ ($\Delta g_{k^*}(x) = M(g(x),.)$).*

$$\forall x \in \mathcal{X},\ \begin{cases} \text{if } M(g(x),.) > 0, & \Delta^* g(x) = \left((2\delta_{k,k^*} - 1)\,\Delta g_{k^*}(x)\right)_{1\leq k\leq Q} \\ \text{if } M(g(x),.) = 0, & \Delta^* g(x) = 0 \end{cases}.$$

where δ is the Kronecker symbol.

Example 1 *Suppose that $g(x) = (-0.1, 0.6, -0.3, -0.2)$. Then*

$$\begin{cases} \Delta g(x) = (-0.35, 0.35, -0.45, -0.4) \\ \Delta^* g(x) = (-0.35, 0.35, -0.35, -0.35) \end{cases}.$$

Before proceeding, it is useful to highlight the way those definitions relate to the bi-class case. Let $\tilde{\mathcal{G}}$ denote the class of real-valued functions implemented by a large margin bi-class classifier. There is a one-to-one map from this class onto a class $\mathcal{G}$ as defined above. To each function $\tilde{g}$ in $\tilde{\mathcal{G}}$, a function $g = (g_1, g_2)$ in $\mathcal{G}$ can be associated such that $g_1 = \tilde{g} = -g_2$. This is precisely to ensure the existence of this one-to-one map that the constraint $\sum_{k=1}^{Q} g_k = 0$ has been introduced. Then, $\Delta g = \Delta^* g = g = (\tilde{g}, -\tilde{g})$. As a consequence, one can consider that when implementing a large margin bi-class classifier, the functions effectively handled are the component functions Δg_1 (or equivalently the component functions $\Delta^* g_1$). 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 these formulas is the one connecting the risk of g with the behaviour of $\Delta^\# g$.

Proposition 9 *The risk of a function g of $\mathcal{G}$ can be expressed as:*

$$R(g) = \mathbb{E}\left[\mathbb{1}_{\{\Delta^\# g_Y(X)\leq 0\}}\right].$$

With these definitions at hand, the margin risk is defined as follows.

Definition 10 (Margin risk) *Let $\gamma \in \mathbb{R}_+^* = (0,\infty)$. The* risk with margin γ *of a function g of $\mathcal{G}$, $R_\gamma(g)$, is defined as:*

$$R_\gamma(g) = \mathbb{E}\left[\mathbb{1}_{\{\Delta^\# g_\gamma(X)<\gamma\}}\right].$$

The empirical risk with margin γ of g, $R_{\gamma,m}(g)$ (or $R_{\gamma,D_m}(g)$ if the sample needs to be specified), and the frequency of errors with margin γ, $R_{\gamma,z^n}(g)$, are defined accordingly.

A consequence of the definition of the margin risk is the fact that knowing the exact behaviour of the component functions $\Delta^\# g_k$ below $-\gamma$ and over γ is useless. On the contrary, one can take benefit from working with classes of functions taking values in $[-\gamma,\gamma]^Q$, which is compact, rather than in $\mathbb{R}^Q$. This advantage will appear in the first place in Section 3, and then more clearly in Section 5. Such a transform is achieved by application of the following piecewise-linear squashing operator.

Definition 11 (π_γ operator, Bartlett, 1998) *For $\gamma \in \mathbb{R}_+^*$, define π_γ as an operator on $\mathcal{G}$ such that:*

$$\begin{aligned}\pi_\gamma : \quad & \mathcal{G} \longrightarrow \pi_\gamma\mathcal{G}, \\ & g \mapsto \pi_\gamma g = \left(\pi_\gamma g_k\right)_{1\le k\le Q},\end{aligned}$$

$$\forall x \in \mathcal{X},\ \pi_\gamma g(x) = \left(sign\left(g_k(x)\right)\cdot\min\left(|g_k(x)|,\gamma\right)\right)_{1\le k\le Q}$$

where the sign function is defined by $sign(t) = 1$ if $t \ge 0$, and $sign(t) = -1$ otherwise.

For $\gamma \in \mathbb{R}_+^*$, let $\Delta_\gamma^\#$ denote $\pi_\gamma \circ \Delta^\#$ and $\Delta_\gamma^\#\mathcal{G} = \left\{\Delta_\gamma^\# g : g \in \mathcal{G}\right\}$.

The capacity measure that will appear in the basic guaranteed risk stated in Section 3 is a covering number. Its definition, and the definition of related concepts, is the subject of the following section. Introductions to the basic notions of functional analysis used in this article can be found in Carl and Stephani (1990), Devroye et al. (1996) and van der Vaart and Wellner (1996).

2.3 Capacity Measures: Covering and Packing Numbers

The notion of covering number is based on the notions of ε-cover and ε-net.

Definition 12 (ε-cover and ε-net, Kolmogorov and Tihomirov, 1961) *Let (E,ρ) be a pseudo-metric space. For $e \in E$ and $r \in \mathbb{R}_+^*$, let $B(e,r)$ be the open ball of center e and radius r in E. Let E' be a subset of E. For $\varepsilon \in \mathbb{R}_+^*$, an* ε-net *of E' is a subset $\overline{E'}$ of E such that:*

$$E' \subset \bigcup_{e\in\overline{E'}} B(e,\varepsilon).$$

$\bigcup_{e\in\overline{E'}} B(e,\varepsilon)$ is then an ε-cover *of E'. $\overline{E'}$ is a* proper ε-net *of E' if it is included in E'.*

Definition 13 (Covering number, Kolmogorov and Tihomirov, 1961) *Let (E,ρ) be a pseudo-metric space. For $\varepsilon \in \mathbb{R}_+^*$, if $E' \subset E$ has an ε-net of finite cardinality, then its* covering number *$\mathcal{N}(\varepsilon,E',\rho)$ is the smallest cardinality of its ε-nets. If there is no such finite net, then the covering number is defined to be ∞. We denote $\mathcal{N}^{(p)}(\varepsilon,E',\rho)$ the covering number obtained by considering proper ε-nets only.*

Hereafter, the pseudo-metric that will be used on the families of functions considered is the following one:

Definition 14 (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:*

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

Definition 15 $\forall n \in \mathbb{N}^*, \forall \varepsilon \in \mathbb{R}_+^*$,

$$\mathcal{N}^{(p)}(\varepsilon, \mathcal{G}, n) = \max_{x^n \in \mathcal{X}^n} \mathcal{N}^{(p)}(\varepsilon, \mathcal{G}, d_{x^n}),$$

the maximum being used in place of a supremum to highlight the fact that we implicitly make the assumption that all the ε-nets considered are of finite cardinality.

There is a close connection between covering and packing properties of bounded subsets in pseudo-metric spaces.

Definition 16 (ε-separation and packing number, Kolmogorov and Tihomirov, 1961) *Let (E,ρ) be a pseudo-metric space and $\varepsilon \in \mathbb{R}_+^*$. A set $E' \subset E$ is* ε-separated *if, for any distinct points e_1 and e_2 in E', $\rho(e_1,e_2) \geq \varepsilon$. The* ε-packing number *of $E'' \subset E$, $\mathcal{M}(\varepsilon, E'', \rho)$, is the maximal size of an ε-separated subset of E''.*

Definition 17 (Separation) *For $n \in \mathbb{N}^*$, let $\mathcal{F}$ be a class of functions on $\mathcal{X}$ taking their values in $[\![-n,n]\!]^Q$ and $\mathcal{F}|_{\mathcal{D}}$ its restriction to a subset $\mathcal{D}$ of $\mathcal{X}$ of finite cardinality. Two functions f and f' in the class $\mathcal{F}|_{\mathcal{D}}$ are* separated *if they are 2-separated in the pseudo-metric $d_{\mathcal{D}}$, that is, if*

$$\max_{x\in\mathcal{D}} \left\| f(x) - f'(x) \right\|_\infty \geq 2.$$

Definition 18 (Pairwise separated set of functions) *Let $\mathcal{F}$, $\mathcal{D}$ and $\mathcal{F}|_{\mathcal{D}}$ be defined as above. $\mathcal{F}|_{\mathcal{D}}$ is* pairwise separated *if any two distinct functions of $\mathcal{F}|_{\mathcal{D}}$ are separated.*

2.4 Additional Definitions

This section gathers definitions which will be used in the proof of our basic uniform convergence result.

Definition 19 ($\overline{\mathcal{G}}(\gamma,x^n)$ and $\overline{\mathcal{G}}(\gamma,D_n)$) *Let $n \in \mathbb{N}^*$ and $\gamma \in \mathbb{R}_+^*$. Let $x^n = (x_i)_{1\leq i\leq n} \in \mathcal{X}^n$. Let us consider any function (deterministic algorithm) f_{net} that takes as input γ, $\mathcal{G}$, and x^n, and returns a subset of $\mathcal{G}$ such that its image by the operator $\Delta_\gamma^\#$ is a proper $\gamma/2$-net of the set $\Delta_\gamma^\#\mathcal{G}$ (in the pseudo-metric d_{x^n}), and this net is of minimal cardinality, that is, of cardinality $\mathcal{N}^{(p)}\left(\gamma/2, \Delta_\gamma^\#\mathcal{G}, d_{x^n}\right)$.*

$$\overline{\mathcal{G}}(\gamma,x^n) = f_{net}(\gamma, \mathcal{G}, x^n).$$

The random variable $\overline{\mathcal{G}}(\gamma,D_n)$ is defined accordingly, by replacing in the definition of $\overline{\mathcal{G}}(\gamma,x^n)$ the sequence x^n with $(X_i)_{1\leq i\leq n}$.

Note that for the sake of simplicity, we use $\overline{\mathcal{G}}(\gamma, D_n)$ in place of $\overline{\mathcal{G}}\left(\gamma, (X_i)_{1\leq i\leq n}\right)$, although the latter formulation is more precise.

Definition 20 (Swapping group $\mathfrak{T}_{2n}$) *For $n \in \mathbb{N}^*$, let $\mathfrak{T}_{2n}$ be the "swapping" subgroup of $\mathfrak{S}_{2n}$, the symmetric group of degree $2n$. $\mathfrak{T}_{2n}$ is the set of all permutations σ over $[\![1,2n]\!]$ that swap i and $n+i$ for all i in some subset of $[\![1,n]\!]$. Precisely, for all i in $[\![1,n]\!]$, $(\sigma(i),\sigma(i+n))$ is either equal to $(i,i+n)$ or to $(i+n,i)$. The permutations σ are regarded as acting on coordinates. For $z^{2n} \in \mathcal{Z}^{2n}$ and $\sigma \in \mathfrak{T}_{2n}$, let $\sigma\left(z^{2n}\right) = \left(\left(x_{\sigma(i)}, y_{\sigma(i)}\right)\right)_{1\leq i\leq 2n}$. $\mathfrak{T}_{2n}$ is endowed with a uniform probability distribution.*

Definition 21 (Bernoulli/Rademacher sequence) *For $n \in \mathbb{N}^*$, a* Bernoulli or Rademacher sequence *is a sequence $\alpha = (\alpha_i)_{1\leq i\leq n}$ of independent real random variables with $\mathbb{P}(\alpha_i = -1) = \mathbb{P}(\alpha_i = 1) = \frac{1}{2}$ for all i.*

3. Uniform Convergence of the Empirical Margin Risk

With the hypotheses and definitions of the previous section at hand, we prove the following uniform convergence result.

3.1 Basic Uniform Convergence Result

Theorem 22 *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:*

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

The proof is given in Appendix B. This theorem can be seen as a multi-class extension of Corollary 9 in Bartlett (1998). Indeed, setting $Q = 2$ (and $\Gamma = 1$), we get a slightly improved version of this corollary. The difference rests on the fact that in the first symmetrization, we took advantage of an idea which is implicitly at the basis of Formula (4.28) in Vapnik (1998). This idea consists in making use of Lemma 49. As a consequence, Theorem 22 can also be seen as a specification for the case of large margin multi-category classifiers of Theorem 4.1 in Vapnik (1998).

3.2 Choice of the Margin Operator

Theorem 22 has been derived for both margin operators, Δ and Δ^*. The choice between them should thus rest on the use which is done of the bound, that is, on the nature of the pathway followed to bound from above the covering number of interest. This question, the nature of which is primarily technical, will turn out to be of central importance in the following sections. At this point, we can already notice that the Δ^* operator provides less information on the behaviour of the function on which it is applied than the Δ operator. Such a difference would appear as an advantage to derive a generalization of Sauer's lemma, and a drawback to compute an upper bound on the corresponding generalized VC dimension. This suggests to implement a hybrid strategy, mixing results involving Δ^* with results involving Δ. This is precisely what will be done here.

4. γ-Ψ-dimensions: the Generalized VC Dimensions of Large Margin Multi-Category Classifiers

Several approaches can be applied to bound from above the covering number of interest for a given class of functions $\mathcal{G}$. The standard one, introduced in Vapnik and Chervonenkis (1971), consists in involving in the process the VC dimension, or one of its extensions. If VC dimensions appear useful in practice, their interest is primarily of theoretical nature. Indeed, they characterize learnability in different settings (see for instance Alon et al., 1997). In this section, the γ-Ψ-dimensions are introduced as the generalized VC dimensions suited for large margin multi-category classifiers. They appear as syntheses of the Ψ-dimensions and the fat-shattering dimension (also known as the γ-dimension). The pertinence of this specification will be established in Section 5.

The basic result relating a covering number (precisely the growth function) to the VC dimension is the Sauer-Shelah lemma (Vapnik and Chervonenkis, 1971; Sauer, 1972; Shelah, 1972). As stated in the introduction, extensions of the standard VC theory, which only deals with the computation of dichotomies with indicator functions, have mainly been proposed for large margin bi-class discriminant models and multi-class discriminant models taking their values in finite sets. In both cases, generalized Sauer-Shelah lemmas have been derived (see for instance Haussler and Long, 1995; Alon et al., 1997), which involve extended notions of VC dimension. For large margin bi-class discriminant models, the generalization of the VC dimension which has given birth to the richest set of theoretical results is a scale-sensitive variant called the fat-shattering dimension (Kearns and Schapire, 1994). In the multi-class case, several alternative solutions were proposed by different authors, such as the graph dimension (Dudley, 1987; Natarajan, 1989), or the Natarajan dimension (Natarajan, 1989). It was proved in Ben-David et al. (1995) that most of these extensions could be gathered in a general scheme, which makes it possible to derive necessary and sufficient conditions for PAC learning (Valiant, 1984). In this scheme, they appear as special cases of Ψ-dimensions.

We introduce scale-sensitive extensions of the Ψ-dimensions. The underlying idea is simple: in the same way as scale-sensitive extensions of the VC dimension, such as the fat-shattering dimension, make it possible to study the generalization capabilities of bi-class discriminant models taking their values in $\mathbb{R}$, scale-sensitive extensions of the Ψ-dimensions should make it possible to study the generalization capabilities of Q-class discriminant models taking their values in $\mathbb{R}^Q$.

4.1 Ψ-dimensions

Definition 23 (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.*

Remark 24 *Let $\mathcal{F}$ and Ψ be defined as above. Extending the definition of the standard VC dimension, VC-dim, so that it applies to classes of functions taking values in $\{-1,1,*\}$, which has no incidence in practice, the following proposition holds true:*

$$\Psi\text{-}dim(\mathcal{F}) = VC\text{-}dim\left(\{(x,\psi) \mapsto \psi \circ f(x) : f \in \mathcal{F}, \psi \in \Psi\}\right).$$

In words, the idea common to all these dimensions is to introduce adequately chosen mappings from $[\![1,Q]\!]$ into $\{-1,1,*\}$ so that the problem of the computation of the capacity measure boils down to the computation of several standard VC dimensions. In that context, the motivation for the choice of one particular dimension (set Ψ) utterly rests on the possibility to derive two tight bounds: a generalized Sauer-Shelah lemma and a bound on the dimension itself. The most frequently used Ψ-dimension is the graph dimension.

Definition 25 (Graph dimension, Natarajan, 1989) *Let $\mathcal{F}$ be a class of functions on a set $\mathcal{X}$ taking their values in $[\![1,Q]\!]$. The* graph dimension *of $\mathcal{F}$, G-dim$(\mathcal{F})$, is the Ψ-dimension of $\mathcal{F}$ in the specific case where $\Psi = \{\psi_k : 1 \leq k \leq Q\}$, such that ψ_k takes the value 1 if its argument is equal to k, and the value -1 otherwise. Reformulated in the context of multi-class discriminant analysis, the functions ψ_k are the indicator functions of the categories.*

Obviously, this notion of Ψ-dimension is connected with one of the standard decomposition schemes implemented to tackle multi-class problems with bi-class classifiers: the *one-against-all* method. Another popular decomposition scheme is the *one-against-one* method. The corresponding Ψ-dimension has been proposed by Natarajan.

Definition 26 (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.*

4.2 Margin Ψ-dimensions

Our scale-sensitive version of the concept of Ψ-dimension is devised so that the corresponding dimensions can alternatively be seen as multivariate extensions of the fat-shattering dimension.

Definition 27 (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*

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

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.*

Let $\wedge$ denote the conjunction of two events. With these definitions at hand, the Ψ-dimensions with margin γ, or γ-Ψ-dimensions, are defined as follows:

Definition 28 (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 *(*Ψ-shattered with margin γ*) 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*

$$\forall i \in [\![1,n]\!], \begin{cases} \text{if } y_i = \quad 1, & \exists k : \psi^{(i)}(k) = \quad 1 \quad \wedge \ \Delta^\# g_{y,k}(x_i) - b_i \geq \gamma \\ \text{if } y_i = -1, & \exists l : \psi^{(i)}(l) = -1 \quad \wedge \ \Delta^\# g_{y,l}(x_i) + b_i \geq \gamma \end{cases} . \tag{2}$$

The γ-Ψ-dimension*, or* Ψ-dimension with margin γ*, of $\Delta^\# \mathcal{G}$, denoted by Ψ-dim$\left(\Delta^\# \mathcal{G}, \gamma\right)$, 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.*

From a theoretical point of view, the one-against-one decomposition method exhibits an advantage over the one-against-all decomposition method: its use makes it easier to extend to the multi-class case bi-class theorems, by application of the pigeonhole principle. Thus, the scale-sensitive Ψ-dimension which will be involved in our generalized Sauer-Shelah lemma is the one extending the Natarajan dimension. Given the definitions of the Natarajan dimension and the scale-sensitive Ψ-dimensions, it can be formulated as:

Definition 29 (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 *(*N-shattered with margin γ*) 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 couples of distinct indexes 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

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

The Natarajan dimension with margin γ *of the class $\Delta^\# \mathcal{G}$, N-dim $\left(\Delta^\# \mathcal{G}, \gamma\right)$, 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.*

4.3 Discussion

In the preceding section, we have given a formulation of the definition of the Natarajan dimension which is inspired from the one in Ben-David et al. (1995) (the definition in Natarajan, 1989, does not involve the $\psi_{k,l}$ mappings). This formulation can be restricted by considering only the mappings $\psi_{k,l}$ such that $k < l$, instead of $k \neq l$. This is possible due to the symmetrical roles played by the indexes of categories $i_1(x_i)$ and $i_2(x_i)$ in the definition. As a consequence, the cardinality of the set Ψ considered can be divided by 2 (reduced from $Q(Q-1)$ to $\binom{Q}{2}$). This is useful indeed, since many theorems dealing with Ψ-dimensions involve the cardinality of Ψ (see for instance Theorem 7 in Ben-David et al., 1995). An equivalent simplification can be performed in the case of the margin Natarajan dimension.

Proposition 30 *The definition of the Natarajan dimension with margin γ is not affected by the introduction of the additional constraint: $\forall i \in [\![1,n]\!]$, $i_1(x_i) < i_2(x_i)$.*

Proof Let $\mathcal{G}_y$ be a subset of $\mathcal{G}$ of cardinality 2^n such that $\Delta^{\#}\mathcal{G}_y$ γ-N-shatters $s_{\mathcal{X}^n}$ with respect to $I(s_{\mathcal{X}^n})$ and v_b. Let $I'(s_{\mathcal{X}^n})$ be the set of n couples of indexes $(i'_1(x_i), i'_2(x_i))$ deduced from $I(s_{\mathcal{X}^n})$ by reordering its elements, that is,

$$\forall i \in [\![1,n]\!],\ \left(i'_1(x_i), i'_2(x_i)\right) = \left(\min\left(i_1(x_i), i_2(x_i)\right), \max\left(i_1(x_i), i_2(x_i)\right)\right).$$

Let $v_{b'} = (b'_i)$ be the vector of $\mathbb{R}^n$ deduced from v_b as follows: $\forall i \in [\![1,n]\!]$, $b'_i = b_i$ if $(i'_1(x_i), i'_2(x_i)) = (i_1(x_i), i_2(x_i))$, $b'_i = -b_i$ otherwise. We establish that $\Delta^{\#}\mathcal{G}_y$ still γ-N-shatters $s_{\mathcal{X}^n}$ with respect to $I'(s_{\mathcal{X}^n})$ and $v_{b'}$. For any vector $v_y = (y_i)$ of $\{-1,1\}^n$, let $g_{y'}$ be the function in $\mathcal{G}_y$ such that $\Delta^{\#}g_{y'}$ "contributes" to the γ-N-shattering of $s_{\mathcal{X}^n}$ with respect to $I(s_{\mathcal{X}^n})$ and v_b for a value of the binary vector equal to $v_{y'} = (y'_i)$, where $y'_i = y_i$ if $(i'_1(x_i), i'_2(x_i)) = (i_1(x_i), i_2(x_i))$, $y'_i = -y_i$ otherwise. According to Definition 29,

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

As a consequence, for the set of indexes i such that $(i'_1(x_i), i'_2(x_i)) = (i_1(x_i), i_2(x_i))$,

$$\left\{ \begin{array}{lll} \text{if } y_i = & 1, & \Delta^{\#}g_{y',i'_1(x_i)}(x_i) - b'_i \geq \gamma \\ \text{if } y_i = & -1, & \Delta^{\#}g_{y',i'_2(x_i)}(x_i) + b'_i \geq \gamma \end{array} \right. . \tag{3}$$

Furthermore, for the set of indexes i such that $(i'_1(x_i), i'_2(x_i)) = (i_2(x_i), i_1(x_i))$,

$$\left\{ \begin{array}{lll} \text{if } y_i = & -1, & \Delta^{\#}g_{y',i'_2(x_i)}(x_i) + b'_i \geq \gamma \\ \text{if } y_i = & 1, & \Delta^{\#}g_{y',i'_1(x_i)}(x_i) - b'_i \geq \gamma \end{array} \right. .$$

This is exactly (3), which thus holds true for all values of i in $[\![1,n]\!]$ (whether the couple $(i'_1(x_i), i'_2(x_i))$ is equal to $(i_1(x_i), i_2(x_i))$ or equal to $(i_2(x_i), i_1(x_i))$). According to Definition 29, the function $\Delta^{\#}g_{y'}$ thus contributes to the γ-N-shattering of $s_{\mathcal{X}^n}$ with respect to $I'(s_{\mathcal{X}^n})$ and $v_{b'}$ for a value of the binary vector equal to v_y. But since the vector v_y has been chosen arbitrarily in $\{-1,1\}^n$, this implies that $\Delta^{\#}\mathcal{G}_y$ γ-N-shatters $s_{\mathcal{X}^n}$ with respect to $I'(s_{\mathcal{X}^n})$ and $v_{b'}$, which, by construction of $I'(s_{\mathcal{X}^n})$, concludes the proof. ■

In the sequel, we will sometimes make use of Proposition 30 implicitly. We now establish that the γ-Ψ-dimensions are actually multivariate extensions of the fat-shattering dimension.

Proposition 31 *Let $\tilde{\mathcal{G}}$ be a class of real-valued functions on a set $\mathcal{X}$. Let $\mathcal{G}$ be the corresponding class of functions from $\mathcal{X}$ into $\mathbb{R}^2$. Then, for all positive value of γ,*

$$P_\gamma\text{-}dim(\tilde{\mathcal{G}}) = \Psi\text{-}dim(\Delta^{\#}\mathcal{G}, \gamma).$$

Proof When $Q = 2$, one can consider that the set Ψ contains only two mappings, ψ_+ and ψ_-, with $\psi_+(1) = 1$, $\psi_+(2) = -1$ and $\psi_-(1) = -1$, $\psi_-(2) = 1$ (adding other mappings, for instance mappings taking the value $*$, would be useless since such mappings either do not take the value 1, or do not take the value -1). Using the same line of reasoning as in the proof of Proposition 30, one

can establish that Ψ can be restricted further to the singleton $\{\psi_+\}$. As a consequence, (2) simplifies into:

$$\forall i \in [\![1,n]\!], \left\{ \begin{array}{ll} \text{if } y_i = \quad 1, & \Delta^{\#} g_{y,1}(x_i) - b_i \geq \gamma \\ \text{if } y_i = -1, & \Delta^{\#} g_{y,2}(x_i) + b_i \geq \gamma \end{array} \right. . \tag{4}$$

Since we have seen in Section 2.2 that $\Delta^{\#} g = (\tilde{g}, -\tilde{g})$, (4) simplifies further into (1) (with $\tilde{g}$ in place of g), which concludes the proof. ■

In both cases (fat-shattering dimension and margin Ψ-dimensions) the introduction of the vector of "biases" v_b could be seen as a simple computational trick, useful to derive the generalized Sauer-Shelah lemma (establish a connection between the property of separation and the capacity to shatter a set of points) at the expense of a more complex computation for the bound on the margin dimension itself. This is partly the case indeed. However, in Section 7, we will see that these extra degrees of freedom can be handled pretty easily.

5. Relating the Covering Number and the Margin Natarajan Dimension

This section is devoted to the formulation of an upper bound on the covering number of interest in terms of the margin Natarajan dimension. Its main result is a generalization of the Sauer-Shelah lemma given by Lemmas 38 and 39. Our basic uniform convergence result, Theorem 22, involves the class of functions $\Delta^{\#}_{\gamma}\mathcal{G}$. However, in the preceding section, the scale-sensitive Ψ-dimensions have been defined for $\Delta^{\#}\mathcal{G}$ (although the extension to $\Delta^{\#}_{\gamma}\mathcal{G}$ is straightforward). The reason for this change, and the way it can be handled, is the subject of the following subsection.

5.1 Switching from $\Delta^{\#}_{\gamma}\mathcal{G}$ to $\Delta^{\#}\mathcal{G}$

As stated in Section 2.2, the advantage of working with the class $\Delta^{\#}_{\gamma}\mathcal{G}$ is obvious: the range of its functions, $[-\gamma,\gamma]^Q$, is optimal (the smallest range that does not affect the value of the margin risk). The seamy side of things is that the nonlinearity introduced by the π_γ operator is difficult to handle when bounding a generalized VC dimension. Furthermore, there is no direct connection between $\Psi\text{-dim}\left(\Delta^{\#}_{\gamma}\mathcal{G},\varepsilon\right)$ and $\Psi\text{-dim}\left(\Delta^{\#}\mathcal{G},\varepsilon\right)$. On the contrary, the transition can be performed very easily at the level of the covering number, thanks to the following lemma.

Lemma 32 *Let $\mathcal{G}$ be a class of functions from a domain $\mathcal{X}$ into $\mathbb{R}^Q$, let γ and ε be two positive real numbers and let $n \in \mathbb{N}^*$. Then,*

$$\mathcal{N}^{(p)}(\varepsilon,\Delta^{\#}_{\gamma}\mathcal{G},n) \leq \mathcal{N}^{(p)}(\varepsilon,\Delta^{\#}\mathcal{G},n).$$

Proof This property directly springs from the fact that π_γ satisfies the Lipschitz condition with constant 1. Thus, $\forall (g,g') \in \mathcal{G}^2$, $\forall x \in \mathcal{X}$, $\forall (\gamma,\varepsilon) \in \left(\mathbb{R}^*_+\right)^2$,

$$\left\|\Delta^{\#} g(x) - \Delta^{\#} g'(x)\right\|_\infty < \varepsilon \Longrightarrow \left\|\Delta^{\#}_{\gamma} g(x) - \Delta^{\#}_{\gamma} g'(x)\right\|_\infty < \varepsilon.$$

■

Since the computations leading to our generalized Sauer-Shelah lemma will require the functions in $\Delta^{\#}\mathcal{G}$ to have a bounded range, to compensate for the elimination of the π_γ operator, from now on, we make the hypothesis that there exists a positive real number M such that the functions g, and by

way of consequence the functions $\Delta^{\#}g$, take their values in $[-M,M]^Q$. Given this hypothesis, the only values of the margin parameter γ corresponding to a nontrivial situation are those inferior or equal to M. As a consequence, we also assume that the parameter Γ of Theorem 22 is set equal to M. To formulate the main combinatorial result of this section, new concepts are to be defined. They correspond to extensions of concepts introduced in Alon et al. (1997).

5.2 Definitions

Definition 33 (η-discretization operator) *Let $\mathcal{G}$ be a class of functions from $\mathcal{X}$ into $[-M,M]^Q$. For $\eta \in \mathbb{R}_+^*$, define the* η-discretization *as an operator on $\Delta^{\#}\mathcal{G}$ such that:*

$$\begin{aligned}(.)^{(\eta)}: \quad & \Delta^{\#}\mathcal{G} \longrightarrow \left(\Delta^{\#}\mathcal{G}\right)^{(\eta)}, \\ & \Delta^{\#}g \mapsto \left(\Delta^{\#}g\right)^{(\eta)} = \left(\left(\Delta^{\#}g_k\right)^{(\eta)}\right)_{1\leq k\leq Q},\end{aligned}$$

$$\forall x \in \mathcal{X},\ \left(\Delta^{\#}g\right)^{(\eta)}(x) = \left(sign\left(\Delta^{\#}g_k(x)\right)\cdot\left\lfloor\frac{\left|\Delta^{\#}g_k(x)\right|}{\eta}\right\rfloor\right)_{1\leq k\leq Q}$$

where the function $\lfloor.\rfloor$ is defined by $\forall t \in \mathbb{R}_+$, $\lfloor t\rfloor = \max\{j\in\mathbb{N}:\ j\leq t\}$.

Note that this definition is not a straightforward extension of the original one to the case of vector-valued functions, since we had to relax the hypothesis of nonnegativity. Fortunately, this generalization does not raise any difficulty.

Definition 34 (Strong Natarajan dimension) *Let $\mathcal{G}$ be a class of functions from $\mathcal{X}$ into $[-M,M]^Q$ and let $\eta \in (0,M]$. A subset $s_{\mathcal{X}^n} = \{x_i : 1\leq i\leq n\}$ of $\mathcal{X}$ is said to be* strongly N-shattered *by $\left(\Delta^{\#}\mathcal{G}\right)^{(\eta)}$ 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 couples of distinct indexes in $[\![1,Q]\!]$ and a vector $v_b = (b_i)$ in $\left[\!\left[-\left\lfloor\frac{M}{\eta}\right\rfloor+1, \left\lfloor\frac{M}{\eta}\right\rfloor-1\right]\!\right]^n$ such that, for each vector $v_y = (y_i)$ in $\{-1,1\}^n$, there is a function g_y in $\mathcal{G}$ satisfying

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

The strong Natarajan dimension *of the class $\left(\Delta^{\#}\mathcal{G}\right)^{(\eta)}$, SN-dim$\left(\left(\Delta^{\#}\mathcal{G}\right)^{(\eta)}\right)$, is the maximal cardinality of a subset of $\mathcal{X}$ strongly N-shattered by $\left(\Delta^{\#}\mathcal{G}\right)^{(\eta)}$, if this cardinality is finite. If no such maximum exists, $\left(\Delta^{\#}\mathcal{G}\right)^{(\eta)}$ is said to have infinite strong Natarajan dimension.*

Obviously, as in the case of the margin Natarajan dimension, the definition remains unchanged if the additional constraint: $\forall i\in[\![1,n]\!]$, $i_1(x_i) < i_2(x_i)$ is introduced.

5.3 Relating Separation and Strong N-shattering

The Sauer-Shelah lemma and its generalizations rest on a simple idea: to establish a connection between the property of separation of two functions and their capacity to "shatter" a singleton. This connection is obvious in the case of binary-valued functions (more precisely functions taking values in $\{-1,1\}$). Then,

$$\left|f(x)-f'(x)\right| \geq 2 \Longleftrightarrow f(x) = -f'(x)$$

and thus f and f' classify x in different categories. Things are more complicated is the case of classifiers taking values in $\mathbb{R}^Q$. The corresponding result in that latter context is the following.

Lemma 35 *Let $\mathcal{G}$ be a class of functions from $\mathcal{X}$ into $[-M,M]^Q$ and let η be a real number belonging to $(0,M]$. Let $\mathcal{D}$ be a subset of $\mathcal{X}$ of finite cardinality and let $\mathcal{F}$ and $\mathcal{F}^*$ be respectively the restrictions of $(\Delta\mathcal{G})^{(\eta)}$ and $(\Delta^*\mathcal{G})^{(\eta)}$ to $\mathcal{D}$. $\mathcal{F}$ and $\mathcal{F}^*$ are endowed with the pseudo-metric $d_{\mathcal{D}}$. If two functions g and g' in $\mathcal{G}$ are such that $f^* = (\Delta^* g)^{(\eta)}\Big|_{\mathcal{D}}$ and $f^{*'} = (\Delta^* g')^{(\eta)}\Big|_{\mathcal{D}}$ are separated, then there exists x in $\mathcal{D}$ such that $\left\{ f = (\Delta g)^{(\eta)}\Big|_{\mathcal{D}}, f' = (\Delta g')^{(\eta)}\Big|_{\mathcal{D}} \right\}$ strongly N-shatters the singleton $\{x\}$. Suppose further, without loss of generality, that $\max_k f_k^*(x) \geq \max_k f_k^{*'}(x)$ and let $k_0 = \operatorname{argmax}_k f_k^*(x)$. Then there is at least one couple $(I(\{x\}), v_b) = (\{(i_1(x), i_2(x))\}, (b_0))$ with $i_1(x) = k_0$ and $b_0 = f_{k_0}^*(x) - 1$ witnessing the strong N-shattering of $\{x\}$ by $\{f, f'\}$.*

Proof We first demonstrate that k_0 is well defined. Indeed, this is the case unless $f^*(x) = 0$. But $f^*(x) = 0$ and $\max_k f_k^*(x) \geq \max_k f_k^{*'}(x)$ implies that $f^{*'}(x) = 0$, which is in contradiction with the hypothesis $\left\| f^*(x) - f^{*'}(x) \right\|_\infty \geq 2$. By definition of the operator Δ^*, there exists an index l_0 different from k_0 such that $f'_{l_0}(x) = f_{l_0}^{*'}(x)$. l_0 is simply the index of a component of $g'(x)$ satisfying $g'_{l_0}(x) = \max_{k \neq k_0} g'_k(x)$. By definition of k_0 and b_0, $f_{k_0}(x) - b_0 = f_{k_0}^*(x) - b_0 = 1$. By construction, $f'_{l_0}(x) + b_0 = f_{l_0}^{*'}(x) + b_0$. Two cases must now be considered. If $f_{l_0}^{*'}(x) = \max_k f_k^{*'}(x)$, then $f_{l_0}^{*'}(x) = 0 \Longrightarrow f^{*'}(x) = 0 \Longrightarrow f_{k_0}^*(x) \geq 2$ (otherwise f^* and $f^{*'}$ would not be separated). As a consequence, $f_{l_0}^{*'}(x) + f_{k_0}^*(x) \geq 2$ and thus $f_{l_0}^{*'}(x) + b_0 \geq 1$, which is equivalent to $f'_{l_0}(x) + b_0 \geq 1$. If $f_{l_0}^{*'}(x) \neq \max_k f_k^{*'}(x)$, then $f_{k_0}^{*'}(x) = \max_k f_k^{*'}(x)$ and $f_{l_0}^{*'}(x) = -f_{k_0}^{*'}(x)$. Necessarily, $f_{k_0}^*(x) - f_{k_0}^{*'}(x) \geq 2$ (otherwise f^* and $f^{*'}$ would not be separated) and finally $f'_{l_0}(x) + b_0 = f_{k_0}^*(x) - f_{k_0}^{*'}(x) - 1 \geq 1$. Thus, the couple $\left(I(\{x\}) = \{(k_0, l_0)\}, v_b = \left(f_{k_0}^*(x) - 1\right)\right)$ witnesses the strong N-shattering of $\{x\}$ by $\{f, f'\}$. ■

Lemma 35 will turn out to be of central importance in the sequel. We consider it as contributing to characterize the specificity of the multi-class case, since it highlights the usefulness of the Δ^* operator (whereas the usefulness of the Δ operator will appear in Section 7).

Remark 36 *Lemma 35 cannot be stated with the operator Δ only.*

Proof To prove this last assertion, it suffices to exhibit a counter example. Let $\mathcal{G}$ be a class of functions from $\mathcal{X}$ into $[-2,2]^4$ and g and g' be two functions in $\mathcal{G}$ such that there exists $\mathcal{D} = \{x\}$ satisfying $g(x) = (1.4, -0.2, -0.2, -1.0)$ and $g'(x) = (1.4, -0.2, -0.6, -0.6)$. Let $\eta = 0.1$. Using the same notations as above, we get $f(x) = (8, -8, -8, -12)$, $f'(x) = (8, -8, -10, -10)$ and $f^*(x) = f^{*'}(x) = (8, -8, -8, -8)$. Although f and f' are separated, they do not strongly N-shatter

$\{x\}$. Indeed, if it were the case, then according to Definition 34, there would be two different indexes k_0 and l_0 in $[\![1,4]\!]$ such that $f_{k_0}(x)+f'_{l_0}(x)\geq 2$, which is not the case. ■

In contrast with this negative result, the hypothesis $\left\| f^*(x)-f^{*'}(x)\right\|_\infty \geq 2$ also implies that $\left\{ f^*, f^{*'}\right\}$ strongly N-shatters $\{x\}$ (Lemma 35 could have been stated with the operator Δ^* only). A tricky thing must be borne in mind. If two pairs $\left(g^{(1)},g^{(2)}\right)$ and $\left(g^{(3)},g^{(4)}\right)$ of functions in $\mathcal{G}$ are such that $\left(f^{*(1)}(x),f^{*(2)}(x)\right)=\left(f^{*(3)}(x),f^{*(4)}(x)\right)$, then if $\left\|f^{*(1)}(x)-f^{*(2)}(x)\right\|_\infty \geq 2$, $\{x\}$ is strongly N-shattered both by $\left\{f^{(1)},f^{(2)}\right\}$ and by $\left\{f^{(3)},f^{(4)}\right\}$. However, those shatterings could require different witnesses $(I(\{x\}),v_b)$. More precisely, using the notations of Definition 34, given the couple $\left(f^{*(1)},f^{*(2)}\right)$, one can exhibit an index $i_1(x)$ and a bias b_0 contributing to both shatterings (by $\left\{f^{(1)},f^{(2)}\right\}$ and by $\left\{f^{(3)},f^{(4)}\right\}$) but the last component of the witness, $i_2(x)$, must be chosen as a function of the values taken by the functions f on x. It is thus a priori different for $\left\{f^{(1)},f^{(2)}\right\}$ and for $\left\{f^{(3)},f^{(4)}\right\}$.

We now prove the main combinatorial result at the basis of our generalization of the Sauer-Shelah lemma, an extension of Lemma 3.3 in Alon et al. (1997).

5.4 Main Combinatorial Result

Lemma 37 *Let $\mathcal{G}$ be a class of functions on $\mathcal{X}$ taking their values in $[-M,M]^Q$ and let η be a real number belonging to $(0,M]$. Let $\mathcal{D}$ be a subset of $\mathcal{X}$ of finite cardinality $|\mathcal{D}|$ and let $\mathcal{F}$ and $\mathcal{F}^*$ be respectively the restrictions of $(\Delta\mathcal{G})^{(\eta)}$ and $(\Delta^*\mathcal{G})^{(\eta)}$ to $\mathcal{D}$. $\mathcal{F}$ and $\mathcal{F}^*$ are endowed with the pseudo-metric $d_{\mathcal{D}}$. Setting $d=SN\text{-}dim(\mathcal{F})$ and $q=\left\lfloor \frac{M}{\eta}\right\rfloor$, the following bound holds true:*

$$\mathcal{M}(2,\mathcal{F}^*,d_{\mathcal{D}}) < 2\left(|\mathcal{D}|\ Q^2(Q-1)\,q^2\right)^{\lceil \log_2(\phi(d,|\mathcal{D}|))\rceil} \tag{5}$$

where $\phi(d,|\mathcal{D}|)=\sum_{i=1}^{d}\binom{|\mathcal{D}|}{i}\left(\binom{Q}{2}(2q-1)\right)^i$.

Proof Let us say that the class $\mathcal{F}$ strongly N-shatters a triplet $(s_{\mathcal{D}},I(s_{\mathcal{D}}),v_b)$ (for a nonempty subset $s_{\mathcal{D}}$ of $\mathcal{D}$, a set of couples of indexes $I(s_{\mathcal{D}})$ and a vector of biases v_b) if $\mathcal{F}$ strongly N-shatters $s_{\mathcal{D}}$ according to $I(s_{\mathcal{D}})$ and v_b. For all integers $l\geq 2$ and $|\mathcal{D}|\geq 1$, let $t(l,|\mathcal{D}|)$ denote the maximum number t such that, for every set $\mathcal{F}_l^*$ of l pairwise separated functions in $\mathcal{F}^*$, $\mathcal{F}_l=\left\{f\in\mathcal{F}: f^*\in\mathcal{F}_l^*\right\}$ strongly N-shatters at least t triplets $(s_{\mathcal{D}},I(s_{\mathcal{D}}),v_b)$. If there is no subset of $\mathcal{F}^*$ of cardinality l pairwise separated, then $t(l,|\mathcal{D}|)$ is infinite.

The number of triplets $(s_{\mathcal{D}},I(s_{\mathcal{D}}),v_b)$ that could be shattered and for which the cardinality of $s_{\mathcal{D}}$ does not exceed $d\geq 1$ is less than $\sum_{i=1}^{d}\binom{|\mathcal{D}|}{i}\left(\binom{Q}{2}(2q-1)\right)^i$, since for $s_{\mathcal{D}}$ of size $i>0$, there are strictly less than $\left(\binom{Q}{2}(2q-1)\right)^i$ possibilities to choose the couple $(I(s_{\mathcal{D}}),v_b)$. It follows that $t(l,|\mathcal{D}|)\geq\phi(d,|\mathcal{D}|)$ for some l implies $t(l,|\mathcal{D}|)=\infty$. By definition of $t(l,|\mathcal{D}|)$, this means that there is no subset of $\mathcal{F}^*$ of cardinality l pairwise separated (otherwise $t(l,|\mathcal{D}|)$ would be finite) and finally, by definition of $\mathcal{M}(2,\mathcal{F}^*,d_{\mathcal{D}})$, $\mathcal{M}(2,\mathcal{F}^*,d_{\mathcal{D}})<l$. Therefore, to finish the proof, it suffices to show that, for all $d\geq 1$ and $|\mathcal{D}|\geq 1$,

$$t\left(2\left(|\mathcal{D}|\ Q^2(Q-1)\,q^2\right)^{\lceil \log_2(\phi(d,|\mathcal{D}|))\rceil},|\mathcal{D}|\right)\geq\phi(d,|\mathcal{D}|). \tag{6}$$

We claim that

$$t(2,|\mathcal{D}|)\geq 1 \tag{7}$$

for all $|\mathcal{D}| \geq 1$ and

$$t\left(2p\ |\mathcal{D}|\ Q^2(Q-1)\,q^2, |\mathcal{D}|\right) \geq 2t(2p, |\mathcal{D}|-1) \tag{8}$$

for all $p \geq 1$ and $|\mathcal{D}| \geq 2$.

The first part of the claim is a direct consequence of Lemma 35.

For the second part, first note that if no set of $2p\ |\mathcal{D}|\ Q^2(Q-1)\,q^2$ pairwise separated functions in $\mathcal{F}^*$ exists, then by definition $t\left(2p\ |\mathcal{D}|\ Q^2(Q-1)\,q^2, |\mathcal{D}|\right) = \infty$ and hence the claim holds. Assume then that there is a set $\mathcal{F}_0^*$ of $2p\ |\mathcal{D}|\ Q^2(Q-1)\,q^2$ pairwise separated functions in $\mathcal{F}^*$. Split it arbitrarily into $p\ |\mathcal{D}|\ Q^2(Q-1)\,q^2$ pairs. For each pair $\left(f^*, f^{*'}\right)$, there exists a singleton $\{x\} \subset \mathcal{D}$ strongly N-shattered by $\{f, f'\}$. Once more, this is a direct consequence of Lemma 35. By definition, a vector $f^*(x)$ has all components of equal magnitude. As a consequence, the number of different values that it can take is equal to $Qq+1$. The numbers of different sets of the form $\left\{f^*(x), f^{*'}(x)\right\}$ such that $\left\|f^*(x) - f^{*'}(x)\right\|_\infty \geq 2$ is bounded from above by $\frac{1}{2}(Qq+1)(Qq-1) < \frac{1}{2}Q^2\,q^2$. Thus, by the pigeonhole principle, switching the indexes in the couples of functions if needed, for each procedure of this type, there exists $x_0 \in \mathcal{D}$ such that at least $\left(2p\ |\mathcal{D}|\ Q^2(Q-1)\,q^2\right) / \left(|\mathcal{D}|\ Q^2\,q^2\right) = 2p\,(Q-1)$ of the resulting couples of functions take the same value on x_0, value satisfying $\left\|f^*(x_0) - f^{*'}(x_0)\right\|_\infty \geq 2$. For all these pairs, the corresponding sets $\{f, f'\}$ all shatter $\{x_0\}$ (shatter at least one triplet of the form $(\{x_0\}, I(\{x_0\}), v_b)$). If the components of the couples are reordered in such a way that all the couples are identical with $\max_k f_k^*(x_0) \geq \max_k f_k^{*'}(x_0)$, this result still holds if one imposes that the values of $i_1(x_0)$ and b_0 are those considered in Lemma 35 ($i_1(x_0) = \text{argmax}_k\, f_k^*(x_0)$ and $b_0 = f^*_{i_1(x_0)}(x_0) - 1$). Once $i_1(x_0)$ is set, $i_2(x_0)$ can take at most $Q-1$ different values. Thus, using once more the pigeonhole principle, among those last couples of functions, there are (at least) $2p\,(Q-1)/(Q-1) = 2p$ of them such that the quintuplet $\left(x_0, f^*(x_0), f^{*'}(x_0), I(\{x_0\}), v_b\right)$ can be the same, that is, a single pair $(I(\{x_0\}), v_b)$ can witness the strong N-shattering of $\{x_0\}$ by all the sets $\{f, f'\}$. To sum up, this means that there are two subclasses of $\mathcal{F}_0^*$ of cardinality at least $2p$, call them $\mathcal{F}_+^*$ and $\mathcal{F}_-^*$, and there are $x_0 \in \mathcal{D}$, two vectors $V_{0,+}$ and $V_{0,-}$ in $[\![-q, q]\!]^Q$ such that $\|V_{0,+} - V_{0,-}\|_\infty \geq 2$, $(k_0, l_0) \in [\![1, Q]\!]^2$ with $k_0 \neq l_0$, and a scalar b_0 in $[\![-q+1, q-1]\!]$ such that:

$$\begin{cases} \forall f_+^* \in \mathcal{F}_+^*, & f_+^*(x_0) & = & V_{0,+} \\ \forall f_-^* \in \mathcal{F}_-^*, & f_-^*(x_0) & = & V_{0,-} \\ \forall f_+ \in \mathcal{F}_+, & f_{+,k_0}(x_0) & \geq & 1+b_0 \\ \forall f_- \in \mathcal{F}_-, & f_{-,l_0}(x_0) & \geq & 1-b_0 \end{cases}$$

where $\mathcal{F}_+ = \left\{f_+ \in \mathcal{F} : f_+^* \in \mathcal{F}_+^*\right\}$ and $\mathcal{F}_- = \left\{f_- \in \mathcal{F} : f_-^* \in \mathcal{F}_-^*\right\}$. Since the members of $\mathcal{F}_+^*$ are pairwise separated on $\mathcal{D}$ but are all equal on x_0, they are pairwise separated on $\mathcal{D} \setminus \{x_0\}$. The same holds for the members of $\mathcal{F}_-^*$. Hence, by definition of the function t, $\mathcal{F}_+$ strongly N-shatters at least $t\,(2p, |\mathcal{D}|-1)$ triplets $(s_\mathcal{D}, I(s_\mathcal{D}), v_b)$ with $s_\mathcal{D} \subseteq \mathcal{D} \setminus \{x_0\}$, and the same holds for $\mathcal{F}_-$. Clearly, $\mathcal{F}_0 = \{f \in \mathcal{F} : f^* \in \mathcal{F}_0^*\}$ strongly N-shatters all triplets strongly N-shattered either by $\mathcal{F}_+$ or by $\mathcal{F}_-$. Moreover, if the same triplet $(s_\mathcal{D}, I(s_\mathcal{D}), v_b)$ is strongly N-shattered both by $\mathcal{F}_+$ and by $\mathcal{F}_-$, then $\mathcal{F}_0$ also strongly N-shatters the triplet $(\{x_0\} \bigcup s_\mathcal{D}, \{(k_0, l_0)\} \bigcup I(s_\mathcal{D}), \overline{v_b})$, where $\overline{v_b}$ is deduced from v_b by adding one component corresponding to the point x_0, component taking the value b_0. Indeed, the sets $\mathcal{F}_+$ and $\mathcal{F}_-$ have been built precisely in that purpose. Suffice it to notice what follows. Let $(s_\mathcal{D}, I(s_\mathcal{D}), v_b)$ be a triplet strongly N-shattered both by $\mathcal{F}_+$ and by $\mathcal{F}_-$. For the sake

of simplicity, reordering the points in $\mathcal{D}$ if needed, we suppose that $s_{\mathcal{D}}$ can be written as follows: $s_{\mathcal{D}} = \{x_i : 1 \leq i \leq |s_{\mathcal{D}}|\}$. Then, for any vector $v_y = (y_i)$ in $\{-1,1\}^{|s_{\mathcal{D}}|}$, there exists (at least) one function $f_{+,y}$ in $\mathcal{F}_+$ such that

$$\forall i \in [\![1, |s_{\mathcal{D}}|]\!], \left\{ \begin{array}{ll} \text{if } y_i = \quad 1, & f_{+,y,i_1(x_i)}(x_i) - b_i \geq 1 \\ \text{if } y_i = -1, & f_{+,y,i_2(x_i)}(x_i) + b_i \geq 1 \end{array} \right.$$

and

$$f_{+,y,k_0}(x_0) - b_0 \geq 1$$

and one function $f_{-,y}$ in $\mathcal{F}_-$ such that

$$\forall i \in [\![1, |s_{\mathcal{D}}|]\!], \left\{ \begin{array}{ll} \text{if } y_i = \quad 1, & f_{-,y,i_1(x_i)}(x_i) - b_i \geq 1 \\ \text{if } y_i = -1, & f_{-,y,i_2(x_i)}(x_i) + b_i \geq 1 \end{array} \right.$$

and

$$f_{-,y,l_0}(x_0) + b_0 \geq 1.$$

Since, once more by construction, neither $\mathcal{F}_+$ nor $\mathcal{F}_-$ strongly N-shatters $\{x_0\} \bigcup s_{\mathcal{D}}$ (whatever the pair $(I(\{x_0\} \bigcup s_{\mathcal{D}}), \bar{v}_b)$ may be), it follows that $t\left(2p\ |\mathcal{D}|\ Q^2(Q-1)\, q^2, |\mathcal{D}|\right) \geq 2t(2p, |\mathcal{D}|-1)$, which is precisely (8).

For any integer number r satisfying $1 \leq r < |\mathcal{D}|$, let

$$l = 2\left(Q^2(Q-1)\, q^2\right)^r\ \Pi_{u=0}^{r-1}(|\mathcal{D}| - u).$$

Applying (8) iteratively and eventually (7), it appears that $t(l, |\mathcal{D}|) \geq 2^r$. Since t is clearly nondecreasing in its first argument, and $2\left(|\mathcal{D}|\ Q^2(Q-1)\, q^2\right)^r \geq l$, this implies

$$t\left(2\left(|\mathcal{D}|\ Q^2(Q-1)\, q^2\right)^r, |\mathcal{D}|\right) \geq 2^r.$$

We make use of this bound by considering separately the case where $\lceil \log_2(\phi(d, |\mathcal{D}|)) \rceil < |\mathcal{D}|$ and the case where $\lceil \log_2(\phi(d, |\mathcal{D}|)) \rceil \geq |\mathcal{D}|$. In the first case, one can set $r = \lceil \log_2(\phi(d, |\mathcal{D}|)) \rceil$. We then get

$$t\left(2\left(|\mathcal{D}|\ Q^2(Q-1)\, q^2\right)^{\lceil \log_2(\phi(d,|\mathcal{D}|)) \rceil}, |\mathcal{D}|\right) \geq 2^{\lceil \log_2(\phi(d,|\mathcal{D}|)) \rceil}$$

and consequently

$$t\left(2\left(|\mathcal{D}|\ Q^2(Q-1)\, q^2\right)^{\lceil \log_2(\phi(d,|\mathcal{D}|)) \rceil}, |\mathcal{D}|\right) \geq 2^{\log_2(\phi(d,|\mathcal{D}|))} = \phi(d, |\mathcal{D}|)$$

which is precisely (6). If on the contrary $\lceil \log_2(\phi(d, |\mathcal{D}|)) \rceil \geq |\mathcal{D}|$, then

$$2\left(|\mathcal{D}|\ Q^2(Q-1)\, q^2\right)^{\lceil \log_2(\phi(d,|\mathcal{D}|)) \rceil} > (Qq+1)^{|\mathcal{D}|}.$$

Since the number of distinct functions in $\mathcal{F}^*$ is bounded from above by $(Qq+1)^{|\mathcal{D}|}$, $\mathcal{F}^*$ cannot contain a set of pairwise separated functions of cardinality larger than this number and hence, by definition of t,

$$t\left(2\left(|\mathcal{D}|\ Q^2(Q-1)\, q^2\right)^{\lceil \log_2(\phi(d,|\mathcal{D}|)) \rceil}, |\mathcal{D}|\right) = \infty.$$

$t\left(2\left(|\mathcal{D}|\ Q^2(Q-1)\, q^2\right)^{\lceil \log_2(\phi(d,|\mathcal{D}|)) \rceil}, |\mathcal{D}|\right)$ is consequently once more superior to $\phi(d, |\mathcal{D}|)$, which completes the proof of (6) and thus concludes the proof of the lemma. ■

Note that expressing Lemma 37 in the bi-class case (by setting $Q = 2$), one obtains almost exactly the expression of Lemma 3.3 in Alon et al. (1997), keeping in mind that our functions and theirs do not take their values in the same intervals.

5.5 Generalized Sauer-Shelah Lemma

Our generalized Sauer-Shelah lemma appears as a direct consequence of Lemma 37.

Lemma 38 (Generalized Sauer-Shelah lemma) *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 $n \geq N\text{-dim}(\Delta\mathcal{G},\varepsilon/6)$, the following bound is true:*

$$\mathcal{N}^{(p)}(\varepsilon,\Delta^*\mathcal{G},n) < 2\left(n\,Q^2(Q-1)\left\lfloor\frac{3M}{\varepsilon}\right\rfloor^2\right)^{\lceil\log_2(\phi(d,n))\rceil} \tag{9}$$

where $d = N\text{-dim}(\Delta\mathcal{G},\varepsilon/6)$ and $\phi(d,n) = \sum_{i=1}^{d}\binom{n}{i}\left(\binom{Q}{2}\left(2\left\lfloor\frac{3M}{\varepsilon}\right\rfloor-1\right)\right)^i$.

Proof $\forall x^n \in \mathcal{X}^n$, applying Lemma 56 (right-hand side inequality) to $\Delta^*\mathcal{G}$ gives:

$$\mathcal{N}^{(p)}\left(\varepsilon,\Delta^*\mathcal{G},d_{x^n}\right) \leq \mathcal{M}\left(\varepsilon,\Delta^*\mathcal{G},d_{x^n}\right).$$

Setting $\eta = \varepsilon/3$ in Proposition 2 of Lemma 57, one obtains:

$$\mathcal{N}^{(p)}\left(\varepsilon,\Delta^*\mathcal{G},d_{x^n}\right) \leq \mathcal{M}\left(2,\left(\Delta^*\mathcal{G}\right)^{(\varepsilon/3)},d_{x^n}\right). \tag{10}$$

Let $\mathcal{D}_n$ denote the smallest subset of $\mathcal{X}$ including all the elements of x^n (its cardinality is inferior or equal to n since x^n can contain multiple copies of some elements of $\mathcal{X}$). We write $\left(\Delta^*\mathcal{G}\right)^{(\varepsilon/3)}\Big|_{\mathcal{D}_n}$ to designate the restriction of $\left(\Delta^*\mathcal{G}\right)^{(\varepsilon/3)}$ to $\mathcal{D}_n$. Since

$$\mathcal{M}\left(2,\left(\Delta^*\mathcal{G}\right)^{(\varepsilon/3)},d_{x^n}\right) = \mathcal{M}\left(2,\left(\Delta^*\mathcal{G}\right)^{(\varepsilon/3)}\Big|_{\mathcal{D}_n},d_{x^n}\right),$$

(10) implies:

$$\mathcal{N}^{(p)}\left(\varepsilon,\Delta^*\mathcal{G},d_{x^n}\right) \leq \mathcal{M}\left(2,\left(\Delta^*\mathcal{G}\right)^{(\varepsilon/3)}\Big|_{\mathcal{D}_n},d_{x^n}\right).$$

The packing numbers of $\left(\Delta^*\mathcal{G}\right)^{(\varepsilon/3)}\Big|_{\mathcal{D}_n}$ can be bounded thanks to Lemma 37, by setting $\mathcal{D} = \mathcal{D}_n$, using n as an upper bound on $|\mathcal{D}|$ (which is possible since the right-hand side of (5) is an increasing function of $|\mathcal{D}|$), $q = \left\lfloor\frac{M}{\eta}\right\rfloor = \left\lfloor\frac{3M}{\varepsilon}\right\rfloor$ and $d = \text{SN-dim}\left(\left(\Delta\mathcal{G}\right)^{(\varepsilon/3)}\Big|_{\mathcal{D}_n}\right)$. Thus, we get:

$$\mathcal{N}^{(p)}\left(\varepsilon,\Delta^*\mathcal{G},d_{x^n}\right) < 2\left(n\,Q^2(Q-1)\left\lfloor\frac{3M}{\varepsilon}\right\rfloor^2\right)^{\lceil\log_2(\phi(d,n))\rceil}, \tag{11}$$

with $\phi(d,n) = \sum_{i=1}^{d}\binom{n}{i}\left(\binom{Q}{2}\left(2\left\lfloor\frac{3M}{\varepsilon}\right\rfloor-1\right)\right)^i$. Since the right-hand side of (11) is a nondecreasing function of d, one can replace d with an upper bound. By definition of $\left(\Delta\mathcal{G}\right)^{(\varepsilon/3)}\Big|_{\mathcal{D}_n}$,

$$\text{SN-dim}\left(\left(\Delta\mathcal{G}\right)^{(\varepsilon/3)}\Big|_{\mathcal{D}_n}\right) \leq \text{SN-dim}\left(\left(\Delta\mathcal{G}\right)^{(\varepsilon/3)}\right).$$

By application of Proposition 1 in Lemma 57,

$$\text{SN-dim}\left((\Delta\mathcal{G})^{(\varepsilon/3)}\right) \leq \text{N-dim}\left(\Delta\mathcal{G},\varepsilon/6\right).$$

Thus, (11) still holds if d is set equal to $\text{N-dim}\left(\Delta\mathcal{G},\varepsilon/6\right)$. Taking the maximum of its left-hand side over $\mathcal{X}^n$ then concludes the proof. ■

To find an upper bound on $\phi(d,n)$, and thus derive a generalized Sauer-Shelah lemma easier to handle than Lemma 38, it suffices to make use of Lemma 58 with $K_1 = d$, $K_2 = n$ and $K_3 = \binom{Q}{2}\left(2\left\lfloor\frac{3M}{\varepsilon}\right\rfloor - 1\right)$. This implies that

$$\phi(d,n) < \Phi\left(d,n,\binom{Q}{2}\left(2\left\lfloor\frac{3M}{\varepsilon}\right\rfloor - 1\right)\right) < \left(\frac{en\binom{Q}{2}\left(2\left\lfloor\frac{3M}{\varepsilon}\right\rfloor - 1\right)}{d}\right)^d$$

and consequently

$$\log_2\left(\phi(d,n)\right) < d\log_2\left(\frac{en\binom{Q}{2}\left(2\left\lfloor\frac{3M}{\varepsilon}\right\rfloor - 1\right)}{d}\right).$$

Substituting the right-hand side of this inequality to its left-hand side in (9), we finally get our master lemma.

Lemma 39 (Final formulation of the generalized Sauer-Shelah lemma) *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$\left(\Delta\mathcal{G},\varepsilon/6\right)$, the following bound is true:*

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

where $d =$ N-dim$\left(\Delta\mathcal{G},\varepsilon/6\right)$.

5.6 Discussion

To sum up, in this section, we have derived a bound on the covering number of interest in terms of one of the γ-Ψ-dimensions, the margin Natarajan dimension. Obviously, such a generalized Sauer-Shelah lemma can be derived in a similar way for other scale-sensitive extensions of a Ψ-dimension, such as the one corresponding to the graph dimension. The bound, by the way, is slightly easier to establish in the latter case. It involves smaller constants. However, as was already pointed out in Section 4.1, the choice of one particular variant of the VC dimension rests on the search for an optimal compromise between two requirements that can be contradictory: the need for a tight bound on the capacity measure in terms of the VC dimension, and the need for a tight bound on the VC dimension itself. In Section 7, it will appear clearly that the connection of the Natarajan dimension with the one-against-one decomposition method is a major advantage. Deriving a bound on the margin Natarajan dimension of the M-SVMs can be performed very simply, by extending in a straightforward way the reasoning of the proof of the standard bound on the fat-shattering dimension of the perceptron (or pattern recognition SVM).

6. Almost Sure Convergence Result

The combination of Theorem 22 and Lemma 39 (applied with $\varepsilon = \gamma/4$ and $n = 2m$) provides us with our master theorem.

Theorem 40 *Let $\mathcal{G}$ be the class of functions from $\mathcal{X}$ into $[-M,M]^Q$ that a large margin Q-category classifier can implement. Let $\delta \in (0,1)$. With probability at least $1-\delta$, for every value of γ in $(0,M]$, the risk of any function g in $\mathcal{G}$ is bounded from above by:*

$$R(g) \leq R_{\gamma,m}(g) +$$

$$\sqrt{\frac{2}{m}\left(\ln\left(4\left(2m\,Q^2(Q-1)\left\lfloor\frac{12M}{\gamma}\right\rfloor^2\right)^{\left\lceil d\log_2\left(emQ(Q-1)\left(2\left\lfloor\frac{12M}{\gamma}\right\rfloor-1\right)/d\right)\right\rceil}\right)+\ln\left(\frac{2M}{\gamma\delta}\right)\right)+\frac{1}{m}}$$

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

With our notation, which designates by P and $\mathbb{P}_{D_m}$ respectively the probability measure characterizing the classification problem of interest, and a probability over the m-sample D_m, Theorem 40 states a distribution-free bound corresponding to a one-sided convergence in probability of the form:

$$\lim_{m\longrightarrow+\infty}\sup_P \mathbb{P}_{D_m}\left(\sup_{g\in\mathcal{G}}\left(R(g)-R_{\gamma,D_m}(g)\right)>\varepsilon\right)=0.$$

In fact, a stronger result can be obtained, since the convergence holds with probability 1.

Proposition 41 (Almost sure convergence)

$$\lim_{m\longrightarrow+\infty}\sup_P \mathbb{P}\left(\sup_{n\geq m}\sup_{g\in\mathcal{G}}\left(R(g)-R_{\gamma,n}(g)\right)>\varepsilon\right)=0.$$

Proof For a class $\mathcal{G}$ of functions taking values in $\mathbb{R}^Q$ and a given value of γ in $\mathbb{R}_+^*$, we obtained the following bound as a partial result in the proof of Theorem 22:

$$\mathbb{P}_{D_m}\left(\sup_{g\in\mathcal{G}}\left(R(g)-R_{\gamma,D_m}(g)\right)>\varepsilon\right)\leq 2\mathcal{N}^{(p)}\left(\gamma/2,\Delta_\gamma^{\#}\mathcal{G},2m\right)\exp\left(-\frac{m}{2}\left(\varepsilon-\frac{1}{m}\right)^2\right).$$

Under the restrictive assumption that the functions in $\mathcal{G}$ take their values in $[-M,M]^Q$, Lemmas 32 and 39 can be applied to bound from above the covering number, which yields:

$$\mathbb{P}_{D_m}\left(\sup_{g\in\mathcal{G}}\left(R(g)-R_{\gamma,D_m}(g)\right)>\varepsilon\right)\leq$$

$$4\left(2m\,Q^2(Q-1)\left\lfloor\frac{6M}{\gamma}\right\rfloor^2\right)^{\left\lceil d\log_2\left(emQ(Q-1)\left(2\left\lfloor\frac{6M}{\gamma}\right\rfloor-1\right)/d\right)\right\rceil}\exp\left(-\frac{m}{2}\left(\varepsilon-\frac{1}{m}\right)^2\right) \qquad (12)$$

where $d = \text{N-dim}\,(\Delta\mathcal{G}, \gamma/12)$. Let us denote by u_m the right-hand side of (12). Obviously,

$$\forall \varepsilon > 0,\ 4\left(2m\,Q^2(Q-1)\left\lfloor\frac{6M}{\gamma}\right\rfloor^2\right)^{\left\lceil d\log_2\left(emQ(Q-1)\left(2\left\lfloor\frac{6M}{\gamma}\right\rfloor-1\right)/d\right)\right\rceil} = o\left(\exp\left(\frac{m\varepsilon^2}{4}\right)\right).$$

As a consequence, $u_m = o\left(\exp\left(-\frac{m\varepsilon^2}{4}\right)\right)$. Since $\sum_{m=1}^{\infty}\exp\left(-\frac{m\varepsilon^2}{4}\right) < \infty$, by transitivity,

$$\sum_{m=1}^{\infty}\mathbb{P}_{D_m}\left(\sup_{g\in\mathcal{G}}\left(R(g) - R_{\gamma,D_m}(g)\right) > \varepsilon\right) < \infty.$$

One may thus apply the Borel-Cantelli lemma (see for instance Theorem A.22. in Devroye et al., 1996) and strengthen to almost sure convergence the convergence stated in Theorem 40. ■

7. Margin Natarajan Dimension of the Multi-Class SVMs

The theoretical results derived so far were dealing with general classes of functions $\mathcal{G}$, from $\mathcal{X}$ into $\mathbb{R}^Q$ or $[-M,M]^Q$, satisfying the mild conditions exposed in Section 2.1. In short, our aim was to establish that for those classes, the γ-Ψ-dimensions characterize learnability in the same way as the VC dimension, the fat-shattering dimension and the Ψ-dimensions characterize learnability for classes of functions taking values respectively in $\{-1,1\}$, $\mathbb{R}$ and $[\![1,Q]\!]$. From now on, we assess the use of the γ-Ψ-dimensions to characterize and control the generalization capabilities of classes of parametric functions. To that end, we focus on the main models of large margin multi-category classifiers, the multi-class SVMs.

Support vector machines (SVMs) are learning systems which have been introduced by Vapnik and co-workers (Boser et al., 1992; Cortes and Vapnik, 1995) as nonlinear extensions of the maximal margin hyperplane (Vapnik, 1982). Originally, they were designed to perform pattern recognition (compute dichotomies). In this context, the principle on which they are based is very simple. First, the examples are mapped into a high-dimensional Hilbert space called the *feature space* thanks to a nonlinear transform, the *feature map*, usually denoted by Φ. Second, the maximal margin hyperplane is computed in that space, to separate the two categories. The problem of performing multi-class discriminant analysis with SVMs was initially tackled through decomposition schemes involving bi-class machines. Such possibilities as the one-against-all method (Rifkin and Klautau, 2004), the one-against-one method (Fürnkranz, 2002) (a variant of which is the DAGSVM of Platt et al., 2000), or those based on error correcting codes (ECOC) (Allwein et al., 2000; Crammer and Singer, 2002) have thus been studied in depth during the last decade. Globally, the multi-class SVMs have been proposed more recently. They are all obtained by combining a multivariate affine model with the feature map Φ.

7.1 M-SVMs: Model and Function Selection

As in the bi-class case, the central element of a M-SVM is a *symmetric positive semidefinite (Mercer) kernel* (Aronszajn, 1950). Such kernels correspond to *positive type functions* (Berlinet and Thomas-Agnan, 2004). Let κ be a Mercer kernel on $\mathcal{X}$ and $\left(H_\kappa, \langle .,. \rangle_{H_\kappa}\right)$ the corresponding reproducing kernel Hilbert space (RKHS) (Berlinet and Thomas-Agnan, 2004). Let Φ be

any of the mappings on $\mathcal{X}$ satisfying:

$$\forall (x,x') \in \mathcal{X}^2,\ \kappa(x,x') = \langle \Phi(x), \Phi(x') \rangle, \tag{13}$$

where $\langle .,. \rangle$ is the dot product of the ℓ_2 space. Let $\Phi(\mathcal{X}) = \{\Phi(x) : x \in \mathcal{X}\}$ and let $\left(E_{\Phi(\mathcal{X})}, \langle .,. \rangle\right)$ be the Hilbert space spanned by $\Phi(\mathcal{X})$. According to the usual abuse of language, in the sequel, "the" *feature space* will designate any of the spaces $E_{\Phi(\mathcal{X})}$. By definition of a RKHS, $\mathcal{H} = \left(\left(H_\kappa, \langle .,. \rangle_{H_\kappa}\right) + \{1\}\right)^Q$ is the class of functions $h = (h_k)_{1 \le k \le Q}$ of the form:

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

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 $\{x_{ik} : 1 \le i \le m_k\}$ become dense in $\mathcal{X}$ in the norm induced by the dot product (see for instance Wahba, 1999). Due to (13), $\mathcal{H}$ can also be seen as a multivariate affine model on $\Phi(\mathcal{X})$. Functions h can then be rewritten as:

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

where the vectors w_k are elements of $E_{\Phi(\mathcal{X})}$. They are thus described by the pair $(\mathbf{w}, \mathbf{b})$ with $\mathbf{w} = (w_k)_{1 \le k \le Q} \in E_{\Phi(\mathcal{X})}^Q$ and $\mathbf{b} = (b_k)_{1 \le k \le Q} \in \mathbb{R}^Q$. Let $\bar{\mathcal{H}}$ stand for the product space H_κ^Q whose functions $\bar{h} = (\langle w_k, . \rangle)_{1 \le k \le Q}$ are seen as functions on $\Phi(\mathcal{X})$. Its norm $\|.\|_{\bar{\mathcal{H}}}$ is given by:

$$\forall \bar{h} \in \bar{\mathcal{H}},\ \left\|\bar{h}\right\|_{\bar{\mathcal{H}}} = \sqrt{\sum_{k=1}^{Q} \|w_k\|^2} = \|\mathbf{w}\|,$$

where $\|w_k\| = \sqrt{\langle w_k, w_k \rangle}$. $\bar{\mathcal{H}}$ also represents the restriction of $\mathcal{H}$ to the functions satisfying $\mathbf{b} = 0$. For convenience, $E_{\Phi(\mathcal{X})}^Q$ is endowed with a second norm, $\|.\|_\infty$. It is defined by $\|\mathbf{w}\|_\infty = \max_{1 \le k \le Q} \|w_k\|$. With these definitions at hand, a generic definition of the M-SVMs can be formulated as follows.

Definition 42 (M-SVM) *Let $((x_i, y_i))_{1 \le i \le m} \in (\mathcal{X} \times [\![1, Q]\!])^m$. 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:*

$$J(h) = \sum_{i=1}^{m} \ell_{\text{M-SVM}}(y_i, h(x_i)) + \lambda \|\mathbf{w}\|^2 \tag{14}$$

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

In accordance with the notations of Section 2.2 and Section 4.3, in what follows, $\tilde{\mathcal{H}}$ will designate the (univariate) affine model corresponding to the bi-class SVMs. The different M-SVMs only differ in the nature of the function $\ell_{\text{M-SVM}}$. This one is systematically built around the standard *hinge loss* of bi-class SVMs. This function, from $\tilde{\mathcal{H}} \times \mathcal{X} \times \{-1, 1\}$ into $\mathbb{R}_+$, maps $\left(\tilde{h}, x, y\right)$ to $\left(1 - y\tilde{h}(x)\right)_+$, where $(t)_+ = \max(0, t)$. Three main models of M-SVMs can be found in literature. The first one in chronological order was introduced independently by Weston and Watkins (1998) and by Blanz and Vapnik (Blanz, personal communication). It corresponds to a loss function ℓ_{WW} given by:

$\ell_{\mathrm{WW}}(y,h(x)) = \sum_{k\neq y}(1-h_y(x)+h_k(x))_+$. Then came the model of Crammer and Singer (2001), model built around $\mathcal{H}$, one advantage of which consists in the fact that it requires one single slack variable per training example. Its loss function is $\ell_{\mathrm{CS}}\left(y,\overline{h}(x)\right) = \left(1-\overline{h}_y(x)+\max_{k\neq y}\overline{h}_k(x)\right)_+$. The last model to date is the one of Lee et al. (2004), where $\ell_{\mathrm{LLW}}(y,h(x)) = \sum_{k\neq y}\left(h_k(x)+\frac{1}{Q-1}\right)_+$. Its specificity is that asymptotically, it implements the optimal classification rule, that is, Bayes decision rule. Indeed, this property of *infinite-sample consistency* is not shared by the two first M-SVMs, as was shown by Zhang (2004) and Tewari and Bartlett (2007). For all three machines, a representer theorem establishes that the function selected by the training procedure is of the form:

$$h(.) = \left(\sum_{i=1}^{m}\beta_{ik}\kappa(x_i,.)+b_k\right)_{1\leq k\leq Q}. \tag{15}$$

The lines of reasoning highlighting the fact that a bi-class SVM is intrinsically a large margin classifier can be extended easily to the M-SVMs. This requires however to discuss the form taken by the penalty component of the objective function. Indeed, the notion of multi-class margin given by Definition 5 involves differences between outputs, which suggests to use such penalty terms as $\max_{k<l}\|w_k-w_l\|^2$ or $\sum_{k<l}\|w_k-w_l\|^2$. However, this raises the difficulty that the function minimizing (14) is then defined up to an additive constant. The solution is provided by the restriction $\sum_k h_k = 0$. Under this hypothesis, the equation $\sum_{k<l}\|w_k-w_l\|^2 = Q\sum_k\|w_k\|^2 = Q\|\mathbf{w}\|^2$ justifies the use of $\|\mathbf{w}\|^2$ as penalty term.

Our generalized Sauer-Shelah lemma, Lemma 39, holds for classes of functions with bounded range (taking values in $[-M,M]^Q$). We now introduce the standard hypotheses on $\mathcal{X}$ ($\Phi(\mathcal{X})$) and $\mathcal{H}$ which will allow us to formulate the upper bound on the margin Natarajan dimension of interest.

Hypotheses 43 *To upper bound the capacity of a Q-category M-SVM, the following hypotheses and constraints are introduced regarding its domain and its parameters:*

1. *$\Phi(\mathcal{X})$ is included in the ball of radius $\Lambda_{\Phi(\mathcal{X})}$ about the origin in $E_{\Phi(\mathcal{X})}$;*
2. *the vector $\mathbf{w}$ satisfies $\|\mathbf{w}\|_\infty \leq \Lambda_w$;*
3. *the vector $\mathbf{b}$ belongs to $[-\beta,\beta]^Q$.*

With these hypotheses at hand, Lemma 39 can be applied to the corresponding subset of $\mathcal{H}$, by setting $M = \Lambda_w\Lambda_{\Phi(\mathcal{X})}+\beta$.

7.2 Switching from $\Delta^{\#}\mathcal{H}$ to $\Delta^{\#}\overline{\mathcal{H}}$

The computation of an upper bound on the margin Natarajan dimension is easier when the model is linear than when it is affine. Exactly as in the case of the transition from the class $\Delta^{\#}_{\gamma}\mathcal{G}$ to the class $\Delta^{\#}\mathcal{G}$ (see Section 5.1), the corresponding change is easier to perform when working with covering numbers. To that end, one can make use of the following lemma, the proof of which is inspired from the proof of Lemma 2.4 in Alon et al. (1997).

Lemma 44 *Let $\underline{\mathcal{H}}$ be the class of functions that a Q-category M-SVM can implement under Hypotheses 43. Let $\mathcal{H}$ be the subset of $\mathcal{H}$ corresponding to the functions satisfying $\mathbf{b}=0$. Let $\varepsilon \in \mathbb{R}^*_+$*

and $n \in \mathbb{N}^*$. Then

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

Proof Let $B =$

$$\left\{-\beta, -\left(\left\lceil \frac{\beta}{\varepsilon} \right\rceil - 1\right)\varepsilon, -\left(\left\lceil \frac{\beta}{\varepsilon} \right\rceil - 2\right)\varepsilon, \ldots, -2\varepsilon, -\varepsilon, 0, \varepsilon, 2\varepsilon, \ldots, \left(\left\lceil \frac{\beta}{\varepsilon} \right\rceil - 2\right)\varepsilon, \left(\left\lceil \frac{\beta}{\varepsilon} \right\rceil - 1\right)\varepsilon, \beta\right\}.$$

By construction, B^Q is a proper $\varepsilon/2$-net of $[-\beta, \beta]^Q$ in the ℓ_∞ norm. For $x^n \in \mathcal{X}^n$, let $\overline{\Delta^{\#}\bar{\mathcal{H}}}(\varepsilon, x^n)$ be a proper $\varepsilon/2$-net of $\Delta^{\#}\bar{\mathcal{H}}$ in the d_{x^n} pseudo-metric. We make the assumption that $\overline{\Delta^{\#}\bar{\mathcal{H}}}(\varepsilon, x^n)$ is of minimal cardinality, that is to say $\left|\overline{\Delta^{\#}\bar{\mathcal{H}}}(\varepsilon, x^n)\right| = \mathcal{N}^{(p)}(\varepsilon/2, \Delta^{\#}\bar{\mathcal{H}}, d_{x^n})$. Then, due to the triangle inequality, $\overline{\Delta^{\#}\bar{\mathcal{H}}}(\varepsilon, x^n) \times B^Q$ is a proper ε-net of $\Delta^{\#}\mathcal{H}$ in the d_{x^n} pseudo-metric. Since the cardinality of B^Q is $\left(2\left\lceil \frac{\beta}{\varepsilon} \right\rceil + 1\right)^Q$, this ε-net is of cardinality $\left(2\left\lceil \frac{\beta}{\varepsilon} \right\rceil + 1\right)^Q \mathcal{N}^{(p)}(\varepsilon/2, \Delta^{\#}\bar{\mathcal{H}}, d_{x^n})$. As a consequence, $\mathcal{N}^{(p)}(\varepsilon, \Delta^{\#}\mathcal{H}, d_{x^n}) \leq \left(2\left\lceil \frac{\beta}{\varepsilon} \right\rceil + 1\right)^Q \mathcal{N}^{(p)}(\varepsilon/2, \Delta^{\#}\bar{\mathcal{H}}, d_{x^n})$. Taking the maximum of both sides of this inequality over all the possible sequences x^n in $\mathcal{X}^n$ thus concludes the proof. ■

Note that, with little additional work, a tighter bound results from exploiting the restriction $\sum_k h_k = 0$.

7.3 Upper Bounding the Margin Natarajan Dimension of $\Delta\bar{\mathcal{H}}$

In this section, we follow the sketch of the proof of Theorem 4.6 in Bartlett and Shawe-Taylor (1999).

Lemma 45 *Let $\bar{\mathcal{H}}$ be the class of functions that a Q-category M-SVM can implement under Hypotheses 43, and the additional constraint $\mathbf{b} = 0$. Let $\underline{\varepsilon} \in \mathbb{R}_+^*$ and $n \in \mathbb{N}^*$. If a subset $s_{\mathcal{X}^n} = \{x_i : 1 \leq i \leq n\}$ of $\mathcal{X}$ is N-shattered with margin ε by $\Delta\bar{\mathcal{H}}$, then there exists a subset $s_{\mathcal{X}^p}$ of $s_{\mathcal{X}^n}$ of cardinality p equal to $\left\lceil \frac{n}{\binom{Q}{2}} \right\rceil$ such that for every partition of $s_{\mathcal{X}^p}$ into two subsets s_1 and s_2, the following bound holds true:*

$$\left\| \sum_{x_i \in s_1} \Phi(x_i) - \sum_{x_i \in s_2} \Phi(x_i) \right\| \geq \frac{\left\lceil \frac{n}{\binom{Q}{2}} \right\rceil}{\Lambda_w} \varepsilon. \tag{16}$$

Proof Suppose that $s_{\mathcal{X}^n} = \{x_i : 1 \leq i \leq n\}$ is a subset of $\mathcal{X}$ N-shattered with margin ε by $\Delta\bar{\mathcal{H}}$. Let $(I(s_{\mathcal{X}^n}), v_b)$ witness this shattering. Thanks to Proposition 30, without loss of generality, we can assume that $I(s_{\mathcal{X}^n})$ satisfies the constraint: $\forall i \in [\![1, n]\!]$, $i_1(x_i) < i_2(x_i)$. According to the pigeonhole principle, there is at least one couple of indexes (k_0, l_0) with $1 \leq k_0 < l_0 \leq Q$ such that there are at least $p = \left\lceil \frac{n}{\binom{Q}{2}} \right\rceil$ points in $s_{\mathcal{X}^n}$ for which the couple $(i_1(x_i), i_2(x_i))$ is (k_0, l_0). For the sake of simplicity, the points in $s_{\mathcal{X}^n}$ are reordered in such a way that the p first of them exhibit this property. The corresponding subset of $s_{\mathcal{X}^n}$ is denoted $s_{\mathcal{X}^p}$. This means that for all vector $v_y = (y_i)$ in $\{-1, 1\}^n$,

there is a function $\bar{h}_y$ in $\bar{\mathcal{H}}$ characterized by the vector $\mathbf{w}_y = (w_{y,k})_{1\leq k\leq Q}$ such that:

$$\forall i \in [\![1,p]\!], \left\{ \begin{array}{ll} \text{if } y_i = & 1, \quad \Delta\bar{h}_{y,k_0}(x_i) - b_i \geq \varepsilon \\ \text{if } y_i = & -1, \quad \Delta\bar{h}_{y,l_0}(x_i) + b_i \geq \varepsilon \end{array} \right. . \tag{17}$$

By definition of $\bar{\mathcal{H}}$ and the margin operator Δ, this is equivalent to:

$$\forall i \in [\![1,p]\!], \left\{ \begin{array}{ll} \text{if } y_i = & 1, \quad \frac{1}{2}\left(\langle w_{y,k_0}, \Phi(x_i)\rangle - \max_{k\neq k_0}\langle w_{y,k}, \Phi(x_i)\rangle\right) - b_i \geq \varepsilon \\ \text{if } y_i = & -1, \quad \frac{1}{2}\left(\langle w_{y,l_0}, \Phi(x_i)\rangle - \max_{k\neq l_0}\langle w_{y,k}, \Phi(x_i)\rangle\right) + b_i \geq \varepsilon \end{array} \right.$$

and thus implies

$$\forall i \in [\![1,p]\!], \left\{ \begin{array}{ll} \text{if } y_i = & 1, \quad \frac{1}{2}\langle w_{y,k_0} - w_{y,l_0}, \Phi(x_i)\rangle - b_i \geq \varepsilon \\ \text{if } y_i = & -1, \quad \frac{1}{2}\langle w_{y,l_0} - w_{y,k_0}, \Phi(x_i)\rangle + b_i \geq \varepsilon \end{array} \right. . \tag{18}$$

Consider now any partition of $s_{\mathcal{X}^p}$ into two subsets s_1 and s_2. Consider any vector v_y in $\{-1,1\}^n$ such that $y_i = 1$ if $x_i \in s_1$ and $y_i = -1$ if $x_i \in s_2$. It results from (18) that:

$$\frac{1}{2}\langle w_{y,k_0} - w_{y,l_0}, \sum_{x_i\in s_1}\Phi(x_i)\rangle - \sum_{x_i\in s_1} b_i + \frac{1}{2}\langle w_{y,l_0} - w_{y,k_0}, \sum_{x_i\in s_2}\Phi(x_i)\rangle + \sum_{x_i\in s_2} b_i \geq |s_{\mathcal{X}^p}|\,\varepsilon$$

which simplifies into

$$\frac{1}{2}\langle w_{y,k_0} - w_{y,l_0}, \sum_{x_i\in s_1}\Phi(x_i) - \sum_{x_i\in s_2}\Phi(x_i)\rangle - \sum_{x_i\in s_1} b_i + \sum_{x_i\in s_2} b_i \geq p\varepsilon.$$

Conversely, consider any vector v_y such that $y_i = -1$ if $x_i \in s_1$ and $y_i = 1$ if $x_i \in s_2$. We have:

$$\frac{1}{2}\langle w_{y,l_0} - w_{y,k_0}, \sum_{x_i\in s_1}\Phi(x_i) - \sum_{x_i\in s_2}\Phi(x_i)\rangle + \sum_{x_i\in s_1} b_i - \sum_{x_i\in s_2} b_i \geq p\varepsilon.$$

As a consequence, if $\sum_{x_i\in s_1} b_i - \sum_{x_i\in s_2} b_i \geq 0$, there is a function $\bar{h}_y$ in $\bar{\mathcal{H}}$ such that

$$\frac{1}{2}\langle w_{y,k_0} - w_{y,l_0}, \sum_{x_i\in s_1}\Phi(x_i) - \sum_{x_i\in s_2}\Phi(x_i)\rangle \geq \left\lceil \frac{n}{\binom{Q}{2}} \right\rceil \varepsilon \tag{19}$$

whereas if $\sum_{x_i\in s_1} b_i - \sum_{x_i\in s_2} b_i < 0$, there is another function $\bar{h}_y$ in $\bar{\mathcal{H}}$ such that

$$\frac{1}{2}\langle w_{y,l_0} - w_{y,k_0}, \sum_{x_i\in s_1}\Phi(x_i) - \sum_{x_i\in s_2}\Phi(x_i)\rangle \geq \left\lceil \frac{n}{\binom{Q}{2}} \right\rceil \varepsilon. \tag{20}$$

Applying the Cauchy-Schwarz inequality to (19) and (20) yields

$$\frac{1}{2}\,\|w_{y,k_0} - w_{y,l_0}\| \left\| \sum_{x_i\in s_1}\Phi(x_i) - \sum_{x_i\in s_2}\Phi(x_i) \right\| \geq \left\lceil \frac{n}{\binom{Q}{2}} \right\rceil \varepsilon,$$

which thus holds true irrespective of the value of $\sum_{x_i\in s_1} b_i - \sum_{x_i\in s_2} b_i$. Finally, (16) directly springs from this last bound, as a consequence of fact that the constraint $\|\mathbf{w}\|_\infty \leq \Lambda_w$ implies $1/2\max_{1\leq k<l\leq Q}\|w_k - w_l\| \leq \Lambda_w$. ■

Remark 46 *The proof of Lemma 45 does not hold any more if one uses the Δ^* operator in place of the Δ operator. Indeed, reformulating (17) with Δ^* in place of Δ, one cannot derive (18) any more. This is precisely the reason why it is specifically the Δ operator which appears in the hypotheses of Lemma 45 and, by way of consequence, the final bound on the margin Natarajan dimension (see Theorem 48 below).*

Lemma 47 (Bartlett and Shawe-Taylor, 1999, Lemma 4.3) *If $\Phi(\mathcal{X})$ is included in the ball of radius $\Lambda_{\Phi(\mathcal{X})}$ about the origin in $E_{\Phi(\mathcal{X})}$, then for all $n \in \mathbb{N}^*$, all subset $s_{\mathcal{X}^n} = \{x_i : 1 \leq i \leq n\}$ of $\mathcal{X}$ can be partitioned into two subsets s_1 and s_2 satisfying*

$$\left\| \sum_{x_i \in s_1} \Phi(x_i) - \sum_{x_i \in s_2} \Phi(x_i) \right\| \leq \sqrt{n} \Lambda_{\Phi(\mathcal{X})}. \tag{21}$$

The following theorem is a direct consequence of Lemma 45 and Lemma 47.

Theorem 48 *Let $\bar{\mathcal{H}}$ be the class of functions that a Q-category M-SVM can implement under Hypotheses 43, and the additional constraint $\mathbf{b} = 0$. Then, for any positive real value ε, the following bound holds true:*

$$N\text{-}dim\left(\Delta\bar{\mathcal{H}}, \varepsilon\right) \leq \binom{Q}{2} \left(\frac{\Lambda_w \Lambda_{\Phi(\mathcal{X})}}{\varepsilon}\right)^2. \tag{22}$$

Proof Let $s_{\mathcal{X}^n}$ be a subset of $\mathcal{X}$ of cardinality n N-shattered with margin ε by $\Delta\bar{\mathcal{H}}$. According to Lemma 45, there is at least a subset $s_{\mathcal{X}^p}$ of $s_{\mathcal{X}^n}$ of cardinality $p = \left\lceil \frac{n}{\binom{Q}{2}} \right\rceil$ satisfying (16) for all its partitions into two subsets s_1 and s_2. Since, according to Lemma 47, there is at least one of these partitions for which (21) holds true,

$$\frac{p}{\Lambda_w} \varepsilon \leq \sqrt{p} \Lambda_{\Phi(\mathcal{X})}$$

which implies that

$$p \leq \left(\frac{\Lambda_w \Lambda_{\Phi(\mathcal{X})}}{\varepsilon}\right)^2.$$

Since $n \leq \binom{Q}{2} p$, one finally obtains

$$n \leq \binom{Q}{2} \left(\frac{\Lambda_w \Lambda_{\Phi(\mathcal{X})}}{\varepsilon}\right)^2$$

which concludes the proof. ■

7.4 Discussion

Proposition 31 states that in the bi-class case, there is only one γ-Ψ-dimension, which corresponds to the fat-shattering dimension. Thus, it is satisfactory to notice that for $Q = 2$, (22) becomes

$$P_\varepsilon\text{-dim}\left(H_\kappa\right) \leq \left(\frac{\Lambda_w \Lambda_{\Phi(\mathcal{X})}}{\varepsilon}\right)^2$$

which is precisely the bound provided by Theorem 4.6 in Bartlett and Shawe-Taylor (1999) (see also Remark 1 in Gurvits, 2001), that is, the tightest bound on the fat-shattering dimension of a linear classifier currently available. In the general case, Theorem 48 tells us that the margin Natarajan dimension of a Q-category M-SVM can be bounded from above by a uniform bound on the fat-shattering dimensions of its separating hyperplanes (defined by the equation $\langle w_k - w_l, \Phi(x) \rangle = 0$) times the number of those hyperplanes, $\binom{Q}{2}$. It must be borne in mind that this expression is directly connected with the idea at the basis of the definition of the Ψ-dimensions (see the discussion in Section 4.1), which is to simulate the implementation of a decomposition scheme, and take benefit of this to make use of standard bi-class results. In the case of the Natarajan dimension, this scheme corresponds to the one-against-one method. The terms $\binom{Q}{2}$ and $\|\mathbf{w}\|_\infty$ then appear in (22) as a consequence of the fact that all the pairs of categories are considered independently one from the other and play an utterly symmetrical part (we need a bound on $1/2 \max_{1 \leq k < l \leq Q} \|w_k - w_l\|$). Obviously, a tighter bound should result from taking into account the fact that the $\binom{Q}{2}$ binary classifiers are not independent, since they are based on a common set of Q vectors w_k. Here appears once more the need to derive original solutions for the multi-class case, instead of simple extensions of bi-class results.

Deriving a nontrivial bound on N-dim $\left(\Delta\bar{\mathcal{H}}, \varepsilon\right)$ in terms of $\|\mathbf{w}\|$, that is, a tighter bound than the one resulting from just replacing in the hypotheses of Theorem 48 $\|\mathbf{w}\|_\infty$ with $\|\mathbf{w}\|$, remains an open problem. The fact that the norm used in the penalty term of the objective function (14) and the one appearing in the upper bound on the margin Natarajan dimension are different is unsatisfactory. The point is that, so far, no one has put forward a theoretical argument (guaranteed risk) to justify the use of $\|\mathbf{w}\|$, whereas the use of $\|\mathbf{w}\|_\infty$ as penalty term, considered only in Guermeur (2002), raises significant technical difficulties. Indeed, in that case, the convex programming problem corresponding to the training algorithm cannot be solved by means of Lagrangian duality any more, since one cannot compute the gradient of the Lagrangian function with respect to the vectors w_k. In that sense, there remains a gap to fill between theory and practice.

8. γ-Ψ-dimensions and Implementation of the SRM Inductive Principle

In this section, we discuss the significance of the main results of the paper. We first summarize the specificities of the multi-class case highlighted by their proofs, and then outline an application of our bound on the risk of M-SVMs for model selection.

8.1 Characterization of Relevant Information

The main results of this article involve two distinct margin operators, Δ and Δ^*. Theorem 22, the basic uniform convergence result on which all this study is based, holds true for both of them. However, we pointed out in Remark 36 the reason why the generalized Sauer-Shelah lemma (Lemmas 38 and 39) requires specifically the use of Δ^*. On the contrary, Remark 46 highlights the fact that the proof of the bound on the margin Natarajan dimension of the M-SVMs, Theorem 48, makes use of a specific property of Δ. Fortunately, the connection between the capacities of $\Delta^* \mathcal{G}$ and $\Delta \mathcal{G}$ is provided by Lemmas 35 and 37. These observations highlight the fact that the link between separation and shattering capacity is more complex in the multi-class case than in the bi-class case (for which we simply have $\Delta = \Delta^*$). At different steps of the reasoning, different pieces of information on the behaviour of the functions of interest are needed. One must provide neither too many nor too few of

them. It is a bit disappointing to notice that the computation of the bound on the margin Natarajan dimension requires more information than simply the index of the highest output and the difference between the two highest outputs, that is, what is relevant to determine both the classification performed and the confidence one can have in the accuracy of this classification. This suggests that some improvement could be made to our generalization of the standard bi-class results, regarding for instance the choice of the functional pseudo-metric. However, it is difficult to figure out how these changes could remain compatible with the whole line of reasoning leading to the bound on the risk of the M-SVMs. Indeed, the choices we made to extend the VC theory to the case of large margin multi-category discriminant models and apply it to M-SVMs were primarily governed by one concern: allowing a natural extension of the proof of Lemma 3.3 in Alon et al. (1997) and the proof of Theorem 4.6 in Bartlett and Shawe-Taylor (1999) to the multi-class case. As a consequence, the question could be now: can we develop our theory without making use of those two pillars of the standard theory?

8.2 Application for Model Selection

When working with SVMs, performing model selection amounts to choosing the value of the "soft margin parameter" C, the kernel κ and the values of its parameters. Cross-validation was initially regarded as the method of choice to perform this task, although it exhibits some drawbacks, as was first pointed out by Stone (1977). This strategy has induced many authors to derive upper bounds on the leave-one-out error of SVMs (see Chapelle et al., 2002, for a survey). The most widely used of them is probably the famous "radius-margin bound", for which several multi-class extensions have been proposed independently by Wang et al. (2005); Darcy and Guermeur (2005); Monfrini and Guermeur (2007), as criteria for the choice of the values of the hyperparameters of M-SVMs (or SVMs involved in decomposition schemes). Care was taken to the fact that they could be differentiated with respect to those parameters, in order to make the optimization procedure tractable.

With that difficulty in mind, it appears that model selection for SVMs, either bi-class or multi-class, made a great stride when Hastie et al. (2004) introduced their algorithm fitting the entire path of SVM solutions for every value of C (see also Lee and Cui, 2006, for an algorithm dedicated to the M-SVM of Lee and co-authors). Indeed, with this algorithm at hand, requirements in computational time are drastically reduced, which makes it possible to use new criteria (tighter bounds on the risk) for the selection of C. The idea is simple: starting with a small value of C, it suffices to follow the path, that is, increase progressively the value of C, and assess the bound at each step. Eventually, the value selected is the one corresponding to the smallest value of the bound. This is precisely what was done in Guermeur et al. (2005). In that paper, taking our inspiration from Williamson et al. (2000), we used a bound on the generalization error of M-SVMs obtained as a function of a bound on the entropy numbers of the evaluation operator. The corresponding experimental protocol provides us with an easy way to assess the usefulness of our new bound for model selection. For a given position in the path (a given value of C), all what has to be done is to optimize the guaranteed risk with respect to the margin parameter γ. With the notation introduced in (15), the formula at the basis of the computation of the upper bound on the margin Natarajan dimension is the following one:

$$\forall (k,l) \in [\![1,Q]\!]^2 \,,\ \|w_k - w_l\|^2 = \sum_{i=1}^{m} \sum_{j=1}^{m} (\beta_{ik} - \beta_{il}) \left(\beta_{jk} - \beta_{jl}\right) \kappa(x_i, x_j)$$

(obviously, one benefits from using in the computations $\frac{1}{4}\max_{k<l}\|w_k - w_l\|^2$ in place of its upper bound $\|\mathbf{w}\|_\infty^2$).

9. Conclusions and Ongoing Research

In this article, the standard theories of large margin bi-class classifiers and Q-class classifiers taking values in $[\![1,Q]\!]$ have been unified to give birth to a VC theory of large margin multi-class classifiers. This could be done in a straightforward way, by extending concepts and results from only four references: Ben-David et al. (1995), Alon et al. (1997), Bartlett (1998), and Bartlett and Shawe-Taylor (1999). The main difficulty was to identify the need to introduce two margin operators, Δ and Δ^*. The generalized VC dimensions at the center of the new theory are the γ-Ψ-dimensions. They can be seen either as scale-sensitive extensions of the Ψ-dimensions, or multivariate extensions of the fat-shattering dimension. In particular, they characterize learnability for the classes of functions of interest.

It is possible to select the most appropriate of these dimensions as a function of the model studied. In the case of the multi-class SVMs, we have found the margin Natarajan dimension to be the easiest to bound from above making use of standard results derived with the fat-shattering dimension. As a consequence, all the M-SVMs proposed so far can now be evaluated in the unifying framework of the implementation of the SRM inductive principle. Indeed, the main practical interest of guaranteed risks based on γ-Ψ-dimensions should regard the implementation of this learning principle. They make it possible to characterize the variation of the capacity of large margin multi-category discriminant models based on classes of parametric functions with respect to the constraints on their domain and parameters. An obvious application of this study is in model selection, for instance to choose the values of the "soft margin parameter" C and the kernel parameters of M-SVMs.

Readers more interested in computing sample complexities than in the characterization of Glivenko-Cantelli classes, capacity control or model selection, should be aware of the fact that sharper bounds should result from using different sources of inspiration, although even in that case, the lessons drawn from the present study should still prove useful. An obvious possibility is represented by new PAC-Bayes bounds (Ambroladze et al., 2007), or, to remain nearer to the present study, new tools of concentration theory and empirical processes (Talagrand, 1995, 1996; Ledoux, 1996; Massart, 2000; Lugosi, 2004). They make it possible, for instance, to work with data dependent capacity measures such as the empirical VC entropy. A great survey of the recent advances in this field, especially focusing on Rademacher averages, is provided by Boucheron et al. (2005). Regarding more specifically pattern recognition SVMs, the results the extension of which appears most promising are those reported in Bousquet (2002), Steinwart and Scovel (2005), and Blanchard et al. (2007). Performing these multi-class extensions is the subject of an ongoing work.

Acknowledgments

This work was initiated with H. Paugam-Moisy and A. Elisseeff. The author would like to thank the anonymous reviewers for their comments. It is also a pleasure to thank M. Sebag, P. Bartlett, S. Kroon, R. Vert and M. Warmuth for instructive discussions and bibliographical help, as well as E. Monfrini and F. Sur for carefully reading this manuscript.

Appendix A. Technical Lemmas

This appendix is devoted to technical lemmas that are at the basis of the proofs of the main theorems of the paper.

Lemma 49 Jogdeo and Samuels, 1968, Theorem 3.2. *Let T be a random variable described by a binomial distribution with parameters n and p ($T \hookrightarrow \mathcal{B}(n,p)$). Then its median is either $\lfloor np \rfloor$ or $\lfloor np \rfloor + 1$. Moreover, if np is an integer, the median is simply np.*

Lemma 50 *Let $D_{2m} = ((X_i,Y_i))_{1\leq i\leq 2m}$ be a $2m$-sample of independent copie of (X,Y). Let $D_m = ((X_i,Y_i))_{1\leq i\leq m}$ and $\tilde{D}_m = \left(\left(\tilde{X}_i,\tilde{Y}_i\right)\right)_{1\leq i\leq m} = ((X_{m+i},Y_{m+i}))_{1\leq i\leq m}$. $\mathbb{P}_{D_m}$ is a probability over the sample D_m, and $\mathbb{P}_{D_{2m}}$ is a probability over D_{2m}. The distribution of the random variable $\sup_{g\in\mathcal{G}}\left(R(g) - R_{\gamma,D_m}(g)\right)$ is connected with the distribution of the random variable $\sup_{g\in\mathcal{G}}\left(R_{\tilde{D}_m}(g) - R_{\gamma,D_m}(g)\right)$ by the inequality*

$$\mathbb{P}_{D_m}\left(\sup_{g\in\mathcal{G}}\left(R(g) - R_{\gamma,D_m}(g)\right) > \varepsilon\right) \leq 2\mathbb{P}_{D_{2m}}\left(\sup_{g\in\mathcal{G}}\left(R_{\tilde{D}_m}(g) - R_{\gamma,D_m}(g)\right) \geq \varepsilon - \frac{1}{m}\right).$$

Proof The proof of this lemma is inspired from the proof of Vapnik's basic lemma in Vapnik (1998, Section 4.5.1). For $n \in \mathbb{N}^*$, let $z^{2n} = ((x_i,y_i))_{1\leq i\leq 2n}$ be an element of $\mathcal{Z}^{2n}$. In what follows, we will use z^n to designate its "first half", whereas $\tilde{z}^n$, will designate its "second half". $\tilde{z}^n = ((\tilde{x}_i,\tilde{y}_i))_{1\leq i\leq n}$, with $(\tilde{x}_i,\tilde{y}_i) = (x_{n+i},y_{n+i})$. Since D_m and $\tilde{D}_m$ are supposed to be independent, by definition:

$$\mathbb{P}_{D_{2m}}\left(\sup_{g\in\mathcal{G}}\left(R_{\tilde{D}_m}(g) - R_{\gamma,D_m}(g)\right) \geq \varepsilon - \frac{1}{m}\right) =$$

$$\int_{\mathcal{Z}^{2m}} \mathbb{1}\left[\sup_{g\in\mathcal{G}}\left(R_{\tilde{z}^m}(g) - R_{\gamma,z^m}(g)\right) \geq \varepsilon - \frac{1}{m}\right] dP^{2m}(z^{2m}),$$

and one can apply Fubini's theorem for nonnegative measurable functions (see Rudin, 1987, Section 8.8) to the product measure P^{2m}, which gives:

$$\mathbb{P}_{D_{2m}}\left(\sup_{g\in\mathcal{G}}\left(R_{\tilde{D}_m}(g) - R_{\gamma,D_m}(g)\right) \geq \varepsilon - \frac{1}{m}\right) =$$

$$\int_{\mathcal{Z}^m} dP^m(z^m) \int_{\mathcal{Z}^m} \mathbb{1}\left[\sup_{g\in\mathcal{G}}\left(R_{\tilde{z}^m}(g) - R_{\gamma,z^m}(g)\right) \geq \varepsilon - \frac{1}{m}\right] dP^m(\tilde{z}^m).$$

In the inner integral, z^m is fixed. Let Q denote the following event:

$$Q = \left\{z^m = ((x_i,y_i))_{1\leq i\leq m} \in \mathcal{Z}^m : \sup_{g\in\mathcal{G}}\left(R(g) - R_{\gamma,z^m}(g)\right) > \varepsilon\right\}.$$

Restricting the integration domain to Q gives

$$\mathbb{P}_{D_{2m}}\left(\sup_{g\in\mathcal{G}}\left(R_{\tilde{D}_m}(g) - R_{\gamma,D_m}(g)\right) \geq \varepsilon - \frac{1}{m}\right) \geq$$

$$\int_{\mathcal{Q}} dP^m(z^m) \underbrace{\int_{\mathcal{Z}^m} \mathbb{1}\left[\sup_{g\in\mathcal{G}}\left(R_{\tilde{z}^m}(g) - R_{\gamma,z^m}(g)\right) \geq \varepsilon - \frac{1}{m}\right] dP^m(\tilde{z}^m)}_{I}. \tag{23}$$

I is an integral which is calculated for a fixed z^m satisfying

$$\sup_{g\in\mathcal{G}}\left(R(g) - R_{\gamma,z^m}(g)\right) > \varepsilon.$$

Consequently, there exists a function g^* in $\mathcal{G}$ such that

$$R(g^*) - R_{\gamma,z^m}(g^*) \geq \varepsilon.$$

By definition of g^*, the following inequality holds

$$I \geq \int_{\mathcal{Z}^m} \mathbb{1}\left[R_{\tilde{z}^m}(g^*) - R_{\gamma,z^m}(g^*) \geq \varepsilon - \frac{1}{m}\right] dP^m(\tilde{z}^m).$$

$$\left\{\begin{array}{l} R(g^*) - R_{\gamma,z^m}(g^*) \geq \varepsilon \\ R_{\tilde{z}^m}(g^*) - R(g^*) \geq -\frac{1}{m} \end{array}\right. \Longrightarrow R_{\tilde{z}^m}(g^*) - R_{\gamma,z^m}(g^*) \geq \varepsilon - \frac{1}{m}.$$

As a consequence

$$I \geq \int_{\mathcal{Z}^m} \mathbb{1}\left[R_{\tilde{z}^m}(g^*) - R(g^*) \geq -\frac{1}{m}\right] dP^m(\tilde{z}^m).$$

Furthermore

$$\int_{\mathcal{Z}^m} \mathbb{1}\left[R_{\tilde{z}^m}(g^*) - R(g^*) \geq -\frac{1}{m}\right] dP^m(\tilde{z}^m) = \mathbb{P}_{\tilde{D}_m}\left(mR_{\tilde{D}_m}(g^*) \geq mR(g^*) - 1\right). \tag{24}$$

By definition of $R(g^*)$ and $R_{\tilde{D}_m}(g^*)$, $mR_{\tilde{D}_m}(g^*)$ has a binomial distribution with parameters m and $R(g^*)$ $\left(mR_{\tilde{D}_m}(g^*) \hookrightarrow \mathcal{B}(m, R(g^*))\right)$. To bound from below the right-hand side of (24), we make use of a result on the median of random variables following a binomial distribution, Lemma 49. According to this lemma, $mR(g^*) - 1$ is inferior or equal to the median of $mR_{\tilde{D}_m}(g^*)$, and thus, by definition of the median, the right-hand side of (24) is superior or equal to $1/2$. By transitivity, I is also greater that $1/2$. Substituting this lower bound on I into (23) yields

$$\mathbb{P}_{D_{2m}}\left(\sup_{g\in\mathcal{G}}\left(R_{\tilde{D}_m}(g) - R_{\gamma,D_m}(g)\right) \geq \varepsilon - \frac{1}{m}\right) \geq \frac{1}{2}\int_{\mathcal{Q}} dP^m(z^m)$$

or equivalently, by definition of Q:

$$\mathbb{P}_{D_{2m}}\left(\sup_{g\in\mathcal{G}}\left(R_{\tilde{D}_m}(g) - R_{\gamma,D_m}(g)\right) \geq \varepsilon - \frac{1}{m}\right) \geq \frac{1}{2}\mathbb{P}_{D_m}\left(\sup_{g\in\mathcal{G}}\left(R(g) - R_{\gamma,D_m}(g)\right) > \varepsilon\right)$$

which is the result announced. ■

Lemma 51 *The distribution of the random variable* $\sup_{g\in\mathcal{G}}\left(R_{\tilde{D}_m}(g)-R_{\gamma,D_m}(g)\right)$ *is connected with the distribution of the random variable* $\max_{\overline{g}\in\overline{\mathcal{G}}(\gamma,D_{2m})}\left(R_{\gamma/2,\tilde{D}_m}(\overline{g})-R_{\gamma/2,D_m}(\overline{g})\right)$ *by the inequality*

$$\mathbb{P}_{D_{2m}}\left(\sup_{g\in\mathcal{G}}\left(R_{\tilde{D}_m}(g)-R_{\gamma,D_m}(g)\right)\geq\varepsilon-\frac{1}{m}\right)\leq$$

$$\mathbb{P}_{D_{2m}}\left(\max_{\overline{g}\in\overline{\mathcal{G}}(\gamma,D_{2m})}\left(R_{\gamma/2,\tilde{D}_m}(\overline{g})-R_{\gamma/2,D_m}(\overline{g})\right)\geq\varepsilon-\frac{1}{m}\right).$$

Proof $\forall g\in\mathcal{G},\forall(x_i,y_i)\in z^{2m}$,

$$\begin{cases}\Delta^{\#}g_{y_i}(x_i)\leq 0\\ d_{x^{2m}}(\Delta^{\#}_{\gamma}g,\Delta^{\#}_{\gamma}\overline{g})<\frac{\gamma}{2}\end{cases}\Longrightarrow\Delta^{\#}\overline{g}_{y_i}(x_i)<\frac{\gamma}{2}. \tag{25}$$

Similarly,

$$\begin{cases}\Delta^{\#}\overline{g}_{y_i}(x_i)<\frac{\gamma}{2}\\ d_{x^{2m}}(\Delta^{\#}_{\gamma}g,\Delta^{\#}_{\gamma}\overline{g})<\frac{\gamma}{2}\end{cases}\Longrightarrow\Delta^{\#}g_{y_i}(x_i)<\gamma. \tag{26}$$

From (25) it results that if $d_{x^{2m}}(\Delta^{\#}_{\gamma}g,\Delta^{\#}_{\gamma}\overline{g})<\frac{\gamma}{2}$, then

$$R_{\tilde{z}^m}(g)\leq R_{\gamma/2,\tilde{z}^m}(\overline{g}).$$

Similarly, it results from (26) that if $d_{x^{2m}}(\Delta^{\#}_{\gamma}g,\Delta^{\#}_{\gamma}\overline{g})<\frac{\gamma}{2}$, then

$$R_{\gamma/2,z^m}(\overline{g})\leq R_{\gamma,z^m}(g).$$

To sum up, for all g in $\mathcal{G}$, there exists $\overline{g}$ in $\overline{\mathcal{G}}(\gamma,x^{2m})$ such that

$$R_{\tilde{z}^m}(g)-R_{\gamma,z^m}(g)\leq R_{\gamma/2,\tilde{z}^m}(\overline{g})-R_{\gamma/2,z^m}(\overline{g})$$

and thus

$$\mathbb{P}_{D_{2m}}\left(\sup_{g\in\mathcal{G}}\left(R_{\tilde{D}_m}(g)-R_{\gamma,D_m}(g)\right)\geq\varepsilon-\frac{1}{m}\right)=$$

$$\int_{\mathcal{Z}^{2m}}\mathbb{1}\left[\sup_{g\in\mathcal{G}}\left(R_{\tilde{z}^m}(\overline{g})-R_{\gamma,z^m}(\overline{g})\right)\geq\varepsilon-\frac{1}{m}\right]dP^{2m}(z^{2m})\leq$$

$$\int_{\mathcal{Z}^{2m}}\mathbb{1}\left[\max_{\overline{g}\in\overline{\mathcal{G}}(\gamma,x^{2m})}\left(R_{\gamma/2,\tilde{z}^m}(\overline{g})-R_{\gamma/2,z^m}(\overline{g})\right)\geq\varepsilon-\frac{1}{m}\right]dP^{2m}(z^{2m})=$$

$$\mathbb{P}_{D_{2m}}\left(\max_{\overline{g}\in\overline{\mathcal{G}}(\gamma,D_{2m})}\left(R_{\gamma/2,\tilde{D}_m}(\overline{g})-R_{\gamma/2,D_m}(\overline{g})\right)\geq\varepsilon-\frac{1}{m}\right).$$

■

Lemma 52 *Let S_{2m} be a random variable described by the uniform distribution on $\mathfrak{T}_{2m}$. Then*

$$\mathbb{P}_{D_{2m}}\left(\max_{\overline{g}\in\overline{\mathcal{G}}(\gamma,D_{2m})}\left(R_{\gamma/2,\tilde{D}_m}(\overline{g})-R_{\gamma/2,D_m}(\overline{g})\right)\geq\varepsilon-\frac{1}{m}\right)\leq$$

$$\max_{z^{2m}\in\mathcal{Z}^{2m}}\sum_{\overline{g}\in\overline{\mathcal{G}}(\gamma,x^{2m})}\mathbb{P}_{S_{2m}}\left(R_{\gamma/2,S_{2m}(\tilde{z}^m)}(\overline{g})-R_{\gamma/2,S_{2m}(z^m)}(\overline{g})\geq\varepsilon-\frac{1}{m}\right).$$

Proof Since coordinate permutations preserve the product distribution P^{2m},

$$\mathbb{P}_{D_{2m}}\left(\max_{\overline{g}\in\overline{\mathcal{G}}(\gamma,D_{2m})}\left(R_{\gamma/2,\tilde{D}_m}(\overline{g})-R_{\gamma/2,D_m}(\overline{g})\right)\geq\varepsilon-\frac{1}{m}\right)$$

is not affected by a permutation σ. One thus obtains:

$$\forall\sigma\in\mathfrak{T}_{2m},\ \mathbb{P}_{D_{2m}}\left(\max_{\overline{g}\in\overline{\mathcal{G}}(\gamma,D_{2m})}\left(R_{\gamma/2,\tilde{D}_m}(\overline{g})-R_{\gamma/2,D_m}(\overline{g})\right)\geq\varepsilon-\frac{1}{m}\right)=$$

$$\int_{\mathcal{Z}^{2m}}\mathbb{1}\left[\max_{\overline{g}\in\overline{\mathcal{G}}(\gamma,x^{2m})}\left(R_{\gamma/2,\sigma(\tilde{z}^m)}(\overline{g})-R_{\gamma/2,\sigma(z^m)}(\overline{g})\right)\geq\varepsilon-\frac{1}{m}\right]dP^{2m}(z^{2m}).$$

Averaging the summand of the right-hand side over the whole set $\mathfrak{T}_{2m}$ gives:

$$\mathbb{P}_{D_{2m}}\left(\max_{\overline{g}\in\overline{\mathcal{G}}(\gamma,D_{2m})}\left(R_{\gamma/2,\tilde{D}_m}(\overline{g})-R_{\gamma/2,D_m}(\overline{g})\right)\geq\varepsilon-\frac{1}{m}\right)=$$

$$\frac{1}{|\mathfrak{T}_{2m}|}\sum_{\sigma\in\mathfrak{T}_{2m}}\int_{\mathcal{Z}^{2m}}\mathbb{1}\left[\max_{\overline{g}\in\overline{\mathcal{G}}(\gamma,x^{2m})}\left(R_{\gamma/2,\sigma(\tilde{z}^m)}(\overline{g})-R_{\gamma/2,\sigma(z^m)}(\overline{g})\right)\geq\varepsilon-\frac{1}{m}\right]dP^{2m}(z^{2m}).$$

Since the cardinality of $\mathfrak{T}_{2m}$ is finite, summation and integration can be interchanged as follows:

$$\mathbb{P}_{D_{2m}}\left(\max_{\overline{g}\in\overline{\mathcal{G}}(\gamma,D_{2m})}\left(R_{\gamma/2,\tilde{D}_m}(\overline{g})-R_{\gamma/2,D_m}(\overline{g})\right)\geq\varepsilon-\frac{1}{m}\right)=$$

$$\int_{\mathcal{Z}^{2m}}\frac{1}{|\mathfrak{T}_{2m}|}\sum_{\sigma\in\mathfrak{T}_{2m}}\mathbb{1}\left[\max_{\overline{g}\in\overline{\mathcal{G}}(\gamma,x^{2m})}\left(R_{\gamma/2,\sigma(\tilde{z}^m)}(\overline{g})-R_{\gamma/2,\sigma(z^m)}(\overline{g})\right)\geq\varepsilon-\frac{1}{m}\right]dP^{2m}(z^{2m})=$$

$$\int_{\mathcal{Z}^{2m}}\mathbb{P}_{S_{2m}}\left(\max_{\overline{g}\in\overline{\mathcal{G}}(\gamma,x^{2m})}\left(R_{\gamma/2,S_{2m}(\tilde{z}^m)}(\overline{g})-R_{\gamma/2,S_{2m}(z^m)}(\overline{g})\right)\geq\varepsilon-\frac{1}{m}\right)dP^{2m}(z^{2m})\leq$$

$$\max_{z^{2m}\in\mathcal{Z}^{2m}}\mathbb{P}_{S_{2m}}\left(\max_{\overline{g}\in\overline{\mathcal{G}}(\gamma,x^{2m})}\left(R_{\gamma/2,S_{2m}(\tilde{z}^m)}(\overline{g})-R_{\gamma/2,S_{2m}(z^m)}(\overline{g})\right)\geq\varepsilon-\frac{1}{m}\right).\tag{27}$$

By application of the union bound, the right-hand side of (27) can be bounded from above as follows:

$$\max_{z^{2m}\in\mathcal{Z}^{2m}}\mathbb{P}_{S_{2m}}\left(\max_{\overline{g}\in\overline{\mathcal{G}}(\gamma,x^{2m})}\left(R_{\gamma/2,S_{2m}(\tilde{z}^m)}(\overline{g})-R_{\gamma/2,S_{2m}(z^m)}(\overline{g})\right)\geq\varepsilon-\frac{1}{m}\right)\leq$$

$$\max_{z^{2m}\in\mathcal{Z}^{2m}} \sum_{\overline{g}\in\overline{\mathcal{G}}(\gamma,x^{2m})} \mathbb{P}_{S_{2m}}\left(R_{\gamma/2,S_{2m}(\tilde{z}^m)}(\overline{g}) - R_{\gamma/2,S_{2m}(z^m)}(\overline{g}) \geq \varepsilon - \frac{1}{m}\right).$$

■

Lemma 53 (Hoeffding's inequality, Hoeffding, 1963) *For $n \in \mathbb{N}^*$, let $(T_i)_{1\leq i\leq n}$ be a sequence of n independent random variables with zero means and bounded ranges: $a_i \leq T_i \leq b_i$. Then, for all $\eta \in \mathbb{R}_+^*$,*

$$\mathbb{P}\left(\sum_{i=1}^n T_i \geq \eta\right) \leq \exp\left(\frac{-2\eta^2}{\sum_{i=1}^n (b_i - a_i)^2}\right).$$

Lemma 54 *Let S_{2m} be a random variable described by the uniform distribution on $\mathfrak{T}_{2m}$. For all z^{2m} in $\mathcal{Z}^{2m}$ and for all $\overline{g}$ in $\overline{\mathcal{G}}\left(\gamma, x^{2m}\right)$,*

$$\mathbb{P}_{S_{2m}}\left(R_{\gamma/2,S_{2m}(\tilde{z}^m)}(\overline{g}) - R_{\gamma/2,S_{2m}(z^m)}(\overline{g}) \geq \varepsilon - \frac{1}{m}\right) \leq \exp\left(-\frac{m}{2}\left(\varepsilon - \frac{1}{m}\right)^2\right).$$

Proof To bound uniformly the probabilities $\mathbb{P}_{S_{2m}}\left(R_{\gamma/2,S_{2m}(\tilde{z}^m)}(\overline{g}) - R_{\gamma/2,S_{2m}(z^m)}(\overline{g}) \geq \varepsilon - \frac{1}{m}\right)$, we appeal to the classical law of large numbers. For any function $\overline{g}$ in $\overline{\mathcal{G}}\left(\gamma, x^{2m}\right)$, let $(\xi_i)_{1\leq i\leq m}$ be the sequence of losses $\left(\mathbb{1}_{\left\{\Delta^{\#}\overline{g}_{y_i}(x_i)<\gamma/2\right\}}\right)_{1\leq i\leq m}$ (sequence of losses on z^m) and $\left(\tilde{\xi}_i\right)_{1\leq i\leq m}$ the corresponding sequence of losses on $\tilde{z}^m$. Let $\alpha = (\alpha_i)_{1\leq i\leq m}$ be a Rademacher sequence. The terms of interest can then be rewritten as:

$$\mathbb{P}_{S_{2m}}\left(R_{\gamma/2,S_{2m}(\tilde{z}^m)}(\overline{g}) - R_{\gamma/2,S_{2m}(z^m)}(\overline{g}) \geq \varepsilon - \frac{1}{m}\right) = \mathbb{P}_\alpha\left(\frac{1}{m}\sum_{i=1}^m \alpha_i\left(\tilde{\xi}_i - \xi_i\right) \geq \varepsilon - \frac{1}{m}\right). \tag{28}$$

To bound from above the right-hand side of (28), Hoeffding's inequality (Lemma 53) can be used. Since the random variables $\alpha_i\left(\tilde{\xi}_i - \xi_i\right)$ take their values in $[-1,1]$ (more precisely in $[\![-1,1]\!]$), this gives:

$$\mathbb{P}_\alpha\left(\frac{1}{m}\sum_{i=1}^m \alpha_i\left(\tilde{\xi}_i - \xi_i\right) \geq \varepsilon - \frac{1}{m}\right) \leq \exp\left(-\frac{m}{2}\left(\varepsilon - \frac{1}{m}\right)^2\right).$$

■

Lemma 55 Kroon, 2003, Theorem 68 *Let $(\Omega, \mathcal{B}, \mathbb{P})$ be a probability space, let $K \in \mathbb{R}_+^*$ and let*

$$\{E(\alpha_1, \alpha_2, \delta) :\ 0 < \alpha_1, \alpha_2 \leq K,\ \delta \leq 1\}$$

be a set of events satisfying the following conditions:

1. *for all $0 < \alpha \leq K$ and $0 < \delta \leq 1$, $\mathbb{P}(E(\alpha,\alpha,\delta)) \leq \delta$;*
2. *for all $0 < a < 1$ and $0 < \delta \leq 1$, $\bigcup_{\alpha\in(0,K]} E(\alpha a, \alpha, \delta\alpha(1-a))$ is measurable;*

3. *for all* $0 < \alpha_1 \leq \alpha \leq \alpha_2 \leq K$ *and* $0 < \delta_1 \leq \delta \leq 1$, $E(\alpha_1, \alpha_2, \delta_1) \subseteq E(\alpha, \alpha, \delta)$.

Then for $(a, \delta) \in (0,1) \times (0,1]$,

$$\mathbb{P}\left(\bigcup_{\alpha \in (0,K]} E\left(\alpha a, \alpha, \frac{\delta \alpha (1-a)}{K} \right) \right) \leq \delta.$$

Lemma 56 Kolmogorov and Tihomirov, 1961, Theorem IV *For every pseudo-metric space* (E, ρ), *every totally bounded subset* E' *of* E *and* $\varepsilon \in \mathbb{R}_+^*$,

$$\mathcal{M}(2\varepsilon, E', \rho) \leq \mathcal{N}^{(p)}(\varepsilon, E', \rho) \leq \mathcal{M}(\varepsilon, E', \rho).$$

Lemma 57 *For any class* $\mathcal{G}$ *of functions on* $\mathcal{X}$ *taking their values in* $[-M, M]^Q$ *and for any real number* η *in* $(0, M]$*:*

1. *for every real number* ε *satisfying* $0 < \varepsilon \leq \eta/2$,

$$\textit{SN-dim}\left((\Delta \mathcal{G})^{(\eta)} \right) \leq \textit{N-dim}\left(\Delta \mathcal{G}, \varepsilon \right);$$

2. *for every real number* ε *satisfying* $\varepsilon \geq 3\eta$ *and every* $x^n = (x_i)_{1 \leq i \leq n} \in \mathcal{X}^n$,

$$\mathcal{M}\left(\varepsilon, \Delta^* \mathcal{G}, d_{x^n} \right) \leq \mathcal{M}\left(2, (\Delta^* \mathcal{G})^{(\eta)}, d_{x^n} \right).$$

Proof To prove the first proposition, it is enough to establish that any set strongly N-shattered by $(\Delta \mathcal{G})^{(\eta)}$ is also N-shattered with margin $\eta/2$ by $\Delta \mathcal{G}$. If $s_{\mathcal{X}^n}$, a subset of $\mathcal{X}$ of cardinality n, is strongly N-shattered by $(\Delta \mathcal{G})^{(\eta)}$, then according to Definition 34, there exists a set $I(s_{\mathcal{X}^n})$ of n couples of distinct indexes of categories and a vector v_b in $[\![-\lfloor M/\eta \rfloor + 1, \lfloor M/\eta \rfloor - 1]\!]^n$ such that for every vector $v_y = (y_i) \in \{-1, 1\}^n$, there is a function g_y in $\mathcal{G}$ satisfying

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

Thus, we are looking for a vector $(b_i')_{1 \leq i \leq n}$ such that $\left(\Delta g_{y, i_1(x_i)} \right)^{(\eta)}(x_i) - b_i \geq 1 \Longrightarrow \Delta g_{y, i_1(x_i)}(x_i) - b_i' \geq \eta/2$ and $\left(\Delta g_{y, i_2(x_i)} \right)^{(\eta)}(x_i) + b_i \geq 1 \Longrightarrow \Delta g_{y, i_2(x_i)}(x_i) + b_i' \geq \eta/2$. To that end, four cases must be considered.

1) $b_i \geq 0$ and $y_i = 1$

$$\left(\Delta g_{y, i_1(x_i)} \right)^{(\eta)}(x_i) > 0 \Longrightarrow \eta \left(\Delta g_{y, i_1(x_i)} \right)^{(\eta)}(x_i) \leq \Delta g_{y, i_1(x_i)}(x_i)$$

thus

$$\left(\Delta g_{y, i_1(x_i)} \right)^{(\eta)}(x_i) - b_i \geq 1 \Longrightarrow \Delta g_{y, i_1(x_i)}(x_i) - \eta(b_i + 1/2) \geq \eta/2.$$

2) $b_i \geq 0$ and $y_i = -1$

$$\left(\Delta g_{y, i_2(x_i)} \right)^{(\eta)}(x_i) + b_i \geq 1 \Longrightarrow \Delta g_{y, i_2(x_i)}(x_i) + \eta b_i \geq 0$$

or equivalently

$$\left(\Delta g_{y,i_2(x_i)}\right)^{(\eta)}(x_i)+b_i\geq 1\Longrightarrow \Delta g_{y,i_2(x_i)}(x_i)+\eta(b_i+1/2)\geq \eta/2.$$

3) $b_i<0$ and $y_i=1$

$$\left(\Delta g_{y,i_1(x_i)}\right)^{(\eta)}(x_i)-b_i\geq 1\Longrightarrow \Delta g_{y,i_1(x_i)}(x_i)-\eta b_i\geq 0$$

or equivalently

$$\left(\Delta g_{y,i_1(x_i)}\right)^{(\eta)}(x_i)-b_i\geq 1\Longrightarrow \Delta g_{y,i_1(x_i)}(x_i)-\eta(b_i-1/2)\geq \eta/2.$$

4) $b_i<0$ and $y_i=-1$

$$\left(\Delta g_{y,i_2(x_i)}\right)^{(\eta)}(x_i)>0\Longrightarrow \eta\left(\Delta g_{y,i_2(x_i)}\right)^{(\eta)}(x_i)\leq \Delta g_{y,i_2(x_i)}(x_i)$$

thus

$$\left(\Delta g_{y,i_2(x_i)}\right)^{(\eta)}(x_i)+b_i\geq 1\Longrightarrow \Delta g_{y,i_2(x_i)}(x_i)+\eta(b_i-1/2)\geq \eta/2.$$

To sum up, a satisfactory solution consists in setting $b'_i=\eta(b_i+1/2)$ if $b_i\geq 0$ and $b'_i=\eta(b_i-1/2)$ otherwise. By definition, the set of functions Δg_y, for v_y in $\{-1,1\}^n$, N-shatters $s_{\mathcal{X}^n}$ with margin $\eta/2$, for a set of couples of indexes and a vector of "biases" respectively equal to $I(s_{\mathcal{X}^n})$ and $v_{b'}=(b'_i)_{1\leq i\leq n}$. As a consequence, any set strongly N-shattered by $(\Delta\mathcal{G})^{(\eta)}$ is also N-shattered with margin $\eta/2$ by $\Delta\mathcal{G}$, which is precisely our claim.

To prove the second proposition, let us first notice that:

$$\forall (g,g')\in\mathcal{G}^2,\ \forall x\in\mathcal{X},\ \forall k\in[\![1,Q]\!],\ \forall\eta\in(0,M],$$

$$\left|\Delta^* g_k(x)-\Delta^* g'_k(x)\right|\geq 3\eta\Longrightarrow\left|\left(\Delta^* g_k\right)^{(\eta)}(x)-\left(\Delta^* g'_k\right)^{(\eta)}(x)\right|\geq 2.$$

Indeed, without loss of generality, we can make the hypothesis that $\Delta^* g_k(x)>\Delta^* g'_k(x)$. Then,

$$\left(\left(\Delta^* g'_k\right)^{(\eta)}(x)-1\right)\eta<\Delta^* g'_k(x)<\Delta^* g_k(x)<\left(\left(\Delta^* g_k\right)^{(\eta)}(x)+1\right)\eta.$$

Thus

$$\left(\left(\Delta^* g_k\right)^{(\eta)}(x)+1\right)\eta-\left(\left(\Delta^* g'_k\right)^{(\eta)}(x)-1\right)\eta>3\eta$$

and finally

$$\left(\Delta^* g_k\right)^{(\eta)}(x)-\left(\Delta^* g'_k\right)^{(\eta)}(x)>1,$$

from which the desired result springs directly, keeping in mind that the η-discretizations are integer numbers $\left(\left(\Delta^* g_k\right)^{(\eta)}(x)-\left(\Delta^* g'_k\right)^{(\eta)}(x)>1\Longrightarrow\left(\Delta^* g_k\right)^{(\eta)}(x)-\left(\Delta^* g'_k\right)^{(\eta)}(x)\geq 2\right)$.

Let $s_{\Delta^*\mathcal{G}}$ be a 3η-separated subset of $\Delta^*\mathcal{G}$ in the pseudo-metric d_{x^n}. It results from the definition of the pseudo-metric that:

$$\forall\ \left(\Delta^* g,\Delta^* g'\right)\in s^2_{\Delta^*\mathcal{G}},\ d_{x^n}\left(\Delta^* g,\Delta^* g'\right)\geq 3\eta\Longrightarrow$$

$$\max_{1\leq i\leq n}\left\|\Delta^* g(x_i)-\Delta^* g'(x_i)\right\|_\infty\geq 3\eta\Longrightarrow$$

$$\max_{1\leq i\leq n}\left\|\left(\Delta^* g\right)^{(\eta)}(x_i)-\left(\Delta^* g'\right)^{(\eta)}(x_i)\right\|_\infty \geq 2 \Longrightarrow$$
$$d_{x^n}\left(\left(\Delta^* g_k\right)^{(\eta)},\left(\Delta^* g'_k\right)^{(\eta)}\right)\geq 2.$$

We have thus proved the second proposition. ■

Note that a more interesting second proposition could have resulted from using a different definition of the η-discretization. Indeed, setting $\left(\Delta^\# g_k\right)^{(\eta)}(x)=\left\lfloor\frac{\Delta^\# g_k(x)}{\eta}\right\rfloor$ irrespective of the sign of $\Delta^\# g_k(x)$, one can easily establish that the following proposition, with a dependence between ε and η identical to the one of Alon et al. (1997), holds true: for every $\varepsilon\geq 2\eta$ and every $x^n\in\mathcal{X}^n$, $\mathcal{M}(\varepsilon,\Delta^*\mathcal{G},d_{x^n})\leq \mathcal{M}(2,(\Delta^*\mathcal{G})^{(\eta)},d_{x^n})$. The reason for our choice is to get an additional useful property, namely:

$$\forall\eta\in(0,M],\ \Delta^\# g_l(x)=-\Delta^\# g_k(x)\Longrightarrow\left(\Delta^\# g_l\right)^{(\eta)}(x)=-\left(\Delta^\# g_k\right)^{(\eta)}(x).$$

This property plays a central role in the derivation of our generalized Sauer-Shelah lemma (see for instance the proofs of Lemmas 35 and 37).

Lemma 58 *For all triplet (K_1,K_2,K_3) of positive integers such that $1\leq K_1\leq K_2$ and $K_3\geq 1$, let*

$$\Phi(K_1,K_2,K_3)=\sum_{i=0}^{K_1}\binom{K_2}{i}K_3^i.$$

The following bound is true:

$$\Phi(K_1,K_2,K_3)<\left(\frac{K_2K_3e}{K_1}\right)^{K_1},$$

where e is the base of the Neperian (or natural) logarithm.

Proof $\sum_{i=0}^{K_1}\binom{K_2}{i}K_3^i\leq K_3^{K_1}\sum_{i=0}^{K_1}\binom{K_2}{i}$. By application of Theorem 13.3. in Devroye et al. (1996), $\sum_{i=0}^{K_1}\binom{K_2}{i}$ can be bounded from above by $\left(\frac{K_2e}{K_1}\right)^{K_1}$, which concludes the proof. ■

Appendix B. Proof of Theorem 22

The proof is divided into several steps, following the structure proposed by Dudley (1978) and Pollard (1984, chap. II), structure also described, with variants, in Devroye et al. (1996, Chap. 12), Vapnik (1998, Chap. 4), Anthony and Bartlett (1999), and Schölkopf and Smola (2002, Chap. 5).

B.1 First Symmetrization

The first step is a symmetrization. The idea is to replace the true risk by an estimate computed on a m-sample $\tilde{D}_m$ independent of D_m. This symmetrization corresponds to Lemma 50, and thus gives:

$$\mathbb{P}_{D_m}\left(\sup_{g\in\mathcal{G}}\left(R(g)-R_{\gamma,D_m}(g)\right)>\varepsilon\right)\leq 2\mathbb{P}_{D_{2m}}\left(\sup_{g\in\mathcal{G}}\left(R_{\tilde{D}_m}(g)-R_{\gamma,D_m}(g)\right)\geq\varepsilon-\frac{1}{m}\right). \tag{29}$$

Note that at this point, the standard pathway consists in applying a second symmetrization to get rid of the "ghost sample" $\tilde{D}_m$ (see for example Pollard, 1984; Devroye et al., 1996). For the sake of simplicity, we do not develop this possibility here. Instead, we apply another symmetrization, to keep one single type of empirical measure of accuracy in the bound.

B.2 Second Symmetrization

The second symmetrization, resulting from Lemma 51, corresponds to the following upper bound :

$$\mathbb{P}_{D_{2m}}\left(\sup_{g\in\mathcal{G}}\left(R_{\tilde{D}_m}(g)-R_{\gamma,D_m}(g)\right)\geq\varepsilon-\frac{1}{m}\right)\leq$$

$$\mathbb{P}_{D_{2m}}\left(\max_{\overline{g}\in\overline{\mathcal{G}}(\gamma,D_{2m})}\left(R_{\gamma/2,\tilde{D}_m}(\overline{g})-R_{\gamma/2,D_m}(\overline{g})\right)\geq\varepsilon-\frac{1}{m}\right). \tag{30}$$

It is useful for two reasons. First, it completes, in some sense, the first one, by replacing the two different empirical measures of accuracy appearing in the right-hand side of (29) with two independent copies of the same random variable. Second, it makes it possible to substitute, in the forthcoming computations, the set $\mathcal{G}$ of possibly infinite cardinality with a subset of it of cardinality no more than $\mathcal{N}^{(p)}(\gamma/2,\Delta_\gamma^{\#}\mathcal{G},2m)$. This is exploited in the next step of the proof, to apply a standard union bound.

B.3 Maximal Inequality

To bound from above the right-hand side of (30) irrespective of P, and thus derive a distribution-free result, we introduce an auxiliary step of randomization. Let S_{2m} be a random variable described by the uniform distribution on $\mathfrak{T}_{2m}$. By application of Lemma 52,

$$\mathbb{P}_{D_{2m}}\left(\max_{\overline{g}\in\overline{\mathcal{G}}(\gamma,D_{2m})}\left(R_{\gamma/2,\tilde{D}_m}(\overline{g})-R_{\gamma/2,D_m}(\overline{g})\right)\geq\varepsilon-\frac{1}{m}\right)\leq$$

$$\max_{z^{2m}\in\mathcal{Z}^{2m}}\sum_{\overline{g}\in\overline{\mathcal{G}}(\gamma,x^{2m})}\mathbb{P}_{S_{2m}}\left(R_{\gamma/2,S_{2m}(\tilde{z}^m)}(\overline{g})-R_{\gamma/2,S_{2m}(z^m)}(\overline{g})\geq\varepsilon-\frac{1}{m}\right). \tag{31}$$

B.4 Exponential Bound

Using Lemma 54, the probabilities in the right-hand side of (31) are bounded uniformly by $\exp\left(-\frac{m}{2}\left(\varepsilon-\frac{1}{m}\right)^2\right)$. As a consequence,

$$\max_{z^{2m}\in\mathcal{Z}^{2m}}\sum_{\overline{g}\in\overline{\mathcal{G}}(\gamma,x^{2m})}\mathbb{P}_{S_{2m}}\left(R_{\gamma/2,S_{2m}(\tilde{z}^m)}(\overline{g})-R_{\gamma/2,S_{2m}(z^m)}(\overline{g})\geq\varepsilon-\frac{1}{m}\right)\leq$$

$$\max_{x^{2m}\in\mathcal{X}^{2m}}\left|\overline{\mathcal{G}}\left(\gamma,x^{2m}\right)\right|\exp\left(-\frac{m}{2}\left(\varepsilon-\frac{1}{m}\right)^2\right).$$

According to Definitions 15 and 19, $\max_{x^{2m}\in\mathcal{X}^{2m}}\left|\overline{\mathcal{G}}\left(\gamma,x^{2m}\right)\right|=\mathcal{N}^{(p)}\left(\gamma/2,\Delta_\gamma^{\#}\mathcal{G},2m\right)$, and thus

$$\max_{z^{2m}\in\mathcal{Z}^{2m}}\sum_{\overline{g}\in\overline{\mathcal{G}}(\gamma,x^{2m})}\mathbb{P}_{S_{2m}}\left(R_{\gamma/2,S_{2m}(\tilde{z}^m)}(\overline{g})-R_{\gamma/2,S_{2m}(z^m)}(\overline{g})\geq\varepsilon-\frac{1}{m}\right)\leq$$

$$\mathcal{N}^{(p)}\left(\gamma/2,\Delta_\gamma^{\#}\mathcal{G},2m\right)\exp\left(-\frac{m}{2}\left(\varepsilon-\frac{1}{m}\right)^2\right). \tag{32}$$

The combination of (29), (30), (31), and (32) provides us with the following bound:

$$\mathbb{P}_{D_m}\left(\sup_{g\in\mathcal{G}}\left(R(g)-R_{\gamma,D_m}(g)\right)>\varepsilon\right)\leq 2\mathcal{N}^{(p)}\left(\gamma/2,\Delta_\gamma^{\#}\mathcal{G},2m\right)\exp\left(-\frac{m}{2}\left(\varepsilon-\frac{1}{m}\right)^2\right). \tag{33}$$

Setting the right-hand side of (33) to δ and solving for ε finally gives:

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

B.5 Uniform Bound Over the Margin Parameter γ

This last bound holds for a value of γ specified in advance. To make the bound useful, we would like to be able to select γ after observation of the trained machine on the training set. This can be done thanks to Lemma 55, extending Proposition 8 in Bartlett (1998), which allows us to produce a result that stands uniformly for all values of the margin parameter γ in the interval $(0,\Gamma]$. To apply Lemma 55 to the case of interest, let us define the function Θ as follows:

$$\Theta(t,u)=\sqrt{\frac{2}{m}\left(\ln\left(2\mathcal{N}^{(p)}\left(t,\Delta_\gamma^{\#}\mathcal{G},2m\right)\right)-\ln(u)\right)}.$$

One can readily verify that the measure $\mathbb{P}_{D_m}$ and the set of events $E(\alpha_1,\alpha_2,\delta)$ given by:

$$\sup_{g\in\mathcal{G}}\left(R(g)-R_{\alpha_2,D_m}(g)\right)\geq\Theta\left(\frac{\alpha_1}{2},\delta\right)+\frac{1}{m}$$

satisfy the hypotheses of Lemma 55. Its application gives, for all choice of the couple (a,δ) in $(0,1)\times(0,1]$,

$$\mathbb{P}_{D_m}\left[\bigcup_{\alpha\in(0,K]}\left(\sup_{g\in\mathcal{G}}\left(R(g)-R_{\alpha,D_m}(g)\right)\geq\Theta\left(\frac{\alpha a}{2},\frac{\delta\alpha(1-a)}{K}\right)+\frac{1}{m}\right)\right]\leq\delta.$$

Setting $\alpha=\gamma$, $K=\Gamma$ and choosing $a=1/2$ yields to:

$$\mathbb{P}_{D_m}\left[\bigcup_{\gamma\in(0,\Gamma]}\left(\sup_{g\in\mathcal{G}}\left(R(g)-R_{\gamma,D_m}(g)\right)\geq\Theta\left(\frac{\gamma}{4},\frac{\gamma\delta}{2\Gamma}\right)+\frac{1}{m}\right)\right]\leq\delta$$

and finally, by definition of Θ,

$$\mathbb{P}_{D_m}\left[\bigcup_{\gamma\in(0,\Gamma]}\left(\sup_{g\in\mathcal{G}}\left(R(g)-R_{\gamma,D_m}(g)\right)\geq\right.\right.$$

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

which concludes the proof of Theorem 22.

References

E.L. Allwein, R.E. Schapire, and Y. Singer. Reducing multiclass to binary: A unifying approach for margin classifiers. *Journal of Machine Learning Research*, 1:113–141, 2000.

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.

A. Ambroladze, E. Parrado-Hernandez, and J. Shawe-Taylor. Tighter PAC-Bayes bounds. In *Advances in Neural Information Processing Systems 19*, 2007. (to appear).

M. Anthony and P.L. Bartlett. *Neural Network Learning: Theoretical Foundations*. Cambridge University Press, Cambridge, 1999.

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

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.

P.L. Bartlett, P.M. Long, and R.C. Williamson. Fat-shattering and the learnability of real-valued functions. *Journal of Computer and System Sciences*, 52(3):434–452, 1996.

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.

G. Blanchard, O. Bousquet, and P. Massart. Statistical performance of support vector machines. *The Annals of Statistics*, 2007. (to appear).

B. Boser, I. Guyon, and V.N. Vapnik. A training algorithm for optimal margin classifiers. In *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*, pages 144–152, 1992.

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

O. Bousquet. *Concentration Inequalities and Empirical Processes Theory Applied to the Analysis of Learning Algorithms*. PhD thesis, Ecole Polytechnique, 2002.

B. Carl and I. Stephani. *Entropy, Compactness and the Approximation of Operators*. Cambridge University Press, Cambridge, 1990.

O. Chapelle, V.N. Vapnik, O. Bousquet, and S. Mukherjee. Choosing multiple parameters for support vector machines. *Machine Learning*, 46(1):131–159, 2002.

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

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

K. Crammer and Y. Singer. On the learnability and design of output codes for multiclass problems. *Machine Learning*, 47(2):201–233, 2002.

Y. Darcy and Y. Guermeur. Radius-margin bound on the leave-one-out error of multi-class SVMs. Technical Report RR-5780, INRIA, 2005.

L. Devroye, L. Györfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition*. Springer-Verlag, New York, 1996.

R.M. Dudley. Central limit theorems for empirical measures. *The Annals of Probability*, 6(6): 899–929, 1978.

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.

R.M. Dudley. Universal Donsker classes and metric entropy. *The Annals of Probability*, 15(4): 1306–1326, 1987.

A. Elisseeff, Y. Guermeur, and H. Paugam-Moisy. Margin error and generalization capabilities of multi-class discriminant models. Technical Report NC-TR-99-051-R, NeuroCOLT2, 1999. (revised in 2001).

J. Fürnkranz. Round robin classification. *Journal of Machine Learning Research*, 2:721–747, 2002.

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

Y. Guermeur, A. Elisseeff, and H. Paugam-Moisy. Estimating the sample complexity of a multi-class discriminant model. In *International Conference on Artificial Neural Networks*, pages 310–315. IEE, 1999.

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.

L. Gurvits. A note on a scale-sensitive dimension of linear bounded functionals in Banach spaces. *Theoretical Computer Science*, 261(1):81–90, 2001.

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.

D. Haussler and P.M. Long. A generalization of Sauer's lemma. *Journal of Combinatorial Theory, Series A*, 71(2):219–240, 1995.

W. Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58:13–30, 1963.

K. Jogdeo and S.M. Samuels. Monotone convergence of binomial probabilities and a generalization of Ramanujan's equation. *The Annals of Mathematical Statistics*, 39(4):1191–1195, 1968.

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.

A.N. Kolmogorov and V.M. Tihomirov. ε-entropy and ε-capacity of sets in functional spaces. *American Mathematical Society Translations, series 2*, 17:277–364, 1961.

R.S. Kroon. Support vector machines, generalization bounds, and transduction. Master's thesis, University of Stellenbosch, South Africa, December 2003. `http://www.cs.sun.ac.za/~skroon/personal/pubs/kroon2003support.ps`.

M. Ledoux. On Talagrand's deviation inequalities for product measures. *ESAIM: Probability and Statistics*, 1:63–87, 1996.

Y. Lee and Z. Cui. Characterizing the solution path of multicategory support vector machines. *Statistica Sinica*, 16:391–409, 2006.

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.

G. Lugosi. Concentration-of-measure inequalities. Lecture notes, Summer School on Machine Learning at the Australian National University, Canberra, 2004.

P. Massart. Some applications of concentration inequalities to statistics. *Annales de la Faculté des Sciences de Toulouse*, 9(2):245–303, 2000.

E. Monfrini and Y. Guermeur. A quadratic loss multi-class SVM. Technical report, LORIA, 2007. (to appear).

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

J.C. Platt, N. Cristianini, and J. Shawe-Taylor. Large margin DAGs for multiclass classification. In *Advances in Neural Information Processing Systems 12*, pages 547–553, 2000.

D. Pollard. *Convergence of Stochastic Processes*. Springer-Verlag, New York, 1984.

R. Rifkin and A. Klautau. In defense of one-vs-all classification. *Journal of Machine Learning Research*, 5:101–141, 2004.

W. Rudin. *Real and Complex Analysis*. McGraw-Hill, New York, third edition, 1987.

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

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

J. Shawe-Taylor, P.L. Bartlett, R.C. Williamson, and M. Anthony. Structural risk minimization over data-dependent hierarchies. *IEEE Transactions on Information Theory*, 44(5):1926–1940, 1998.

S. Shelah. A combinatorial problem: Stability and order for models and theories in infinitary languages. *Pacific Journal of Mathematics*, 41(1):247–261, 1972.

I. Steinwart and C. Scovel. Fast rates for support vector machines. In *Proceedings of the eighteenth annual Conference on Learning Theory*, pages 279–294, 2005.

M. Stone. Asymptotics for and against cross-validation. *Biometrika*, 64(1):29–35, 1977.

M. Talagrand. Concentration of measure and isoperimetric inequalities in product spaces. *Publications mathématiques de l'I.H.E.S.*, 81:73–205, 1995.

M. Talagrand. A new look at independence. *The Annals of Probability*, 24(1):1–34, 1996.

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

L.G. Valiant. A theory of the learnable. *Communications of the ACM*, 27(11):1134–1142, 1984.

A.W. van der Vaart and J.A. Wellner. *Weak Convergence and Empirical Processes - With Applications to Statistics*. Springer Series in Statistics. Springer-Verlag, New York, 1996.

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

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

V.N. Vapnik and A.Ya. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and its Applications*, XVI(2):264–280, 1971.

G. Wahba. Support vector machines, reproducing kernel Hilbert spaces, and randomized GACV. In B. Schölkopf, C.J.C. Burges, and A.J. Smola, editors, *Advances in Kernel Methods - Support Vector Learning*, chapter 6, pages 69–88. The MIT Press, Cambridge, MA, 1999.

L. Wang, P. Xue, and K.L. Chan. Generalized radius-margin bounds for model selection in multiclass SVMs. Technical report, School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore, 639798, 2005.

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.

R.C. Williamson, A.J. Smola, and B. Schölkopf. Entropy numbers of linear function classes. In *Proceedings of the Thirteenth Annual Workshop on Computational Learning Theory*, pages 309–319, 2000.

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