INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Large Margin Multi-category Discriminant Models and Scale-sensitive Ψ-dimensions

Yann Guermeur

N° 5314 – version 2

version initiale September 24, 2004 – version révisée September 12, 2006

Thème BIO

Rapport de recherche

ISSN 0249-6399

Large Margin Multi-category Discriminant Models and Scale-sensitive Ψ-dimensions

Yann Guermeur*

Thème BIO — Systèmes biologiques
Projet MODBIO

Rapport de recherche n° 5314 – version 2† — version initiale September 24, 2004 — version révisée September 12, 2006 49 pages

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 polychotomies with functions taking their values in finite sets. The case of classes of vector-valued functions used to compute polychotomies 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, it represents the situation most commonly met in practice.

In this report, a new uniform convergence bound for large margin multi-class discriminant models is derived, which extends in a straightforward way a famous theorem by Bartlett. The capacity measure involved in this bound is a covering number. To bound from above this measure, original scale-sensitive extensions of the Ψ-dimensions are introduced. The covering numbers of interest can be bounded in terms of these 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 architecture of the multi-class SVMs.

Key-words: Multi-class discriminant analysis, uniform strong laws of large numbers, generalized VC dimensions, structural risk minimization inductive principle, multi-class SVMs

* UMR 7503 - CNRS

† The definition of the margin Ψ-dimensions has been modified to establish the connection between the generalized Sauer-Shelah lemma and the bound on the scale-sensitive Natarajan dimension of the M-SVMs.

Unité de recherche INRIA Lorraine
LORIA, Technopôle de Nancy-Brabois, Campus scientifique,
615, rue du Jardin Botanique, BP 101, 54602 Villers-Lès-Nancy (France)
Téléphone : +33 3 83 59 30 00 — Télécopie : +33 3 83 27 83 19

Systèmes discriminants multi-classes à grande marge et Ψ-dimensions paramétrées

Résumé : En discrimination, la théorie statistique de l'apprentissage proposée par Vapnik a principalement été développée suivant trois axes : celui du calcul des dichotomies par des fonctions à valeurs binaires, celui du calcul des dichotomies par des fonctions à valeurs réelles et celui du calcul des polychotomies par des fonctions prenant leurs valeurs dans des ensembles finis. Le cas des familles de fonctions à valeurs vectorielles utilisées pour calculer des polychotomies a rarement été considéré de manière indépendante, ce qui représente un manque important, pour trois raisons principales. Tout d'abord, ce dernier cas englobe les précédents, ensuite, il ne peut être traité de manière satisfaisante par une extension naïve des résultats dédiés au calcul des dichotomies, enfin, il constitue la situation la plus fréquemment rencontrée en pratique.

Dans ce rapport, nous dérivons une nouvelle borne de convergence uniforme pour les modèles de discrimination à grande marge dans le cas multi-classe. Elle étend de manière directe un célèbre théorème de Bartlett. La mesure de capacité apparaissant dans cette borne est un nombre de couverture. Afin de majorer cette mesure, une extension paramétrée des Ψ-dimensions est introduite. Le lien entre ces deux types de notions de complexité est établi par le biais de lemmes de Sauer généralisés. Une illustration en est donnée dans le cas spécifique de la dimension de Natarajan à marge. Une borne sur cette dernière dimension est ensuite calculée pour l'architecture commune à toutes les SVM multi-classes.

Mots-clés : Analyse discriminante à catégories multiples, lois fortes des grands nombres uniformes, dimensions VC généralisées, principe inductif de minimisation structurelle du risque, SVM multi-classes

1 Introduction

One of the central domains of Vapnik's statistical learning theory [73] is the theory of bounds, which is at the origin of the structural risk minimization (SRM) inductive principle [71, 64] and, as such, has not only a theoretical interest, but also a practical one. This theory has been developed for discriminant analysis, regression and density estimation. The first results in the field of discrimination, exposed in [74], were dealing with the computation of dichotomies with binary-valued functions. Later on, several studies were devoted to the case of multi-class $\{1, \dots, Q\}$-valued classifers [11], and large margin classifiers computing dichotomies [3, 8, 10] (see also [9] for the case of regression). However, the case of large margin classifiers computing polychotomies (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 [31].

In this report, we extend some of our previous works on the statistical theory of large margin multi-class discriminant systems, reported for instance in [27, 30, 33]. The main idea is to unify two complementary and well established theories: the theory of large margin (bi-class) classifiers and the theory of multi-class $\{1, \dots, Q\}$-valued classifers. To that end, we first introduce a new extension of Bartlett's famous theorem on the sample complexity of large margin classifiers [8]. Then, we extend the notion of Ψ-dimensions, central in the context of multi-class discriminant analysis, by making it scale-sensitive, on the model of the fat-shattering dimension. These new capacity measures can be used to bound from above the covering numbers appearing in the confidence interval of the uniform convergence result. The corresponding generalization of Sauer's lemma is established in the particular case of the margin Natarajan dimension. This dimension is then upper bounded for the architecture shared by all the multi-class SVMs (M-SVMs) proposed so far, which makes it possible to justify a posteriori the choice of the control term appearing in their objective functions.

The organization of the paper is as follows. Section 2 introduces the notion of margin risk for multi-class discriminant models, as well as the capacity measure that will appear in the confidence interval of the guaranteed risk. Section 3 is devoted to the formulation of our new uniform convergence result and its proof. Scale-sensitive extensions of the Ψ-dimensions are introduced in Section 4. The extension of Sauer's lemma relating the covering numbers of interest to the margin Natarajan dimension is established in Section 5. Section 6 is devoted to the computation of a bound on the margin Natarajan dimension of the architecture shared by all the M-SVMs. Section 7 deals with the synthesis which can be done of the results derived in the preceding sections. It specifically addresses the question of the specificities of the multi-class case, and the comparison which can be done between the different M-SVMs. At last, we draw conclusions and outline our future work in Section 8.

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 $Q \geq 3$ to exclude the degenerate case of dichotomies. Let $\mathcal{X}$ be the space of description and $\mathcal{Y} = \{1, \dots, k, \dots, Q\}$ the set of categories. We make the assumption, standard in statistical learning theory, that there is a joint probability distribution F, fixed but unknown, on $\mathcal{X} \times \mathcal{Y}$. This distribution utterly characterizes the problem of interest. Our goal is to find, in a given set $\mathcal{H}$ of functions $h = (h_k)_{1 \leq k \leq Q}$ from $\mathcal{X}$ into $\mathbb{R}^Q$, a function with the lowest "error rate" on this problem. The "error rate" of a function h is the error rate or *expected risk* of the corresponding discrimination function f, from $\mathcal{X}$ into $\mathcal{Y}$, obtained by assigning each pattern x to the category k satisfying: $h_k(x) = \max_l h_l(x)$. The patterns for which this assignation is ambiguous are assigned to a dummy category, so that they contribute to the computation of the different risks considered below. f must thus be as close as possible to Bayes' decision rule. In the common case where the outputs of the function selected are estimates of the class posterior probabilities, which happens for instance when $\mathcal{H}$ is the set of functions computed by a multi-layer perceptron and the training criterion has been adequately chosen (see for instance [57]), applying this decision function is especially natural since it simply amounts to implementing Bayes' estimated decision rule. The class $\mathcal{H}$ is supposed to satisfy some mild measurability conditions which will appear implicitly in the sequel. A suitable such condition could for instance result from slightly adapting the "image admissible Suslin" property (see for instance [26], Section 5.3 or [29]). Hereafter, $\mathcal{S}$ will designate the product space $\mathcal{X} \times \mathcal{Y}$.

2.2 Multi-class margin risk

The uniform convergence result established in the following section is based on an extended notion of risk. The standard risk is simply the probability of error. Formally, it is thus defined as follows:

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

$$R(f) = \mathbb{P}\left\{(x, y) : f(x) \neq y\right\} = \int_{\mathcal{X} \times \mathcal{Y}} \mathbb{1}_{\{f(x) \neq k\}} dF(x, k) \tag{1}$$

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

The empirical risk is the frequency of error measured on a sample:

Definition 2 (Empirical risk) *Let $s_m = \{(x_i, y_i) : 1 \leq i \leq m\}$ be a m-sample of examples independently drawn from F. The* empirical risk *of f on s_m is defined as:*

$$R_{s_m}(f) = \frac{1}{m} \# \{(x_i, y_i) \in s_m : f(x_i) \neq y_i\}, \tag{2}$$

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

As stated above, the expected risk (resp. empirical risk) of a function h from $\mathcal{X}$ into $\mathbb{R}^Q$ is the expected risk (resp. empirical risk) of the corresponding discriminant function f. For such functions, the element that will appear central to measure the quality of the discrimination is the value of a multi-class margin. This notion of margin has been studied independently by different groups of authors (see for instance [27, 2]).

Definition 3 (Multi-class margin) *Let h be a function from $\mathcal{X}$ into $\mathbb{R}^Q$ and (x, y) an element of $\mathcal{X} \times \mathcal{Y}$. Then the* margin *of h on (x, y), $M(h, x, y)$, is given by:*

$$M(h, x, y) = \frac{1}{2} \left\{ h_y(x) - \max_{k \neq y} h_k(x) \right\}. \tag{3}$$

To take this margin into account, the following operator is introduced:

Definition 4 (Δ^* operator) *Define Δ^* as an operator on $\mathcal{H}$ such that:*

$$\begin{aligned} \Delta^* : \quad & \mathcal{H} \longrightarrow \Delta^* \mathcal{H} \\ & h = (h_k) \mapsto \Delta^* h = (\Delta^* h_k) \end{aligned}$$

$\forall (h, x) \in \mathcal{H} \times \mathcal{X}$, let $M(h, x) = \frac{1}{2} \max_k \{h_k(x) - \max_{l \neq k} h_l(x)\}$.

$$\forall k \in \{1, \dots, Q\}, \ \Delta^* h_k(x) = \begin{cases} M(h, x) & if \ \frac{1}{2} \{h_k(x) - \max_{l \neq k} h_l(x)\} = M(h, x) \\ -M(h, x) & otherwise \end{cases}. \tag{4}$$

Note that the introduction of the 1/2 coefficient makes Δ^* a projection operator, i.e. an operator satisfying $\Delta^{*2} = \Delta^*$. With this definition at hand, we define the margin risk as follows.

Definition 5 (Margin risk) *Let $\mathcal{H}$ be a class of functions from $\mathcal{X}$ into $\mathbb{R}^Q$ and $\gamma \in \mathbb{R}_+^*$. The* risk with margin γ *of a function h of $\mathcal{H}$ is defined as:*

$$R_\gamma(h) = \mathbb{P} \{(x, y) : \Delta^* h_y(x) < \gamma\} = \int_{\mathcal{S}} \mathbb{1}_{\{\Delta^* h_k(x) < \gamma\}} dF(x, k). \tag{5}$$

The empirical margin risk is defined accordingly.

Definition 6 (Empirical margin risk) *The* empirical risk with margin $\gamma \in \mathbb{R}_+^*$ *of h on a m-sample s_m is*

$$R_{\gamma, s_m}(h) = \frac{1}{m} \# \{(x_i, y_i) \in s_m : \Delta^* h_{y_i}(x_i) < \gamma\}. \tag{6}$$

The control term that will be added to this empirical margin risk to bound from above, with high probability, the expected risk, involves covering numbers as capacity measure. Their definition is the subject of the following subsection. Introductions to the basic notions of functional analysis used in this document can be found in [18, 19, 23, 70].

2.3 Capacity measure: covering numbers

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

Definition 7 (ϵ-cover or ϵ-net) *Let (E, ρ) be a pseudo-metric space, and $B(e, r)$ the open ball of center e and radius r in E. Let E' be a subset of E. For $\epsilon \in \mathbb{R}_+^*$, an* ϵ-cover *of E' is a subset $\overline{E'}$ of E such that:*

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

$\overline{E'}$ is a proper ϵ-cover *of E' if it is included in E'.*

Definition 8 (Covering numbers) *Let (E, ρ) be a pseudo-metric space. If $E' \subset E$ has an ϵ-cover of finite cardinality, then its* covering number $\mathcal{N}(\epsilon, E', \rho)$ *is the smallest cardinality of its ϵ-covers. If there is no such finite cover, then the covering number is defined to be ∞. We denote $\mathcal{N}^{(p)}(\epsilon, E', \rho)$ the covering number obtained by considering proper ϵ-covers only.*

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

Definition 9 ($d_{\ell_\infty, \ell_\infty(s_{\mathcal{X}^m})}$ pseudo-metric) *Let $\mathcal{H}$ be a set of functions from $\mathcal{X}$ into $\mathbb{R}^Q$. For an element $s_{\mathcal{X}^m}$ of $\mathcal{X}^m$, define the pseudo-metric $d_{\ell_\infty, \ell_\infty(s_{\mathcal{X}^m})}$ on $\mathcal{H}$ as:*

$$\forall (h, h') \in \mathcal{H}^2, \; d_{\ell_\infty, \ell_\infty(s_{\mathcal{X}^m})}(h, h') = \max_{x \in s_{\mathcal{X}^m}} \|h(x) - h'(x)\|_\infty . \tag{7}$$

For technical reasons, linked in particular to the computation of the upper bound on the covering numbers (more precisely the generalized Sauer-Shelah lemma), it is useful to bound the component functions of $\Delta^* h$ in the interval $[-\gamma, \gamma]$, where γ is the parameter of the margin risk. This is achieved by application of the π_γ operator [8].

Definition 10 (π_γ operator) *Let $\mathcal{H}$ be a class of functions from $\mathcal{X}$ into $\mathbb{R}^Q$. For $\gamma \in \mathbb{R}_+^*$, let $\pi_\gamma : h = (h_k) \mapsto \pi_\gamma(h) = (\pi_\gamma(h_k))$ be the piecewise-linear squashing operator defined as:*

$$\forall h \in \mathcal{H}, \; \forall x \in \mathcal{X}, \; \forall k \in \{1, \ldots, Q\}, \; \pi_\gamma(h_k)(x) = \begin{cases} \gamma . sign\left(h_k(x)\right) & if \; |h_k(x)| \geq \gamma \\ h_k(x) & otherwise \end{cases} . \tag{8}$$

Note that π_γ is also a projection operator. In the sequel, Δ_γ^* will designate $\pi_\gamma \circ \Delta^*$, once more a projection operator. Furthermore, $\Delta_\gamma^* \mathcal{H}$ will represent the set of functions $\left\{\Delta_\gamma^* h : h \in \mathcal{H}\right\}$. $\mathcal{N}_{\infty,\infty}^{(p)}(\epsilon, \Delta_\gamma^* \mathcal{H}, m)$ will denote $\sup_{s_{\mathcal{X}^m} \in \mathcal{X}^m} \mathcal{N}^{(p)}(\epsilon, \Delta_\gamma^* \mathcal{H}, d_{\ell_\infty, \ell_\infty(s_{\mathcal{X}^m})})$.

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.

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

$$R(h) \leq R_{\gamma,s_m}(h) + \sqrt{\frac{2}{m}\left(\ln\left(2\mathcal{N}^{(p)}_{\infty,\infty}(\gamma/4, \Delta^*_\gamma\mathcal{H}, 2m)\right) + \ln\left(\frac{2}{\gamma\delta}\right)\right)} + \frac{1}{m}. \tag{9}$$

This theorem can be seen as an extension of Corollary 9 in [8], Theorem 4.1 in [73], and more generally an extension of the Glivenko-Cantelli theorem (see for instance [55, 23, 73, 69]). Its proof is divided into several steps, following the structure proposed in [24, 55, 62].

3.1 First symmetrization

In this first step of the proof, standard techniques are used to replace the problem of matching the empirical measure R_{γ,s_m} against the distribution R with the problem of matching R_{γ,s_m} against an independent empirical measure, $R_{\tilde{s}_m}$, the parent of which is R. Precisely, taking our inspiration from the proof of Vapnik's basic lemma in [73] (Section 4.5.1), we prove the following result:

Lemma 1 *The distribution of the random variable $\sup_{h\in\mathcal{H}}(R(h) - R_{\gamma,s_m}(h))$ is connected with the distribution of the random variable $\sup_{h\in\mathcal{H}}(R_{\tilde{s}_m}(h) - R_{\gamma,s_m}(h))$ by the inequality*

$$\mathbb{P}_{s_m}\left(\sup_{h\in\mathcal{H}}(R(h) - R_{\gamma,s_m}(h)) > \epsilon\right) \leq 2\mathbb{P}_{s_m,\tilde{s}_m}\left(\sup_{h\in\mathcal{H}}(R_{\tilde{s}_m}(h) - R_{\gamma,s_m}(h)) \geq \epsilon - \frac{1}{m}\right) \tag{10}$$

where $\tilde{s}_m$ is a m-sample independent of s_m, $\mathbb{P}_{s_m}$ is a probability over the sample s_m, and $\mathbb{P}_{s_m,\tilde{s}_m}$ is a probability over $s_{2m} = s_m \cup \tilde{s}_m$.

Proof By definition:

$$\mathbb{P}_{s_m,\tilde{s}_m}\left(\sup_{h\in\mathcal{H}}(R_{\tilde{s}_m}(h) - R_{\gamma,s_m}(h)) \geq \epsilon - \frac{1}{m}\right) =$$

$$\int_{\mathcal{S}^{2m}} \mathbb{1}\left[\sup_{h\in\mathcal{H}}(R_{\tilde{s}_m}(h) - R_{\gamma,s_m}(h)) \geq \epsilon - \frac{1}{m}\right] dF(s_m, \tilde{s}_m).$$

Since s_m and $\tilde{s}_m$ are supposed to be independent, $\mathcal{S}^{2m}$ is the direct product of the space to which s_m belongs and the space to which $\tilde{s}_m$ belongs. One can thus apply Fubini's theorem for nonnegative measurable functions [28, 7] to the product measure $\mathbb{P}_{s_m,\tilde{s}_m}$, which gives:

$$\mathbb{P}_{s_m,\tilde{s}_m}\left(\sup_{h\in\mathcal{H}}\left(R_{\tilde{s}_m}(h) - R_{\gamma,s_m}(h)\right) \geq \epsilon - \frac{1}{m}\right) =$$

$$\int_{\mathcal{S}^m} dF(s_m) \int_{\mathcal{S}^m} \mathbb{1}\left[\sup_{h\in\mathcal{H}}\left(R_{\tilde{s}_m}(h) - R_{\gamma,s_m}(h)\right) \geq \epsilon - \frac{1}{m}\right] dF(\tilde{s}_m).$$

In the integral over $\tilde{s}_m$, the set s_m is fixed. Let $\mathcal{Q}$ denote the following event in the space $\mathcal{S}^m$:

$$\mathcal{Q} = \left\{s_m \in \mathcal{S}^m : \sup_{h\in\mathcal{H}}\left(R(h) - R_{\gamma,s_m}(h)\right) > \epsilon\right\}.$$

Restricting the integration domain to $\mathcal{Q}$ gives

$$\mathbb{P}_{s_m,\tilde{s}_m}\left(\sup_{h\in\mathcal{H}}\left(R_{\tilde{s}_m}(h) - R_{\gamma,s_m}(h)\right) \geq \epsilon - \frac{1}{m}\right) \geq$$

$$\int_{\mathcal{Q}} dF(s_m) \underbrace{\int_{\mathcal{S}^m} \mathbb{1}\left[\sup_{h\in\mathcal{H}}\left(R_{\tilde{s}_m}(h) - R_{\gamma,s_m}(h)\right) \geq \epsilon - \frac{1}{m}\right] dF(\tilde{s}_m)}_{I}. \tag{11}$$

I is an integral which is calculated for a fixed s_m satisfying

$$\sup_{h\in\mathcal{H}}\left(R(h) - R_{\gamma,s_m}(h)\right) > \epsilon.$$

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

$$R(h^*) - R_{\gamma,s_m}(h^*) \geq \epsilon.$$

By definition of h^*, the following inequality holds

$$I \geq \int_{\mathcal{S}^m} \mathbb{1}\left[R_{\tilde{s}_m}(h^*) - R_{\gamma,s_m}(h^*) \geq \epsilon - \frac{1}{m}\right] dF(\tilde{s}_m).$$

$$\begin{cases} R(h^*) - R_{\gamma,s_m}(h^*) \geq \epsilon \\ R_{\tilde{s}_m}(h^*) - R(h^*) \geq -\frac{1}{m} \end{cases} \Longrightarrow R_{\tilde{s}_m}(h^*) - R_{\gamma,s_m}(h^*) \geq \epsilon - \frac{1}{m}.$$

As a consequence

$$I \geq \int_{\mathcal{S}^m} \mathbb{1}\left[R_{\tilde{s}_m}(h^*) - R(h^*) \geq -\frac{1}{m}\right] dF(\tilde{s}_m).$$

Furthermore

$$\int_{\mathcal{S}^m} \mathbb{1}\left[R_{\tilde{s}_m}(h^*) - R(h^*) \geq -\frac{1}{m}\right] dF(\tilde{s}_m) = \mathbb{P}_{\tilde{s}_m}\left(mR_{\tilde{s}_m}(h^*) \geq mR(h^*) - 1\right). \tag{12}$$

By definition of $R(h^*)$ and $R_{\tilde{s}_m}(h^*)$, $mR_{\tilde{s}_m}(h^*)$ has a binomial distribution with parameters m and $R(h^*)$ ($mR_{\tilde{s}_m}(h^*) \hookrightarrow \mathcal{B}(m, R(h^*))$). To bound from below the right-hand side of (12), we make use of a result on the median of random variables following a binomial distribution.

Lemma 2 *Let X be a random variable described by a binomial distribution with parameters n and p ($X \hookrightarrow \mathcal{B}(n,p)$). Then its median is either $\lfloor np \rfloor$ or $\lfloor np \rfloor + 1$. Moreover, if np is an integer, the median is simply np.*

The proof of this result can for instance be found in [40] (see also Appendix B in [49]). It springs from Lemma 2 that $mR(h^*) - 1$ is inferior or equal to the median of $mR_{\tilde{s}_m}(h^*)$, and thus, by definition of the median, that the right-hand side of (12) is superior or equal to $1/2$. By way of consequence, I is also greater that $1/2$. Substituting this lower bound of I into (11) yields

$$\mathbb{P}_{s_m, \tilde{s}_m}\left(\sup_{h \in \mathcal{H}} \left(R_{\tilde{s}_m}(h) - R_{\gamma, s_m}(h) \right) \geq \epsilon - \frac{1}{m} \right) \geq \frac{1}{2} \int_{\mathcal{Q}} dF(s_m)$$

or equivalently, by definition of $\mathcal{Q}$:

$$\mathbb{P}_{s_m, \tilde{s}_m}\left(\sup_{h \in \mathcal{H}} \left(R_{\tilde{s}_m}(h) - R_{\gamma, s_m}(h) \right) \geq \epsilon - \frac{1}{m} \right) \geq \frac{1}{2} \mathbb{P}_{s_m}\left(\sup_{h \in \mathcal{H}} \left(R(h) - R_{\gamma, s_m}(h) \right) > \epsilon \right)$$

which is the result announced. ∎

Note that at this point, the standard pathway consists in applying a second symmetrization to get rid of the "ghost sample" $\tilde{s}_m$ (see for example [55, 23]). 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 in the bound.

3.2 Second symmetrization

Let $s_n = \{(x_i, y_i) : 1 \leq i \leq n\}$ be a n-sample of examples i.i.d. according to the probability distribution function F. For the sake of simplicity, in what follows, we will make the slight abuse of notation consisting in using $d_{\ell_\infty, \ell_\infty(s_n)}$ in place of $d_{\ell_\infty, \ell_\infty(\{x_i : 1 \leq i \leq n\})}$.

Lemma 3 *Let $s_{2m} = (s_m, \tilde{s}_m) \in \mathcal{S}^{2m}$ be a $2m$-sample and $\overline{\Delta_\gamma^* \mathcal{H}}(s_{2m})$ a proper $\gamma/2$-cover of the set $\Delta_\gamma^* \mathcal{H}$ with respect to the pseudo-metric $d_{\ell_\infty, \ell_\infty(s_{2m})}$. This cover is supposed to be of minimal cardinality, i.e. $\# \overline{\Delta_\gamma^* \mathcal{H}}(s_{2m}) = \mathcal{N}^{(p)}(\gamma/2, \Delta_\gamma^* \mathcal{H}, d_{\ell_\infty, \ell_\infty(s_{2m})})$. Let $\overline{\mathcal{H}}(s_{2m})$ be a subset of $\mathcal{H}$ of cardinality $\mathcal{N}^{(p)}(\gamma/2, \Delta_\gamma^* \mathcal{H}, d_{\ell_\infty, \ell_\infty(s_{2m})})$ the image of which by Δ_γ^* is precisely $\overline{\Delta_\gamma^* \mathcal{H}}(s_{2m})$. To put it in another way, $\left\{ \Delta_\gamma^* \overline{h} : \overline{h} \in \overline{\mathcal{H}}(s_{2m}) \right\} = \overline{\Delta_\gamma^* \mathcal{H}}(s_{2m})$, i.e. there is a one-to-one map between the elements of $\overline{\mathcal{H}}(s_{2m})$ and $\overline{\Delta_\gamma^* \mathcal{H}}(s_{2m})$. Then*

$$\mathbb{P}_{s_m, \tilde{s}_m}\left(\sup_{h \in \mathcal{H}} \left(R_{\tilde{s}_m}(h) - R_{\gamma, s_m}(h) \right) \geq \epsilon - \frac{1}{m} \right) \leq$$

$$\mathbb{P}_{s_m,\tilde{s}_m}\left(\sup_{\overline{h}\in\overline{\mathcal{H}}(s_{2m})}\left(R_{\gamma/2,\tilde{s}_m}\left(\overline{h}\right)-R_{\gamma/2,s_m}\left(\overline{h}\right)\right)\geq\epsilon-\frac{1}{m}\right). \tag{13}$$

Proof $\forall h\in\mathcal{H}$, $\forall(\tilde{x}_i,\tilde{y}_i)\in\tilde{s}_m$,

$$\begin{cases}\Delta^* h_{\tilde{y}_i}(\tilde{x}_i)\leq 0\\ d_{\ell_\infty,\ell_\infty(s_{2m})}(\Delta^*_\gamma h,\Delta^*_\gamma\overline{h})<\frac{\gamma}{2}\end{cases}\Longrightarrow\begin{cases}\Delta^*_\gamma h_{\tilde{y}_i}(\tilde{x}_i)\leq 0\\ d_{\ell_\infty,\ell_\infty(s_{2m})}(\Delta^*_\gamma h,\Delta^*_\gamma\overline{h})<\frac{\gamma}{2}\end{cases}.$$

$$\begin{cases}\Delta^*_\gamma h_{\tilde{y}_i}(\tilde{x}_i)\leq 0\\ d_{\ell_\infty,\ell_\infty(s_{2m})}(\Delta^*_\gamma h,\Delta^*_\gamma\overline{h})<\frac{\gamma}{2}\end{cases}\Longrightarrow\Delta^*_\gamma\overline{h}_{\tilde{y}_i}(\tilde{x}_i)<\frac{\gamma}{2}\Longrightarrow\Delta^*\overline{h}_{\tilde{y}_i}(\tilde{x}_i)<\frac{\gamma}{2}. \tag{14}$$

Similarly, $\forall h\in\mathcal{H}$, $\forall(x_i,y_i)\in s_m$,

$$\begin{cases}\Delta^*\overline{h}_{y_i}(x_i)<\frac{\gamma}{2}\\ d_{\ell_\infty,\ell_\infty(s_{2m})}(\Delta^*_\gamma h,\Delta^*_\gamma\overline{h})<\frac{\gamma}{2}\end{cases}\Longrightarrow\begin{cases}\Delta^*_\gamma\overline{h}_{y_i}(x_i)<\frac{\gamma}{2}\\ d_{\ell_\infty,\ell_\infty(s_{2m})}(\Delta^*_\gamma h,\Delta^*_\gamma\overline{h})<\frac{\gamma}{2}\end{cases}.$$

$$\begin{cases}\Delta^*_\gamma\overline{h}_{y_i}(x_i)<\frac{\gamma}{2}\\ d_{\ell_\infty,\ell_\infty(s_{2m})}(\Delta^*_\gamma h,\Delta^*_\gamma\overline{h})<\frac{\gamma}{2}\end{cases}\Longrightarrow\Delta^*_\gamma h_{y_i}(x_i)<\gamma\Longrightarrow\Delta^* h_{y_i}(x_i)<\gamma. \tag{15}$$

From (14) it springs that if $d_{\ell_\infty,\ell_\infty(s_{2m})}(\Delta^*_\gamma h,\Delta^*_\gamma\overline{h})<\frac{\gamma}{2}$, then

$$R_{\tilde{s}_m}(h)\leq R_{\gamma/2,\tilde{s}_m}\left(\overline{h}\right).$$

Similarly, from (15) it springs that if $d_{\ell_\infty,\ell_\infty(s_{2m})}(\Delta^*_\gamma h,\Delta^*_\gamma\overline{h})<\frac{\gamma}{2}$, then

$$R_{\gamma/2,s_m}\left(\overline{h}\right)\leq R_{\gamma,s_m}(h).$$

To sum up, for all h in $\mathcal{H}$, there exists $\overline{h}$ in $\overline{\mathcal{H}}(s_{2m})$ such that

$$R_{\tilde{s}_m}(h)-R_{\gamma,s_m}(h)\leq R_{\gamma/2,\tilde{s}_m}\left(\overline{h}\right)-R_{\gamma/2,s_m}\left(\overline{h}\right)$$

and thus

$$R_{\tilde{s}_m}(h)-R_{\gamma,s_m}(h)\leq\sup_{\overline{h}\in\overline{\mathcal{H}}(s_{2m})}\left(R_{\gamma/2,\tilde{s}_m}\left(\overline{h}\right)-R_{\gamma/2,s_m}\left(\overline{h}\right)\right).$$

With this last inequality at hand, (13) then directly results from taking the supremum over $\mathcal{H}$. ■

Lemma 3 will prove useful for two reasons. First, it completes, in some sense, Lemma 1, by replacing the two different empirical measures appearing in the right-hand side of (10) with two independent copies of the same random variable. Second, it makes it possible to substitute, in the forthcoming computations, the set $\mathcal{H}$ of possibly infinite cardinality with a subset of it of cardinality no more than $\mathcal{N}^{(p)}_{\infty,\infty}(\gamma/2,\Delta^*_\gamma\mathcal{H},2m)$. This will be exploited to apply a standard union bound.

3.3 Maximal inequality

To bound from above the right-hand side of (13), we introduce an auxiliary step of randomization. To that end, let us consider a set $\mathfrak{S}$ of permutations σ over $\{1, \ldots, 2m\}$. For every sample $s_{2m} = (s_m, \tilde{s}_m) \in S^{2m}$, $s_{2m}^{\sigma} = (s_m^{\sigma}, \tilde{s}_m^{\sigma}) = \left\{(x_{\sigma(1)}, y_{\sigma(1)}), \ldots, (x_{\sigma(2m)}, y_{\sigma(2m)})\right\}$ denotes its "range" by σ. Since the set $(s_m, \tilde{s}_m)$ is chosen according to the product probability measure $P_{s_m, \tilde{s}_m}$ over S^{2m}, the right-hand side of (13) is not affected by a permutation σ. One thus obtains:

$$\forall \sigma \in \mathfrak{S},\ \mathbb{P}_{s_m, \tilde{s}_m}\left(\sup_{\overline{h} \in \overline{\mathcal{H}}(s_{2m})} \left(R_{\gamma/2, \tilde{s}_m}\left(\overline{h}\right) - R_{\gamma/2, s_m}\left(\overline{h}\right)\right) \geq \epsilon - \frac{1}{m}\right) =$$

$$\mathbb{P}_{s_m, \tilde{s}_m}\left(\sup_{\overline{h} \in \overline{\mathcal{H}}(s_{2m})} \left(R_{\gamma/2, \tilde{s}_m^{\sigma}}\left(\overline{h}\right) - R_{\gamma/2, s_m^{\sigma}}\left(\overline{h}\right)\right) \geq \epsilon - \frac{1}{m}\right).$$

Averaging the summand of the right-hand side over the whole set $\mathfrak{S}$ gives:

$$\mathbb{P}_{s_m, \tilde{s}_m}\left(\sup_{\overline{h} \in \overline{\mathcal{H}}(s_{2m})} \left(R_{\gamma/2, \tilde{s}_m}\left(\overline{h}\right) - R_{\gamma/2, s_m}\left(\overline{h}\right)\right) \geq \epsilon - \frac{1}{m}\right) =$$

$$\int_{\mathcal{S}^{2m}} \frac{1}{\#\mathfrak{S}} \# \left\{\sigma \in \mathfrak{S} : \sup_{\overline{h} \in \overline{\mathcal{H}}(s_{2m})} \left(R_{\gamma/2, \tilde{s}_m^{\sigma}}\left(\overline{h}\right) - R_{\gamma/2, s_m^{\sigma}}\left(\overline{h}\right)\right) \geq \epsilon - \frac{1}{m}\right\} dF(s_{2m}). \tag{16}$$

The interest of this last expression rests in the fact that it involves an event,

$$E\left(\epsilon, s_{2m}, \overline{\mathcal{H}}(s_{2m})\right) = \left\{\sigma \in \mathfrak{S} : \sup_{\overline{h} \in \overline{\mathcal{H}}(s_{2m})} \left(R_{\gamma/2, \tilde{s}_m^{\sigma}}\left(\overline{h}\right) - R_{\gamma/2, s_m^{\sigma}}\left(\overline{h}\right)\right) \geq \epsilon - \frac{1}{m}\right\},$$

which can be simply expressed in terms of events each of which involves one single function $\overline{h}$ in $\overline{\mathcal{H}}(s_{2m})$. Indeed, setting for all $\overline{h}$ in $\overline{\mathcal{H}}(s_{2m})$,

$$E\left(\epsilon, s_{2m}, \overline{h}\right) = \left\{\sigma \in \mathfrak{S} :\ R_{\gamma/2, \tilde{s}_m^{\sigma}}\left(\overline{h}\right) - R_{\gamma/2, s_m^{\sigma}}\left(\overline{h}\right) \geq \epsilon - \frac{1}{m}\right\},$$

we get

$$E\left(\epsilon, s_{2m}, \overline{\mathcal{H}}(s_{2m})\right) = \bigcup_{\overline{h} \in \overline{\mathcal{H}}(s_{2m})} E\left(\epsilon, s_{2m}, \overline{h}\right). \tag{17}$$

Using a uniform distribution over $\mathfrak{S}$ in (16) yields to:

$$\mathbb{P}_{s_m, \tilde{s}_m}\left(\sup_{\overline{h} \in \overline{\mathcal{H}}(s_{2m})} \left(R_{\gamma/2, \tilde{s}_m}\left(\overline{h}\right) - R_{\gamma/2, s_m}\left(\overline{h}\right)\right) \geq \epsilon - \frac{1}{m}\right) =$$

$$\int_{\mathcal{S}^{2m}} \mathbb{P}_\sigma \left(\sup_{\overline{h} \in \overline{\mathcal{H}}(s_{2m})} \left(R_{\gamma/2, \tilde{s}_m^\sigma} \left(\overline{h}\right) - R_{\gamma/2, s_m^\sigma} \left(\overline{h}\right) \right) \geq \epsilon - \frac{1}{m} \right) dF(s_{2m}). \tag{18}$$

Due to (17), the right-hand side of (18) is equal to

$$\int_{\mathcal{S}^{2m}} \mathbb{P}_\sigma \left(\bigcup_{\overline{h} \in \overline{\mathcal{H}}(s_{2m})} E\left(\epsilon, s_{2m}, \overline{h}\right) \right) dF(s_{2m}).$$

By application of the union bound,

$$\mathbb{P}_\sigma \left(\bigcup_{\overline{h} \in \overline{\mathcal{H}}(s_{2m})} E\left(\epsilon, s_{2m}, \overline{h}\right) \right) \leq \sum_{\overline{h} \in \overline{\mathcal{H}}(s_{2m})} \mathbb{P}_\sigma \left(E\left(\epsilon, s_{2m}, \overline{h}\right) \right). \tag{19}$$

We now bound uniformly the terms appearing in the right-hand side sum. To that end, we appeal to the classical law of large numbers. $\mathfrak{S}$ is chosen to be the set of all permutations that swap some corresponding elements from the first and second half of $\{1, \ldots, 2m\}$. Precisely, for all i in $\{1, \ldots, m\}$, $(\sigma(i), \sigma(i+m))$ is either equal to $(i, i+m)$ or to $(i+m, i)$. For any function $\overline{h}$ in $\overline{\mathcal{H}}(s_{2m})$, let $(\xi_i)_{1 \leq i \leq m}$ be the sequence of losses $\left(\mathbb{1}_{\left\{ \Delta^* \overline{h}_{y_i}(x_i) < \gamma/2 \right\}} \right)_{1 \leq i \leq m}$ (sequence of losses on s_m) and $\left(\tilde{\xi}_i\right)_{1 \leq i \leq m}$ the corresponding sequence of losses on $\tilde{s}_m$. We have then

$$\mathbb{P}_\sigma \left(E\left(\epsilon, s_{2m}, \overline{h}\right) \right) = \mathbb{P} \left(\frac{1}{m} \sum_{i=1}^m \alpha_i \left(\tilde{\xi}_i - \xi_i \right) \geq \epsilon - \frac{1}{m} \right) \tag{20}$$

where the coefficients α_i, $(1 \leq i \leq m)$, are chosen independently and uniformly on $\{-1, 1\}$. To bound from above the right-hand side of (20), an exponential bound can be applied.

3.4 Exponential bound

Hoeffding's inequality (see for example [38, 55]) is a consequence of Chernoff's inequality [50].

Theorem 2 (Hoeffding's inequality) *Let $X_1, X_2, \ldots, X_n$ be n independent random variables with zero means and bounded ranges: $a_i \leq X_i \leq b_i$. Then, for all $\eta > 0$,*

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

Since the random variables $\alpha_i \left(\tilde{\xi}_i - \xi_i \right)$ take their values in $[-1, 1]$, applying this bound to the right-hand side of (20) gives:

$$\mathbb{P}_\sigma \left(E\left(\epsilon, s_{2m}, \overline{h}\right) \right) \leq \exp \left(-\frac{m}{2} \left(\epsilon - \frac{1}{m} \right)^2 \right).$$

By substitution into the right-hand side of (19) we get:

$$\mathbb{P}_\sigma\left(\bigcup_{\overline{h}\in\overline{\mathcal{H}}(s_{2m})} E\left(\epsilon, s_{2m}, \overline{h}\right)\right) \leq \mathcal{N}^{(p)}(\gamma/2, \Delta^*_\gamma\mathcal{H}, d_{\ell_\infty,\ell_\infty(s_{2m})}) \exp\left(-\frac{m}{2}\left(\epsilon - \frac{1}{m}\right)^2\right).$$

From (18) it then springs that:

$$\mathbb{P}_{s_m,\tilde{s}_m}\left(\sup_{\overline{h}\in\overline{\mathcal{H}}(s_{2m})}\left(R_{\gamma/2,\tilde{s}_m}\left(\overline{h}\right) - R_{\gamma/2,s_m}\left(\overline{h}\right)\right) \geq \epsilon - \frac{1}{m}\right) \leq$$

$$\int_{S^{2m}} \mathcal{N}^{(p)}(\gamma/2, \Delta^*_\gamma\mathcal{H}, d_{\ell_\infty,\ell_\infty(s_{2m})}) \exp\left(-\frac{m}{2}\left(\epsilon - \frac{1}{m}\right)^2\right) dF(s_{2m}).$$

The right-hand side is simply equal to

$$\exp\left(-\frac{m}{2}\left(\epsilon - \frac{1}{m}\right)^2\right)\mathbb{E}\left(\mathcal{N}^{(p)}(\gamma/2, \Delta^*_\gamma\mathcal{H}, d_{\ell_\infty,\ell_\infty(s_{2m})})\right).$$

By definition, $\mathcal{N}^{(p)}_{\infty,\infty}(\gamma/2, \Delta^*_\gamma\mathcal{H}, 2m) \geq \mathbb{E}\left(\mathcal{N}^{(p)}(\gamma/2, \Delta^*_\gamma\mathcal{H}, d_{\ell_\infty,\ell_\infty(s_{2m})})\right)$. Thus, applying Lemma 3, we get

$$\mathbb{P}_{s_m,\tilde{s}_m}\left(\sup_{h\in\mathcal{H}}\left(R_{\tilde{s}_m}(h) - R_{\gamma,s_m}(h)\right) \geq \epsilon - \frac{1}{m}\right) \leq$$

$$\mathcal{N}^{(p)}_{\infty,\infty}(\gamma/2, \Delta^*_\gamma\mathcal{H}, 2m)\exp\left(-\frac{m}{2}\left(\epsilon - \frac{1}{m}\right)^2\right)$$

and consequently, by application of Lemma 1,

$$\mathbb{P}_{s_m}\left(\sup_{h\in\mathcal{H}}\left(R(h) - R_{\gamma,s_m}(h)\right) > \epsilon\right) \leq$$

$$2\mathcal{N}^{(p)}_{\infty,\infty}(\gamma/2, \Delta^*_\gamma\mathcal{H}, 2m)\exp\left(-\frac{m}{2}\left(\epsilon - \frac{1}{m}\right)^2\right). \tag{21}$$

Setting the right-hand side of (21) to δ and solving for ϵ finally gives:

Proposition 1 *Suppose that s_m is chosen by m independent draws from F. Then with probability at least $1-\delta$, every h in $\mathcal{H}$ has*

$$R(h) \leq R_{\gamma,s_m}(h) + \sqrt{\frac{2}{m}\left(\ln\left(2\mathcal{N}^{(p)}_{\infty,\infty}(\gamma/2, \Delta^*_\gamma\mathcal{H}, 2m)\right) - \ln(\delta)\right)} + \frac{1}{m}. \tag{22}$$

3.5 Uniform bound over the margin parameter γ

Making use of the proposition above requires to specify the quantity γ in advance. As pointed out by Bartlett in [8], this seems unnatural. For instance, this constraint makes it difficult to use bounds devoted to the case of a null empirical margin risk (see for instance [64]). This is a significant difference indeed, since faster rates of convergence can be derived either in this case, or in the case where there exists at least one function in $\mathcal{H}$ with zero probability of error, what Vapnik calls the *optimistic case* in [73]. Fortunately, this difficulty can be overcome thanks to the following proposition, proved in [8], and extended in [45], which allows us to give a result that stands uniformly for all values of the margin γ in the interval $(0,1]$.

Proposition 2 (Bartlett, Proposition 8 in [8]) *Let $(\Omega, \mathcal{B}, \mathbb{P})$ be a probability space, and let*

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

be a set of events satisfying the following conditions:

1. *for all $0 < \alpha \leq 1$ and $0 < \delta \leq 1$, $\mathbb{P}(E(\alpha, \alpha, \delta)) \leq \delta$;*
2. *for all $0 < a < 1$ and $0 < \delta \leq 1$, $\bigcup_{\alpha \in (0,1]} E(\alpha a, \alpha, \delta\alpha(1-a))$ is measurable;*
3. *for all $0 < \alpha_1 \leq \alpha \leq \alpha_2 \leq 1$ and $0 < \delta_1 \leq \delta \leq 1$, $E(\alpha_1, \alpha_2, \delta_1) \subseteq E(\alpha, \alpha, \delta)$.*

Then for $0 < a, \delta < 1$

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

To apply Proposition 2 to the case of interest, let us define the function Φ as follows:

$$\Phi(t, u) = \sqrt{\frac{2}{m}\left(\ln\left(2\mathcal{N}^{(p)}_{\infty,\infty}(t, \Delta^*_\gamma \mathcal{H}, 2m)\right) - \ln(u)\right)}.$$

The set of events $E(\alpha_1, \alpha_2, \delta)$ given by:

$$E(\alpha_1, \alpha_2, \delta) = \left\{ s_m \in \mathcal{S}^m : \sup_{h \in \mathcal{H}} \left(R(h) - R_{\alpha_2, s_m}(h)\right) \geq \Phi(\alpha_1/2, \delta) + \frac{1}{m} \right\}$$

satisfies the hypotheses of Proposition 2.

1. For all α in $(0,1]$ and all δ in $(0,1]$,

$$E(\alpha, \alpha, \delta) = \left\{ s_m \in \mathcal{S}^m : \sup_{h \in \mathcal{H}} \left(R(h) - R_{\alpha, s_m}(h)\right) \geq \Phi(\alpha/2, \delta) + \frac{1}{m} \right\}$$

so that $\mathbb{P}_{s_m}\left(E(\alpha, \alpha, \delta)\right) \leq \delta$ by Proposition 1.

2. This requirement follows since all sets of samples are measurable.
3. From the definition of the empirical margin risk,

$$\alpha \leq \alpha_2 \Longrightarrow R_{\alpha,s_m}(h) \leq R_{\alpha_2,s_m}(h).$$

Similarly, by definition of the covering numbers,

$$\alpha_1 \leq \alpha \Longrightarrow \leq \mathcal{N}^{(p)}_{\infty,\infty}(\alpha/2, \Delta^*_\gamma \mathcal{H}, 2m) \leq \mathcal{N}^{(p)}_{\infty,\infty}(\alpha_1/2, \Delta^*_\gamma \mathcal{H}, 2m).$$

Thus, $0 < \alpha_1 \leq \alpha \leq \alpha_2$ and $0 < \delta_1 \leq \delta$ implies that

$$\Phi(\alpha/2, \delta) \leq \Phi(\alpha_1/2, \delta_1).$$

Putting this together yields:

$$R_{\alpha,s_m}(h) + \Phi(\alpha/2, \delta) \leq R_{\alpha_2,s_m}(h) + \Phi(\alpha_1/2, \delta_1)$$

and finally

$$E(\alpha_1, \alpha_2, \delta_1) \subseteq E(\alpha, \alpha, \delta).$$

The application of Proposition 2 gives, for all choice of the couple (a, δ) in $(0,1) \times (0,1]$,

$$\mathbb{P}_{s_m}\left[\bigcup_{\alpha \in (0,1]} \left(\sup_{h \in \mathcal{H}} \left(R(h) - R_{\alpha,s_m}(h)\right) \geq \Phi(\alpha a/2, \delta\alpha(1-a)) + \frac{1}{m}\right)\right] \leq \delta.$$

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

$$\mathbb{P}_{s_m}\left[\bigcup_{\gamma \in (0,1]} \left(\sup_{h \in \mathcal{H}} \left(R(h) - R_{\gamma,s_m}(h)\right) \geq \Phi(\gamma/4, \delta\gamma/2) + \frac{1}{m}\right)\right] \leq \delta$$

and finally, by definition of Φ,

$$\mathbb{P}_{s_m}\left[\bigcup_{\gamma \in (0,1]} \left(\sup_{h \in \mathcal{H}} \left(R(h) - R_{\gamma,s_m}(h)\right) \geq \sqrt{\frac{2}{m}\left(\ln\left(2\mathcal{N}^{(p)}_{\infty,\infty}(\gamma/4, \Delta^*_\gamma \mathcal{H}, 2m)\right) - \ln\left(\frac{\gamma\delta}{2}\right)\right)} + \frac{1}{m}\right)\right] \leq \delta,$$

which concludes the proof of Theorem 1.

3.6 Choice of the "margin" operator

Theorem 1 has been derived for the margin operator specified in Definition 4. In earlier works on the generalization capabilities of multi-class discriminant models, we used a slightly different definition of this operator, namely:

Definition 11 (Δ operator [30]) *Define Δ as an operator on $\mathcal{H}$ such that:*

$$\begin{aligned} \Delta : \quad & \mathcal{H} \longrightarrow \Delta\mathcal{H} \\ & h = (h_k) \mapsto \Delta h = (\Delta h_k) \end{aligned}$$

$$\forall x \in \mathcal{X},\ \forall k \in \{1, \dots, Q\},\ \Delta h_k(x) = \frac{1}{2}\left\{h_k(x) - \max_{l \neq k} h_l(x)\right\}. \qquad (23)$$

It is easy to check that the proof of Theorem 1 still holds if one substitutes Δ to Δ^*. The choice between the two operators should thus rest on the use which is done of the bound, i.e. on the subsequent computations required to bound the covering numbers of interest. This question, the nature of which is primarily technical, will appear 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 Vapnik-Chervonenkis dimension. This suggests to implement a hybrid strategy, mixing results involving Δ^* and results involving Δ. This is precisely what will be done here. In what follows, $\Delta^{\#}$ will be used in place of Δ and Δ^*, in the formulation of the results standing for both operators.

4 Scale-sensitive Ψ-dimensions

Several approaches can be applied to bound from above the covering numbers of interest for a given family of functions $\mathcal{H}$. In this report, we focus on the standard pathway, in which the covering numbers are first related to an extended notion of Vapnik-Chervonenkis (VC) dimension [74], for which an upper bound is computed afterwards. The basic result relating a covering number (precisely the growth function) to the VC dimension is the Sauer-Shelah lemma [74, 60, 65]. 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 [37, 3]), 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 [41, 42]. In the multi-class case, several alternative solutions were proposed by different authors, such as the graph dimension [25, 53], or the Natarajan dimension [53]. It was proved in [11] 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 [68]. In this scheme, they appear as special cases of Ψ-dimensions.

In this section, we consider 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 12 (Ψ-shattering [11]) *Let $\mathcal{F}$ be a class of functions on a set $\mathcal{X}$ taking their values in the finite set $\{1, \dots, Q\}$. Let Ψ be a set of mappings ψ from $\{1, \dots, Q\}$ into $\{-1, 1, *\}$, where $*$ is thought of as a null element. A subset $s_{\mathcal{X}^m} = \{x_i : 1 \leq i \leq m\}$, of $\mathcal{X}$ is said to be* Ψ-shattered *by $\mathcal{F}$ if there is a mapping $\psi^m = \left(\psi^{(1)}, \dots, \psi^{(i)}, \dots, \psi^{(m)}\right)$ in Ψ^m such that for each vector v_y of $\{-1, 1\}^m$, there is a function f_y in $\mathcal{F}$ satisfying*

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

Definition 13 (Ψ-dimension [11]) *Let $\mathcal{F}$ and Ψ be defined as above. The* Ψ-dimension *of $\mathcal{F}$, denoted by Ψ-dim($\mathcal{F}$), is the maximal cardinality of a subset of $\mathcal{X}$ Ψ-shattered by $\mathcal{F}$, if it is finite, or infinity otherwise.*

In words, the idea common to all these dimensions is to introduce adequately chosen mappings from $\{1, \dots, Q\}$ into $\{-1, 1, *\}$ so that the problem of the computation of the capacity measure boils down to the computation of a standard VC dimension. 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, defined as follows:

Definition 14 (Graph dimension [25, 53]) *Let $\mathcal{F}$ be a class of functions on a set $\mathcal{X}$ taking their values in a countable set. For any $f \in \mathcal{F}$, the* graph *of f is $\mathcal{G}(f) = \{(x, f(x)) : x \in \mathcal{X}\}$ and the* graph space *of $\mathcal{F}$ is $\mathcal{G}(\mathcal{F}) = \{\mathcal{G}(f) : f \in \mathcal{F}\}$. Then the* graph dimension *of $\mathcal{F}$, G-dim($\mathcal{F}$), is defined to be the VC dimension of the space $\mathcal{G}(\mathcal{F})$.*

When the functions in $\mathcal{F}$ have a finite range, the reformulation of this definition as the one of a Ψ-dimension is the following:

Definition 15 (Graph dimension) *Let $\mathcal{F}$ be a class of functions on a set $\mathcal{X}$ taking their values in $\{1, \ldots, Q\}$. The* graph dimension *of $\mathcal{F}$ is the Ψ-dimension of $\mathcal{F}$ in the specific case where Ψ is the set of Q mappings ψ_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.*

In the sequel, the scale-sensitive Ψ-dimension which will be considered more specifically is an extension of the Natarajan dimension.

Definition 16 (Natarajan dimension [53]) *Let $\mathcal{F}$ be a class of functions on a set $\mathcal{X}$ taking their values in $\{1, \ldots, Q\}$. The* Natarajan dimension *of $\mathcal{F}$, N-dim($\mathcal{F}$), is the Ψ-dimension of $\mathcal{F}$ in the specific case where Ψ is the set of $Q(Q-1)$ mappings $\psi_{k,l}$, $(1 \leq k \neq l \leq Q)$, such that $\psi_{k,l}$ takes the value 1 if its argument is equal to k, the value -1 if its argument is equal to l, and $*$ otherwise.*

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. We introduce the definition of this latter dimension progressively.

Definition 17 (Vapnik dimension [72]) *Let $\mathcal{H}$ be a class of real-valued functions on a set $\mathcal{X}$. A subset $s_{\mathcal{X}^m} = \{x_i : 1 \leq i \leq m\}$ of $\mathcal{X}$ is said to be* V-shattered *by $\mathcal{H}$ if there is a scalar b such that, for each binary vector $v_y = (y_i) \in \{-1, 1\}^m$, there is a function $h_y \in \mathcal{H}$ satisfying*

$$\forall i \in \{1, \ldots, m\}, \begin{cases} h_y(x_i) - b \geq 0 & if\ y_i = 1 \\ h_y(x_i) - b < 0 & if\ y_i = -1 \end{cases} .$$

The Vapnik dimension *of $\mathcal{H}$, V-dim($\mathcal{H}$), is the maximal cardinality of a subset of $\mathcal{X}$ V-shattered by $\mathcal{H}$, if it is finite, or infinity otherwise.*

The Vapnik dimension is a uniform variant of Pollard's pseudo-dimension.

Definition 18 (Pollard's pseudo-dimension [56, 36]) *Let $\mathcal{H}$ be a class of real-valued functions on a set $\mathcal{X}$. A subset $s_{\mathcal{X}^m} = \{x_i : 1 \leq i \leq m\}$ of $\mathcal{X}$ is said to be P-*shattered *by $\mathcal{H}$ if there is a vector $v_b = (b_i) \in \mathbb{R}^m$ such that, for each binary vector $v_y = (y_i) \in \{-1, 1\}^m$, there is a function $h_y \in \mathcal{H}$ satisfying*

$$\forall i \in \{1, \ldots, m\}, \begin{cases} h_y(x_i) - b_i \geq 0 & if\ y_i = \quad 1 \\ h_y(x_i) - b_i < 0 & if\ y_i = \ -1 \end{cases}.$$

The pseudo-dimension *of $\mathcal{H}$, P-dim$(\mathcal{H})$, is the maximal cardinality of a subset of $\mathcal{X}$ P-shattered by $\mathcal{H}$, if it is finite, or infinity otherwise.*

The V_γ dimension is a scale-sensitive variant of the Vapnik dimension.

Definition 19 (V_γ dimension [3, 35]) *Let $\mathcal{H}$ be a class of real-valued functions on a set $\mathcal{X}$. For $\gamma > 0$, a subset $s_{\mathcal{X}^m} = \{x_i : 1 \leq i \leq m\}$ of $\mathcal{X}$ is said to be V_γ-*shattered *by $\mathcal{H}$ if there is a scalar b such that, for each binary vector $v_y = (y_i) \in \{-1, 1\}^m$, there is a function $h_y \in \mathcal{H}$ satisfying*

$$(h_y(x_i) - b)\, y_i \geq \gamma, \ (1 \leq i \leq m).$$

The V_γ dimension *of $\mathcal{H}$, V_γ-dim$(\mathcal{H})$, is the maximal cardinality of a subset of $\mathcal{X}$ V_γ-shattered by $\mathcal{H}$, if it is finite, or infinity otherwise.*

In the same way as the Vapnik dimension can be seen as a uniform variant of the pseudo-dimension, the V_γ dimension can be seen as a uniform variant of the fat-shattering dimension.

Definition 20 (fat-shattering dimension [41, 42]) *Let $\mathcal{H}$ be a class of real-valued functions on a set $\mathcal{X}$. For $\gamma > 0$, a subset $s_{\mathcal{X}^m} = \{x_i : 1 \leq i \leq m\}$ of $\mathcal{X}$ is said to be γ-*shattered *by $\mathcal{H}$ if there is a vector $v_b = (b_i) \in \mathbb{R}^m$ such that, for each binary vector $v_y = (y_i) \in \{-1, 1\}^m$, there is a function $h_y \in \mathcal{H}$ satisfying*

$$(h_y(x_i) - b_i)\, y_i \geq \gamma, \ (1 \leq i \leq m).$$

The fat-shattering dimension with margin γ, *or P_γ* dimension *of the class $\mathcal{H}$, P_γ-dim$(\mathcal{H})$, is the maximal cardinality of a subset of $\mathcal{X}$ γ-shattered by $\mathcal{H}$, if it is finite, or infinity otherwise.*

With these definitions at hand, the Ψ-dimensions with margin γ, or γ-Ψ-dimensions, are defined as follows:

Definition 21 (γ-Ψ-shattering) *Let $\mathcal{H}$ be a class of functions on a set $\mathcal{X}$ taking their values in $\mathbb{R}^Q$. Let Ψ be a family of mappings ψ from $\{1, \ldots, Q\}$ into $\{-1, 1, *\}$. For $\gamma > 0$, a subset $s_{\mathcal{X}^m} = \{x_i : 1 \leq i \leq m\}$ of $\mathcal{X}$ is said to be γ-Ψ-*shattered *(Ψ-shattered with margin γ) by $\Delta\mathcal{H}$ if there is a mapping $\psi^m = \left(\psi^{(1)}, \ldots, \psi^{(i)}, \ldots, \psi^{(m)}\right)$ in Ψ^m and a vector $v_b = (b_i)$ in $\mathbb{R}^m$ such that, for each vector $v_y = (y_i)$ of $\{-1, 1\}^m$, there is a function h_y in $\mathcal{H}$ satisfying*

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

Definition 22 (Ψ-dimension with margin γ) *Let $\mathcal{H}$, Ψ and γ be defined as above. The Ψ-dimension of $\Delta\mathcal{H}$ with margin γ, denoted by Ψ-dim($\Delta\mathcal{H}, \gamma$), is the maximal cardinality of a subset of $\mathcal{X}$ γ-Ψ-shattered by $\Delta\mathcal{H}$, if it is finite, or infinity otherwise.*

Given the definitions of the Natarajan dimension and the scale-sensitive Ψ-dimensions, the margin Natarajan dimension, the generalized VC dimension which will be involved in our extended Sauer-Shelah lemma, can be formulated as:

Definition 23 (Natarajan dimension with margin γ) *Let $\mathcal{H}$ be a class of functions on a set $\mathcal{X}$ taking their values in $\mathbb{R}^Q$. For $\gamma > 0$, a subset $s_{\mathcal{X}^m} = \{x_i : 1 \leq i \leq m\}$ of $\mathcal{X}$ is said to be* γ-N-shattered *(N-shattered with margin γ) by $\Delta\mathcal{H}$ if there is a set*

$$I(s_{\mathcal{X}^m}) = \{(i_1(x_i), i_2(x_i)) : 1 \leq i \leq m\}$$

of m couples of distinct indexes in $\{1, \dots, Q\}$ and a vector $v_b = (b_i)$ in $\mathbb{R}^m$ such that, for each binary vector $v_y = (y_i) \in \{-1, 1\}^m$, there is a function h_y in $\mathcal{H}$ satisfying

$$\forall i \in \{1, \dots, m\}, \left\{ \begin{array}{lrl} if\ y_i = & 1, & \Delta h_{y,i_1(x_i)}(x_i) - b_i \geq \gamma \\ if\ y_i = & -1, & \Delta h_{y,i_2(x_i)}(x_i) + b_i \geq \gamma \end{array} \right. .$$

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

4.3 Discussion

When applied to the bi-class case, Definition 22 corresponds to the fat-shattering dimension. Indeed, in that case, one can consider that any real-valued function $\tilde{h}$ computed by the model $\tilde{\mathcal{H}}$ is simply equal to $1/2\,(h_1 - h_2)$, where h_1 and h_2 are the two components of an underlying vector-valued function h in a class $\mathcal{H}$. The simplest such configuration corresponds to the choice $h_1 = \tilde{h} = -h_2$. Then, $\Delta h_1 = \tilde{h}$ and $\Delta h_2 = -\tilde{h}$. As a consequence, P_γ-dim($\tilde{\mathcal{H}}$) = Ψ-dim($\Delta\mathcal{H}, \gamma$), this result holding irrespective of the choice of the class of mappings Ψ and the specific mappings $\psi^{(i)}$ associated with the set of points to be shattered. 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 link 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 6, we will see that these extra degrees of freedom can be handled pretty easily.

In the preceding section, we have given a formulation of the definition of the Natarajan dimension which is inspired from the one in [11] (the definition in [53] 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 $k \neq l$. Obviously, this change does not modify the definition. On the other hand, it highlights the fact that the cardinality of the set Ψ considered could be devided 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 [11]). An equivalent simplification can be performed in the case of the margin Natarajan dimension.

Proposition 3 *The definition of the Natarajan dimension with margin γ is not affected by the introduction of the additional constraint $i_1(x_i) < i_2(x_i)$, $(1 \le i \le m)$.*

Proof Let $\mathcal{H}_y$ be a subset of $\mathcal{H}$ of cardinality 2^m such that $\Delta\mathcal{H}_y$ γ-N-shatters $s_{\mathcal{X}^m}$ with respect to $I(s_{\mathcal{X}^m})$ and v_b. Let $I'(s_{\mathcal{X}^m})$ be a set of m couples of indexes $(i'_1(x_i), i'_2(x_i))$ deduced from $I(s_{\mathcal{X}^m})$ by reordering its elements, i.e. $\forall i \in \{1,\dots,m\}$, $(i'_1(x_i), i'_2(x_i)) = (\min(i_1(x_i), i_2(x_i)), \max(i_1(x_i), i_2(x_i)))$. Let $v_{b'} = (b'_i)$ be the vector of $\mathbb{R}^m$ deduced from v_b as follows: $\forall i \in \{1,\dots,m\}$, $b'_i = b_i$ 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 $\Delta\mathcal{H}_y$ still γ-N-shatters $s_{\mathcal{X}^m}$ with respect to $I'(s_{\mathcal{X}^m})$ and $v_{b'}$. For any vector $v_y = (y_i)$ of $\{-1,1\}^m$, let $h_{y'}$ be the function in $\mathcal{H}_y$ such that $\Delta h_{y'}$ "contributes" to the γ-N-shattering of $s_{\mathcal{X}^m}$ with respect to $I(s_{\mathcal{X}^m})$ 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 23,

$$\forall i \in \{1,\dots,m\}, \begin{cases} if\ y'_i = 1, & \Delta h_{y',i_1(x_i)}(x_i) - b_i \ge \gamma \\ if\ y'_i = -1, & \Delta h_{y',i_2(x_i)}(x_i) + b_i \ge \gamma \end{cases}.$$

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

$$\begin{cases} if\ y_i = 1, & \Delta h_{y',i'_1(x_i)}(x_i) - b'_i \ge \gamma \\ if\ y_i = -1, & \Delta h_{y',i'_2(x_i)}(x_i) + b'_i \ge \gamma \end{cases}. \tag{24}$$

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

$$\begin{cases} if\ y_i = -1, & \Delta h_{y',i'_2(x_i)}(x_i) + b'_i \ge \gamma \\ if\ y_i = 1, & \Delta h_{y',i'_1(x_i)}(x_i) - b'_i \ge \gamma \end{cases}.$$

This is exactly (24), which thus holds true for all values of i in $\{1,\dots,m\}$ (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 23, function $\Delta h_{y'}$ thus contributes to the γ-N-shattering of $s_{\mathcal{X}^m}$ with respect to $I'(s_{\mathcal{X}^m})$ 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 $\{-1,1\}^m$, this implies that $\Delta\mathcal{H}_y$ γ-N-shatters $s_{\mathcal{X}^m}$ with respect to $I'(s_{\mathcal{X}^m})$ and $v_{b'}$, which, by construction of $I'(s_{\mathcal{X}^m})$, concludes the proof. ∎

In the sequel, when needed, we will implicitly make use of Proposition 3.

5 Relating the Covering Numbers and the Margin Natarajan Dimension

To introduce the central result of this section, straightforward extensions of several lemmas in [3] must first be derived. These lemmas involve additional concepts which are defined below. For the sake of simplicity and efficiency, in what follows, the concepts and lemmas are not considered or expressed in their full generality, but rather formulated in the specific context in which they will be used.

5.1 Definitions

Definition 24 (η-discretization) *Let $h = (h_k)$ be a function from $\mathcal{X}$ into $\mathbb{R}^Q$ and let (γ, η) be a couple of reals such that $0 < \eta \leq \gamma$. The η-*discretization *of $\Delta^{\#}_{\gamma} h$, denoted by $\left(\Delta^{\#}_{\gamma} h\right)^{(\eta)} = \left(\left(\Delta^{\#}_{\gamma} h_k\right)^{(\eta)}\right)$, is the function from $\mathcal{X}$ into $\mathbb{Z}^Q$ such that*

$$\forall k \in \{1, \ldots, Q\}, \ \left(\Delta^{\#}_{\gamma} h_k\right)^{(\eta)}(x) = \begin{cases} \left\lfloor \frac{\Delta^{\#}_{\gamma} h_k(x)}{\eta} \right\rfloor & \text{if } \Delta^{\#} h_k(x) \geq 0 \\ - \left\lfloor \frac{\left|\Delta^{\#}_{\gamma} h_k(x)\right|}{\eta} \right\rfloor & \text{otherwise} \end{cases}$$

or equivalently, for all k in $\{1, \ldots, Q\}$, $\left(\Delta^{\#}_{\gamma} h_k\right)^{(\eta)}(x) = \max\left\{j \in \mathbb{N} : \ j\eta \leq \Delta^{\#}_{\gamma} h_k(x)\right\}$ if $\Delta^{\#} h_k(x) \geq 0$, $\left(\Delta^{\#}_{\gamma} h_k\right)^{(\eta)}(x) = -\max\left\{j \in \mathbb{N} : \ j\eta \leq \left|\Delta^{\#}_{\gamma} h_k(x)\right|\right\}$ otherwise. For a set $\mathcal{H}$ of vector-valued functions, let $\left(\Delta^{\#}_{\gamma} \mathcal{H}\right)^{(\eta)} = \left\{\left(\Delta^{\#}_{\gamma} h\right)^{(\eta)} : \ h \in \mathcal{H}\right\}$.

Note that this definition is not a straightforward extension of the original one, which can be found in [3], 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 25 (Strong Natarajan dimension) *Let $\mathcal{H}$ be a class of functions from $\mathcal{X}$ into $\mathbb{R}^Q$ and let (γ, η) be a couple of reals such that $0 < \eta \leq \gamma$ and the functions in $(\Delta_{\gamma} \mathcal{H})^{(\eta)}$ take their values in $S = \{-n, \ldots, n\}^Q$. A subset $s_{\mathcal{X}^m} = \{x_i : 1 \leq i \leq m\}$ of $\mathcal{X}$ is said to be* strongly N-shattered *by $(\Delta_{\gamma} \mathcal{H})^{(\eta)}$ if there exists a set of couples of indexes*

$$I(s_{\mathcal{X}^m}) = \{(i_1(x_i), i_2(x_i)) : 1 \leq i \leq m\}$$

with $1 \leq i_1(x_i) < i_2(x_i) \leq Q$, $(1 \leq i \leq m)$, and a vector $v_b = (b_i)$ in $\{-n+1, \ldots, n-1\}^m$ such that, for each binary vector $v_y = (y_i) \in \{-1, 1\}^m$, there is a function $(\Delta_{\gamma} h_y)^{(\eta)} = \left((\Delta_{\gamma} h_{y,k})^{(\eta)}\right)_{1 \leq k \leq Q}$ in $(\Delta_{\gamma} \mathcal{H})^{(\eta)}$ satisfying

$$\forall i \in \{1, \ldots, m\}, \ \begin{cases} \text{if } y_i = \ \ 1, & \left(\Delta_{\gamma} h_{y, i_1(x_i)}\right)^{(\eta)}(x_i) - b_i \geq 1 \\ \text{if } y_i = -1, & \left(\Delta_{\gamma} h_{y, i_2(x_i)}\right)^{(\eta)}(x_i) + b_i \geq 1 \end{cases} .$$

The strong Natarajan dimension *of the class* $(\Delta_\gamma \mathcal{H})^{(\eta)}$*, SN-dim* $\left((\Delta_\gamma \mathcal{H})^{(\eta)}\right)$*, is the maximal cardinality of a subset of* $\mathcal{X}$ *strongly N-shattered by* $(\Delta_\gamma \mathcal{H})^{(\eta)}$*, if it is finite, or infinity otherwise.*

Definition 26 (Packing numbers) *Let* (E, ρ) *be a pseudo-metric space and* $\epsilon \in \mathbb{R}_+^*$*. A set* $E' \subset E$ *is* ϵ-separated *if, for any distinct points* e_1 *and* e_2 *in* E'*,* $\rho(e_1, e_2) \geq \epsilon$*. The* ϵ-packing number *of* $E'' \subset E$*,* $\mathcal{M}(\epsilon, E'', \rho)$*, is the maximal size of an* ϵ*-separated subset of* E''*.*

Definition 27 (Separation) *Let* $\mathcal{F}$ *be a class of functions on a domain* $\mathcal{X}$ *taking their values in* $\mathbb{R}^Q$ *and* $\mathcal{F}|_{\mathcal{D}}$ *its restriction to* $\mathcal{D} \subseteq \mathcal{X}$*. Two functions* f *and* f' *in the class* $\mathcal{F}|_{\mathcal{D}}$ *are* separated *if they are 2-separated with respect to the pseudo-metric* $d_{\ell_\infty, \ell_\infty(\mathcal{D})}$*, i.e. if*

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

Definition 28 (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.*

5.2 Lemmas

There is a close connection between covering and packing properties of bounded subsets in metric spaces. The following well-known lemma, introduced in [43] (see also [23, 3, 5] for more recent references), will prove useful in what follows.

Lemma 4 *For every pseudo-metric space* (E, ρ)*, every totally bounded subset* E' *of* E *and* $\epsilon > 0$*,*

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

Proof We are interested here in covering numbers computed from proper covers. For this reason, we establish a slightly different result, namely $\mathcal{M}(2\epsilon, E', \rho) \leq \mathcal{N}^{(p)}(\epsilon, E', \rho) \leq \mathcal{M}(\epsilon, E', \rho)$.

We first prove the left-hand side inequality. Let $E'' = \{e_i'' : 1 \leq i \leq \mathcal{M}(2\epsilon, E', \rho)\}$ be a 2ϵ-separated subset of E' of maximal cardinality. Let $E''' = \left\{e_i''' : 1 \leq i \leq \mathcal{N}^{(p)}(\epsilon, E', \rho)\right\}$ be a proper ϵ-cover of E' of minimal cardinality. Suppose that $\#E''' < \#E''$. Then, according to the pigeonhole principle, there exists at least a triplet (i, j, k) in $\{1, \ldots, \mathcal{M}(2\epsilon, E', \rho)\}^2 \times \{1, \ldots, \mathcal{N}(\epsilon, E', \rho)\}$, with $i \neq j$, such that $\rho(e_i'', e_k''') < \epsilon$ and $\rho\left(e_j'', e_k'''\right) < \epsilon$. By application of the triangle inequality, this implies that $\rho\left(e_i'', e_j''\right) < 2\epsilon$, which is in contradiction with the definition of E''. Thus, the hypothesis $\#E''' < \#E''$ leads to a contradiction, which concludes the proof. Note that we did not make use of the fact that $E''' \subset E'$. This is due to the fact that it suffices to establish $\mathcal{M}(2\epsilon, E', \rho) \leq \mathcal{N}(\epsilon, E', \rho)$ to get $\mathcal{M}(2\epsilon, E', \rho) \leq \mathcal{N}^{(p)}(\epsilon, E', \rho)$.

We now turn to the right-hand side inequality, which is precisely the one we will make use of. Let $E'' = \{e_i : 1 \leq i \leq \mathcal{M}(\epsilon, E', \rho)\}$ be an ϵ-separated subset of E' of maximal cardinality. Suppose that there exists e_0 in E' such that $\min_{1 \leq i \leq \mathcal{M}(\epsilon, E', \rho)} \rho(e_0, e_i) \geq \epsilon$. Then $E'' \cup \{e_0\}$ is an ϵ-separated subset of E', which is in contradiction with the definition of E''. As a consequence, all points e in E' satisfy $\min_{1 \leq i \leq \mathcal{M}(\epsilon, E', \rho)} \rho(e, e_i) < \epsilon$, and by definition of the ϵ-covers, E'' is an ϵ-cover of E'. Since $E'' \subset E'$, E'' is even a proper ϵ-cover of E'. By definition of the covering numbers, we have thus $\mathcal{N}^{(p)}(\epsilon, E', \rho) \leq \# E'' = \mathcal{M}(\epsilon, E', \rho)$, which concludes the proof. ∎

With the above definitions at hand, we can prove the following lemma, which extends to the multivariate case Lemma 3.2 in [3]:

Lemma 5 *For any class $\mathcal{H}$ of functions on $\mathcal{X}$ taking their values in $\mathbb{R}^Q$ and for any couple of reals (γ, η) satisfying $0 < \eta \leq \gamma$:*

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

$$\text{SN-dim}\left((\Delta_\gamma \mathcal{H})^{(\eta)} \right) \leq \text{N-dim}(\Delta_\gamma \mathcal{H}, \epsilon);$$

2. *for every $\epsilon \geq 3\eta$ and every $s_{\mathcal{X}^m} \in \mathcal{X}^m$,*

$$\mathcal{M}(\epsilon, \Delta_\gamma^* \mathcal{H}, d_{\ell_\infty, \ell_\infty(s_{\mathcal{X}^m})}) \leq \mathcal{M}(2, \left(\Delta_\gamma^* \mathcal{H}\right)^{(\eta)}, d_{\ell_\infty, \ell_\infty(s_{\mathcal{X}^m})}).$$

Proof To prove the first proposition, it is enough to establish that any set strongly N-shattered by $(\Delta_\gamma \mathcal{H})^{(\eta)}$ is also N-shattered with margin $\eta/2$ by $\Delta_\gamma \mathcal{H}$. If $s_{\mathcal{X}^m}$, a subset of $\mathcal{X}$ of cardinality m, is strongly N-shattered by $(\Delta_\gamma \mathcal{H})^{(\eta)}$, then according to Definition 25, there exists a set of couples of indexes $I(s_{\mathcal{X}^m})$ and a vector v_b in $\{-\lfloor \gamma/\eta \rfloor + 1, \dots, \lfloor \gamma/\eta \rfloor - 1\}^m$ such that for every vector $v_y = (y_i) \in \{-1, 1\}^m$, there is a function $(\Delta_\gamma h_y)^{(\eta)} = \left((\Delta_\gamma h_{y,k})^{(\eta)} \right)$ in $(\Delta_\gamma \mathcal{H})^{(\eta)}$, i.e. a function h_y in $\mathcal{H}$ satisfying

$$\forall i \in \{1, \dots, m\}, \begin{cases} if\ y_i = \quad 1, & \left(\Delta_\gamma h_{y, i_1(x_i)}\right)^{(\eta)}(x_i) - b_i \geq 1 \\ if\ y_i = -1, & \left(\Delta_\gamma h_{y, i_2(x_i)}\right)^{(\eta)}(x_i) + b_i \geq 1 \end{cases}.$$

Thus, we are looking for a scalar b'_i such that $\left(\Delta_\gamma h_{y, i_1(x_i)}\right)^{(\eta)}(x_i) - b_i \geq 1 \Longrightarrow \Delta_\gamma h_{y, i_1(x_i)}(x_i) - b'_i \geq \eta/2$ and $\left(\Delta_\gamma h_{y, i_2(x_i)}\right)^{(\eta)}(x_i) + b_i \geq 1 \Longrightarrow \Delta_\gamma h_{y, i_2(x_i)}(x_i) + b'_i \geq -\eta/2$. To that end, four cases must be considered.

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

$$\left(\Delta_\gamma h_{y, i_1(x_i)}\right)^{(\eta)}(x_i) > 0 \Longrightarrow \eta \left(\Delta_\gamma h_{y, i_1(x_i)}\right)^{(\eta)}(x_i) \leq \Delta_\gamma h_{y, i_1(x_i)}(x_i)$$

thus

$$\left(\Delta_\gamma h_{y, i_1(x_i)}\right)^{(\eta)}(x_i) - b_i \geq 1 \Longrightarrow \Delta_\gamma h_{y, i_1(x_i)}(x_i) - \eta b_i \geq \eta$$

or equivalently

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

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

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

or equivalently

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

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

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

or equivalently

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

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

$$\left(\Delta_\gamma h_{y,i_2(x_i)}\right)^{(\eta)}(x_i) > 0 \Longrightarrow \eta \left(\Delta_\gamma h_{y,i_2(x_i)}\right)^{(\eta)}(x_i) \leq \Delta_\gamma h_{y,i_2(x_i)}(x_i)$$

thus

$$\left(\Delta_\gamma h_{y,i_2(x_i)}\right)^{(\eta)}(x_i) + b_i \geq 1 \Longrightarrow \Delta_\gamma h_{y,i_2(x_i)}(x_i) + \eta b_i \geq \eta$$

or equivalently

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

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 $\Delta_\gamma h_y$, for v_y in $\{-1, 1\}^m$, N-shatters $s_{\mathcal{X}^m}$ with margin $\eta/2$, for a set of couples of indexes and a vector of "biases" respectively equal to $I(s_{\mathcal{X}^m})$ and $v_{b'} = (b'_i)_{1 \leq i \leq m}$. As a consequence, any set strongly N-shattered by $(\Delta_\gamma \mathcal{H})^{(\eta)}$ is also N-shattered by $\Delta_\gamma \mathcal{H}$ with margin $\eta/2$, which is precisely our claim.

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

$$\forall (h, h') \in \mathcal{H}^2,\ \forall x \in \mathcal{X},\ \forall k \in \{1, \ldots, Q\},\ \forall (\gamma, \eta) : \gamma \geq \eta > 0,$$

$$\left|\Delta^*_\gamma h_k(x) - \Delta^*_\gamma h'_k(x)\right| \geq 3\eta \Longrightarrow \left|\left(\Delta^*_\gamma h_k\right)^{(\eta)}(x) - \left(\Delta^*_\gamma h'_k\right)^{(\eta)}(x)\right| \geq 2.$$

Indeed, without loss of generality, we can make the hypothesis that $\Delta_\gamma^* h_k(x) > \Delta_\gamma^* h'_k(x)$. Then,

$$\left(\left(\Delta_\gamma^* h'_k\right)^{(\eta)}(x) - 1 \right)\eta < \Delta_\gamma^* h'_k(x) < \Delta_\gamma^* h_k(x) < \left(\left(\Delta_\gamma^* h_k\right)^{(\eta)}(x) + 1 \right)\eta.$$

Thus

$$\left(\left(\Delta_\gamma^* h_k\right)^{(\eta)}(x) + 1 \right)\eta - \left(\left(\Delta_\gamma^* h'_k\right)^{(\eta)}(x) - 1 \right)\eta > 3\eta$$

and finally

$$\left(\Delta_\gamma^* h_k\right)^{(\eta)}(x) - \left(\Delta_\gamma^* h'_k\right)^{(\eta)}(x) > 1,$$

from which the desired result springs directly, keeping in mind that the η-discretizations are integers $\left(\left(\Delta_\gamma^* h_k\right)^{(\eta)}(x) - \left(\Delta_\gamma^* h'_k\right)^{(\eta)}(x) > 1 \Longrightarrow \left(\Delta_\gamma^* h_k\right)^{(\eta)}(x) - \left(\Delta_\gamma^* h'_k\right)^{(\eta)}(x) \geq 2 \right)$.

Let $s_{\Delta_\gamma^* \mathcal{H}}$ be a 3η-separated subset of $\Delta_\gamma^* \mathcal{H}$ with respect to $d_{\ell_\infty, \ell_\infty(s_{\mathcal{X}^m})}$. It results from the definition of the pseudo-metric that:

$$\forall \ \left(\Delta_\gamma^* h, \Delta_\gamma^* h'\right) \in s^2_{\Delta_\gamma^* \mathcal{H}}, \ d_{\ell_\infty, \ell_\infty(s_{\mathcal{X}^m})}\left(\Delta_\gamma^* h, \Delta_\gamma^* h'\right) \geq 3\eta \Longrightarrow$$

$$\max_{x \in s_{\mathcal{X}^m}} \left\| \Delta_\gamma^* h(x) - \Delta_\gamma^* h'(x) \right\|_\infty \geq 3\eta \Longrightarrow$$

$$\max_{x \in s_{\mathcal{X}^m}} \left\| \left(\Delta_\gamma^* h\right)^{(\eta)}(x) - \left(\Delta_\gamma^* h'\right)^{(\eta)}(x) \right\|_\infty \geq 2 \Longrightarrow$$

$$d_{\ell_\infty, \ell_\infty(s_{\mathcal{X}^m})}\left(\left(\Delta_\gamma^* h_k\right)^{(\eta)}, \left(\Delta_\gamma^* h'_k\right)^{(\eta)} \right) \geq 2.$$

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 $\left(\Delta_\gamma^\# h_k\right)^{(\eta)}(x) = \lfloor \frac{\Delta_\gamma^\# h_k(x)}{\eta} \rfloor$ irrespective of the sign of $\Delta h_k(x)$, one can easily establish that the following proposition (with a dependence between ϵ and η identical to the one of [3]) holds true: for every $\epsilon \geq 2\eta$ and every $s_{\mathcal{X}^m} \in \mathcal{X}^m$, $\mathcal{M}(\epsilon, \Delta_\gamma^* \mathcal{H}, d_{\ell_\infty, \ell_\infty(s_{\mathcal{X}^m})}) \leq \mathcal{M}(2, \left(\Delta_\gamma^* \mathcal{H}\right)^{(\eta)}, d_{\ell_\infty, \ell_\infty(s_{\mathcal{X}^m})})$. The reason for our choice is to get an additional property, namely:

$$\forall (\gamma, \eta) : 0 < \eta \leq \gamma, \ \ \Delta_\gamma^\# h_l(x) = -\Delta_\gamma^\# h_k(x) \Longrightarrow \left(\Delta_\gamma^\# h_l\right)^{(\eta)}(x) = -\left(\Delta_\gamma^\# h_k\right)^{(\eta)}(x).$$

This property is primarily of practical interest, since it leads to proofs that are easier to read. This advantage will appear no later than in the next lemma.

5.3 Relating separation and strong N-shattering

Lemma 6 *Let $\mathcal{H}$ be a class of functions on $\mathcal{X}$ taking their values in $\mathbb{R}^Q$ and let (γ, η) be a couple of reals such that $0 < \eta \leq \gamma$. Let $\mathcal{D} \subseteq \mathcal{X}$ and let $\mathcal{F}$ and $\mathcal{F}^*$ be respectively the restrictions of $(\Delta_\gamma \mathcal{H})^{(\eta)}$ and $\left(\Delta^*_\gamma \mathcal{H}\right)^{(\eta)}$ to $\mathcal{D}$. $\mathcal{F}$ and $\mathcal{F}^*$ are endowed with the pseudo-metric $d_{\ell_\infty, \ell_\infty(\mathcal{D})}$. If two functions h and h' in $\mathcal{H}$ are such that $f^* = \left(\Delta^*_\gamma h\right)^{(\eta)}\Big|_{\mathcal{D}}$ and $f^{*'} = \left(\Delta^*_\gamma h'\right)^{(\eta)}\Big|_{\mathcal{D}}$ are separated, then there exists x in $\mathcal{D}$ such that $\left\{f = (\Delta_\gamma h)^{(\eta)}\Big|_{\mathcal{D}}, f' = (\Delta_\gamma h')^{(\eta)}\Big|_{\mathcal{D}}\right\}$ 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 $n_0 = \operatorname{argmax}_k f^*_k(x)$. Then there is at least one couple $(I(\{x\}), v_b) = (\{(i_1(x), i_2(x))\}, (b_0))$ with $i_1(x) = n_0$ and $b_0 = f^*_{n_0}(x) - 1$ witnessing the strong N-shattering of $\{x\}$ by $\{f, f'\}$.*

Proof By definition of the separation, it springs from the hypotheses that there exists a couple (x, k_0) in $\mathcal{D} \times \{1, \dots, Q\}$ such that $\left|f^*_{k_0}(x) - f^{*'}_{k_0}(x)\right| \geq 2$. We can make the assumption that $\max_k f^*_k(x) \geq \max_k f^{*'}_k(x)$. We first demonstrate that $\operatorname{argmax}_k f^*_k(x)$ is well defined. Indeed, this is the case unless $f^*(x) = 0_Q$. But $f^*(x) = 0_Q$ and $\max_k f^*_k(x) \geq \max_k f^{*'}_k(x)$ implies that $f^{*'}(x) = 0_Q$, which is in contradiction with the hypothesis $\left\|f^*(x) - f^{*'}(x)\right\|_\infty \geq 2$. To finish the proof of the assertion, four cases must be considered.

1) $\left(f^*_{k_0}(x) = \max_k f^*_k(x)\right) \wedge \left(f^{*'}_{k_0}(x) = \max_k f^{*'}_k(x)\right)$

It springs from the hypotheses that $f^*_{k_0}(x) - f^{*'}_{k_0}(x) \geq 2$. Let $b = f^*_{k_0}(x) - 1$. Then $f_{k_0}(x) - b = f^*_{k_0}(x) - b = 1$. Furthermore, there exists an index l_0 different from k_0 such that $f'_{l_0}(x) = f^{*'}_{l_0}(x) = -f^{*'}_{k_0}(x)$ (l_0 is simply the index of a component of $h'(x)$ satisfying $h'_{l_0}(x) = \max_{k \neq k_0} h'_k(x)$). $f'_{l_0}(x) + b = f^*_{k_0}(x) - f^{*'}_{k_0}(x) - 1 \geq 1$. 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 $\{f, f'\}$.

2) $\left(f^*_{k_0}(x) = \max_k f^*_k(x)\right) \wedge \left(f^{*'}_{k_0}(x) \neq \max_k f^{*'}_k(x)\right)$

It springs from the hypotheses that $f^*_{k_0}(x) - f^{*'}_{k_0}(x) \geq 2$. Let $b = f^*_{k_0}(x) - 1$. Then $f_{k_0}(x) - b = f^*_{k_0}(x) - b = 1$. Furthermore, there exits an index l_0 different from k_0 such that $f^{*'}_{l_0}(x) = \max_k f^{*'}_k(x)$. $f'_{l_0}(x) + b = f^{*'}_{l_0}(x) + f^*_{k_0}(x) - 1 = f^*_{k_0}(x) - f^{*'}_{k_0}(x) - 1 \geq 1$. 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 $\{f, f'\}$.

3) $\left(f^*_{k_0}(x) \neq \max_k f^*_k(x)\right) \wedge \left(f^{*'}_{k_0}(x) = \max_k f^{*'}_k(x)\right)$

It springs from the hypotheses that $f^*_{k_0}(x) - f^{*'}_{k_0}(x) \leq -2$. There exits an index l_0 different from k_0 such that $f^*_{l_0}(x) = \max_k f^*_k(x)$. Let $b = f^*_{l_0}(x) - 1$. Then $f_{l_0}(x) - b = f^*_{l_0}(x) - b = 1$. Furthermore, $f'_{k_0}(x) + b = f^{*'}_{k_0}(x) + f^*_{l_0}(x) - 1 = f^{*'}_{k_0}(x) - f^*_{k_0}(x) - 1 \geq 1$. Thus, the couple

$\left(I\left(\{x\}\right)=\left\{(l_0,k_0)\right\},v_b=\left(f^*_{l_0}(x)-1\right)\right)$ witnesses the strong N-shattering of $\{x\}$ by $\{f,f'\}$.

4) $\left(f^*_{k_0}(x)\neq \max_k f^*_k(x)\right)\wedge\left(f^{*'}_{k_0}(x)\neq \max_k f^{*'}_k(x)\right)$

It springs from the hypotheses that $f^*_{k_0}(x)-f^{*'}_{k_0}(x)\leq -2$. There exists a couple of distinct indexes (l_0,m_0) such that $f^*_{l_0}(x)=\max_k f^*_k(x)$ and $f'_{m_0}(x)=f^{*'}_{m_0}(x)=\max_{k\neq l_0} f^{*'}_k(x)$ (m_0 is simply the index of a component of $h'(x)$ satisfying both $m_0\neq l_0$ and $h'_{m_0}(x)=\max_{k\neq l_0} h'_k(x)$). Let $b=f^*_{l_0}(x)-1$. Then $f_{l_0}(x)-b=f^*_{l_0}(x)-b=1$. Furthermore, $f'_{m_0}(x)+b=f^{*'}_{m_0}(x)+f^*_{l_0}(x)-1=f^{*'}_{m_0}(x)-f^*_{k_0}(x)-1$. Since by construction $f^{*'}_{m_0}(x)\geq f^{*'}_{k_0}(x)$ (keeping in mind that $k_0\in\{1,\dots,Q\}\setminus\{l_0\}$), $f'_{m_0}(x)+b\geq f^{*'}_{k_0}(x)-f^*_{k_0}(x)-1$ and finally $f'_{m_0}(x)+b\geq 1$. Thus, the couple $\left(I\left(\{x\}\right)=\left\{(l_0,m_0)\right\},v_b=\left(f^*_{l_0}(x)-1\right)\right)$ witnesses the strong N-shattering of $\{x\}$ by $\{f,f'\}$. ∎

Lemma 6 will appear of central importance in the sequel. We consider it as contributing to highlight the specificity of the multi-class case, since it involves both margin operators, and it cannot be stated with the operator Δ only. To prove this last assertion, it suffices to exhibit a counter example. Let h and h' be two functions in $\mathcal{H}$ such that there exists $\mathcal{D}=\{x\}$ satisfying $h(x)=(0.5,-0.5,-0.9)$, $h'(x)=(0.5,-0.5,-0.5)$. Let $\gamma=1$ and $\eta=0.1$. Using the same notations as above, we get $f(x)=(5,-5,-7)$, $f'(x)=(5,-5,-5)$, $f^*(x)=f^{*'}(x)=(5,-5,-5)$. Also f and f' are separated, they do not strongly N-shatter $\{x\}$. Indeed, if it were the case, then according to Definition 25, there would be two different indexes k_0 and l_0 in $\{1,2,3\}$ such that $f_{k_0}(x)+f'_{l_0}(x)\geq 2$, which is not the case. As a verification measure, the interested reader will easily establish that on the contrary, if f^* and $f^{*'}$ are separated on $\mathcal{D}$, then necessarily there exists x in $\mathcal{D}$ such that there exists two different indexes k_0 and l_0 in $\{1,\dots,Q\}$ such that $f_{k_0}(x)+f'_{l_0}(x)\geq 2$. The proof is very similar to the one of Lemma 6 itself. A tricky thing must be borne in mind. If two pairs $\left(h^{(1)},h^{(2)}\right)$ and $\left(h^{(3)},h^{(4)}\right)$ of functions in $\mathcal{H}$ are such that $f^{*(1)}=f^{*(3)}$ and $f^{*(2)}=f^{*(4)}$, then if $f^{*(1)}$ and $f^{*(2)}$ are separated, there exists $\{x\}\subset\mathcal{D}$ 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 $\left(I\left(\{x\}\right),v_b\right)$. More precisely, using the notations of Definition 25, given the couple $\left(f^*,f^{*'}\right)$, 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. It is thus a priori different for $\left\{f^{(1)},f^{(2)}\right\}$ and for $\left\{f^{(3)},f^{(4)}\right\}$.

We now prove our main combinatorial result, an extension of Lemma 3.3 in [3], which gives a new generalization of the Sauer-Shelah lemma.

5.4 Generalized Sauer-Shelah lemma

Lemma 7 *Let $\mathcal{H}$ be a class of functions on $\mathcal{X}$ taking their values in $\mathbb{R}^Q$ and let (γ, η) be a couple of reals such that $0 < \eta \leq \gamma$. 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 $\left(\Delta_\gamma \mathcal{H}\right)^{(\eta)}$ and $\left(\Delta^*_\gamma \mathcal{H}\right)^{(\eta)}$ to $\mathcal{D}$. $\mathcal{F}$ and $\mathcal{F}^*$ are endowed with the pseudo-metric $d_{\ell_\infty, \ell_\infty(\mathcal{D})}$. Setting $n = \left\lfloor \frac{\gamma}{\eta} \right\rfloor$ and $d =$ SN-dim$(\mathcal{F})$, the following bound holds true:*

$$\mathcal{M}(2, \mathcal{F}^*, d_{\ell_\infty, \ell_\infty(\mathcal{D})}) < 2\left(\#\mathcal{D}\; Q^2(Q-1)\; n^2\right)^{\lceil \log_2(\phi(d, \#\mathcal{D})) \rceil} \tag{25}$$

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

Proof Let us say that the class $\mathcal{F}$ strongly N-shatters a triplet $(s_{\mathcal{D}^m}, I(s_{\mathcal{D}^m}), v_b)$ (for a nonempty subset $s_{\mathcal{D}^m}$ of $\mathcal{D}$ of cardinality m, a set of couples of indexes $I(s_{\mathcal{D}^m})$ and a vector of biases v_b, if $\mathcal{F}$ strongly N-shatters $s_{\mathcal{D}^m}$ according to $I(s_{\mathcal{D}^m})$ 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 from $\mathcal{F}^*$, $\mathcal{F}_l = \{f \in \mathcal{F} : f^* \in \mathcal{F}^*_l\}$ 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 $\sum_{i=1}^{d} \binom{\#\mathcal{D}}{i} \left(\binom{Q}{2}(2n-1)\right)^i$, since for $s_\mathcal{D}$ of size $i > 0$, there are strictly less than $\left(\binom{Q}{2}(2n-1)\right)^i$ 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 and SN-dim$(\mathcal{F}) \leq d$ implies $t(l, \#\mathcal{D}) = \infty$. As a consequence, by definition of $t(l, \#\mathcal{D})$, there is no set $\mathcal{F}^*_l$ of l pairwise separated functions in $\mathcal{F}^*$ (otherwise $t(l, \#\mathcal{D})$ would be finite) and finally, by definition of $\mathcal{M}(2, \mathcal{F}^*, d_{\ell_\infty, \ell_\infty(\mathcal{D})})$, $\mathcal{M}(2, \mathcal{F}^*, d_{\ell_\infty, \ell_\infty(\mathcal{D})}) < l$. Therefore, to finish the proof, it suffices to show that, for all $d \geq 1$ and $\#\mathcal{D} \geq 1$,

$$t\left(2\left(\#\mathcal{D}\; Q^2(Q-1)\; n^2\right)^{\lceil \log_2(\phi(d, \#\mathcal{D})) \rceil}, \#\mathcal{D}\right) \geq \phi(d, \#\mathcal{D}). \tag{26}$$

We claim that

$$t(2, \#\mathcal{D}) \geq 1 \tag{27}$$

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

$$t\left(2m\; \#\mathcal{D}\; Q^2(Q-1)\; n^2, \#\mathcal{D}\right) \geq 2t(2m, \#\mathcal{D} - 1) \tag{28}$$

for all $m \geq 1$ and $\#\mathcal{D} \geq 2$.

The first part of the claim is a direct consequence of Lemma 6.

For the second part, first note that if no set of $2m\ \#\mathcal{D}\ Q^2(Q-1)\ n^2$ pairwise separated functions in $\mathcal{F}^*$ exists, then by definition $t\left(2m\ \#\mathcal{D}\ Q^2(Q-1)\ n^2, \#\mathcal{D}\right) = \infty$ and hence the claim holds. Assume then that there is a set $\mathcal{F}_0^*$ of $2m\ \#\mathcal{D}\ Q^2(Q-1)\ n^2$ pairwise separated functions in $\mathcal{F}^*$. Split it arbitrarily into $m\ \#\mathcal{D}\ Q^2(Q-1)\ n^2$ pairs. For each pair $\left(f^*, f^{*'}\right)$, there exists a point $x \in \mathcal{D}$ strongly N-shattered by $\{f, f'\}$. Once more, this is a direct consequence of Lemma 6. The number of different values that a vector $f^*(x)$ can take is equal to $Qn+1$. 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 $\frac{1}{2}(Qn+1)(Qn-1) < \frac{1}{2}Q^2\ n^2$. 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 $\left(2m\ \#\mathcal{D}\ Q^2(Q-1)\ n^2\right) / \left(\#\mathcal{D}\ Q^2\ n^2\right) = 2m\ (Q-1)$ 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 such that all the couples are identical with $\max_k f_k^*(x) \geq \max_k f_k^{*'}(x)$, this result still holds if one imposes that the values of $i_1(x_0)$ and b_0 are those considered in Lemma 6 ($i_1(x_0) = \mathrm{argmax}_k f_k^*(x_0)$ and $b_0 = f^*_{i_1(x_0)}(x_0) - 1$). Once $i_1(x_0)$ is set, $i_2(x_0)$ can take only $Q-1$ different values. Thus, using once more the pigeonhole principle, among those last couples of functions, there are (at least) $2m\ (Q-1)/(Q-1) = 2m$ 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, i.e. 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 sub-classes of $\mathcal{F}_0^*$ of cardinality at least $2m$, call them $\mathcal{F}_+^*$ and $\mathcal{F}_-^*$, and there are $x_0 \in \mathcal{D}$, two vectors $V_{0,+}$ and $V_{0,-}$ in $\{-n, \dots, n\}^Q$ such that $\|V_{0,+} - V_{0,-}\|_\infty \geq 2$, $(k_0, l_0) \in \{1, \dots, Q\}^2$ with $k_0 \neq l_0$, and a scalar b_0 in $\{-n+1, \dots, n-1\}$ such that:

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

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, $\mathcal{F}_+ = \left\{f_+ \in \mathcal{F} : f_+^* \in \mathcal{F}_+^*\right\}$ strongly N-shatters at least $t(2m, \#\mathcal{D} - 1)$ 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 $\mathcal{F}_- = \left\{f_- \in \mathcal{F} : f_-^* \in \mathcal{F}_-^*\right\}$. 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 $(\{x_0\} \bigcup s_\mathcal{D}, \{(k_0, l_0)\} \bigcup I(s_\mathcal{D}), \bar{v}_b,)$, where $\bar{v}_b$ is deduced from v_b by adding one component corresponding to the point x_0, component taking the value b_0. Indeed, $\mathcal{F}_+$ 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 $\mathcal{F}_-$. 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

$$\forall i \in \{1,\dots,\#s_{\mathcal{D}}\}, \begin{cases} if\ y_i = 1, & f_{+,y,i_1(x_i)}(x_i) - b_i \geq 1 \\ if\ y_i = -1, & f_{+,y,i_2(x_i)}(x_i) + b_i \geq 1 \end{cases}$$

and

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

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

$$\forall i \in \{1,\dots,\#s_{\mathcal{D}}\}, \begin{cases} if\ y_i = 1, & f_{-,y,i_1(x_i)}(x_i) - b_i \geq 1 \\ if\ y_i = -1, & f_{-,y,i_2(x_i)}(x_i) + b_i \geq 1 \end{cases}$$

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 $t\left(2m\ \#\mathcal{D}\ Q^2(Q-1)\ n^2, \#\mathcal{D}\right) \geq 2t(2m, \#\mathcal{D}-1)$ which is precisely (28).

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

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

Applying (28) iteratively and eventually (27), it appears that $t(l, \#\mathcal{D}) \geq 2^r$. Since t is clearly nondecreasing in its first argument, and $2\left(\#\mathcal{D}\ Q^2(Q-1)\ n^2\right)^r \geq l$, this implies

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

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

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

and consequently

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

which is precisely (26). If on the contrary $\lceil \log_2(\phi(d, \#\mathcal{D})) \rceil \geq \#\mathcal{D}$, then

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

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

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

$t\left(2\left(\#\mathcal{D}\,Q^2(Q-1)\,n^2\right)^{\lceil\log_2(\phi(d,\#\mathcal{D}))\rceil},\#\mathcal{D}\right)$ is consequently once more superior to $\phi(d,\#\mathcal{D})$, which completes the proof of (26) and thus concludes the proof of the lemma. ∎

Note that expressing Lemma 7 in the bi-class case (by setting $Q=2$), one obtains almost exactly the expression of Lemma 3.3 in [3], keeping in mind that our functions and theirs do not take their values in the same intervals.

5.5 First upper bound on the covering numbers of $\Delta_\gamma^*\mathcal{H}$

For all $0<\epsilon,\eta\leq\gamma\leq 1$, let

$$\mathcal{M}_{\infty,\infty}(\epsilon,\Delta_\gamma^*\mathcal{H},2m)=\sup_{s_{\mathcal{X}^{2m}}\in\mathcal{X}^{2m}}\mathcal{M}(\epsilon,\Delta_\gamma^*\mathcal{H},d_{\ell_\infty,\ell_\infty(s_{\mathcal{X}^{2m}})})$$

and

$$\mathcal{M}_{\infty,\infty}(2,\left(\Delta_\gamma^*\mathcal{H}\right)^{(\eta)},2m)=\sup_{s_{\mathcal{X}^{2m}}\in\mathcal{X}^{2m}}\mathcal{M}(2,\left(\Delta_\gamma^*\mathcal{H}\right)^{(\eta)},d_{\ell_\infty,\ell_\infty(s_{\mathcal{X}^{2m}})}).$$

Applying Lemma 4 (right-hand side inequality) to $\Delta_\gamma^*\mathcal{H}$ gives:

$$\mathcal{N}^{(p)}_{\infty,\infty}(\gamma/4,\Delta_\gamma^*\mathcal{H},2m)\leq\mathcal{M}_{\infty,\infty}(\gamma/4,\Delta_\gamma^*\mathcal{H},2m).$$

Setting $\epsilon=\gamma/4$ ($\eta=\gamma/12$) in Proposition 2 of Lemma 5, one establishes that:

$$\mathcal{M}_{\infty,\infty}(\gamma/4,\Delta_\gamma^*\mathcal{H},2m)\leq\mathcal{M}_{\infty,\infty}(2,\left(\Delta_\gamma^*\mathcal{H}\right)^{(\gamma/12)},2m).$$

Similarly, the packing numbers of the discretized set of functions can be bounded thanks to Lemma 7, by setting $\#\mathcal{D}=2m$. and $n=\left\lfloor\frac{\gamma}{\eta}\right\rfloor=12$. Thus, we get:

$$\mathcal{M}_{\infty,\infty}(2,\left(\Delta_\gamma^*\mathcal{H}\right)^{(\gamma/12)},2m)<2\left(288\,m\,Q^2(Q-1)\right)^{\lceil\log_2(\phi(d,2m))\rceil}. \tag{29}$$

In the right-hand side of (29), $\phi(d,2m)=\sum_{i=1}^d\binom{2m}{i}\left(23\binom{Q}{2}\right)^i$, where d is the strong Natarajan dimension of $(\Delta_\gamma\mathcal{H})^{(\gamma/12)}$. Since we are interested in upper bounding the capacity measure, one can also make use of Proposition 1 in Lemma 5 to put the Natarajan dimension with margin $\eta/2=\gamma/24$ of $\Delta_\gamma\mathcal{H}$, N-dim$(\Delta_\gamma\mathcal{H},\gamma/24)$, in place of d. Combining all the partial results in this subsection thus produces the following theorem.

Theorem 3 *Let $\mathcal{H}$ be a class of functions from a domain $\mathcal{X}$ into $\mathbb{R}^Q$. For every value of γ in $(0,1]$ and every integer value of m satisfying $2m\geq$ N-dim$(\Delta_\gamma\mathcal{H},\gamma/24)$, the following bound is true:*

$$\mathcal{N}^{(p)}_{\infty,\infty}(\gamma/4,\Delta_\gamma^*\mathcal{H},2m)<2\left(288\,m\,Q^2(Q-1)\right)^{\lceil\log_2(\phi(d,2m))\rceil} \tag{30}$$

where $d=$ N-dim$(\Delta_\gamma\mathcal{H},\gamma/24)$ and $\phi(d,2m)=\sum_{i=1}^d\binom{2m}{i}\left(23\binom{Q}{2}\right)^i$.

5.6 Standard bound on function ϕ

To find an upper bound on $\phi(d, 2m)$, we make use of a famous corollary of Sauer's lemma.

Lemma 8 *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:

$$\Phi(K_1, K_2, K_3) < \left(\frac{K_2 K_3 e}{K_1}\right)^{K_1}, \tag{31}$$

where e is the base of the Neperian 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 [23], $\sum_{i=0}^{K_1} \binom{K_2}{i}$ can be bounded from above by $\left(\frac{K_2 e}{K_1}\right)^{K_1}$, which concludes the proof. ∎

5.7 Main theorem and discussion

To derive the final formula relating the covering numbers of interest to the margin Natarajan dimension of $\Delta_\gamma \mathcal{H}$, it suffices to make use of Lemma 8 with $K_1 = d$, $K_2 = 2m$ and $K_3 = 23\binom{Q}{2}$. This implies that

$$\phi(d, 2m) < \Phi\left(d, 2m, 23\binom{Q}{2}\right) < (23emQ(Q-1)/d)^d$$

and consequently

$$\log_2(\phi(d, 2m)) < d \log_2(23emQ(Q-1)/d).$$

Substituting this last expression in (30), we finally get our master theorem.

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

$$\mathcal{N}_{\infty,\infty}^{(p)}(\gamma/4, \Delta_\gamma^* \mathcal{H}, 2m) < 2\left(288\, m\, Q^2(Q-1)\right)^{\lceil d \log_2(23emQ(Q-1)/d)\rceil} \tag{32}$$

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

To sum up, in this section, we have derived a bound on the covering numbers of interest in terms of a scale-sensitive extension of one of the Ψ-dimensions, the Natarajan dimension. Obviously, such a generalized Sauer-Shelah lemma can be derived in a similar way for other

Ψ-dimensions, such as 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 the following section, the main advantage of the margin Natarajan dimension will appear clearly. 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 Margin Natarajan Dimension of the Multi-class SVMs

Support vector machines (SVMs) are learning systems which have been introduced by Vapnik and co-workers [14, 20] as a nonlinear extension of the maximal margin hyperplane [71]. 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, usually denoted by Φ. Second, the maximal margin hyperplane is computed in that space, to separate the two categories.

6.1 Architecture and training of the M-SVMs

The problem of performing multi-class discriminant analysis with SVMs was initially tackled through decomposition schemes [61, 52, 73, 44, 54, 2, 4, 58]. The multi-class SVMs are globally more recent. They are all obtained by combining a multivariate affine model with the nonlinear mapping Φ into the feature space [73, 77, 17, 32, 21, 22, 30, 39, 47, 48]. Formally, the functions $h = (h_k)_{1 \leq k \leq Q}$ that a Q-category M-SVM can implement have the general form:

$$\forall x \in \mathcal{X},\ \forall k \in \{1, \dots, Q\},\ h_k(x) = \langle w_k, \Phi(x) \rangle + b_k, \tag{33}$$

where the values of the vectors w_k and the scalars b_k are constrained by the content of the training set. As usual, the mapping Φ does not appear explicitely in the computations. Thanks to the "kernel trick", it is replaced with the *reproducing kernel function* κ, a positive type function [12] which computes the ℓ_2 dot product in the feature space. Let $\left(H_\kappa, \langle ., . \rangle_{H_\kappa}\right)$ denote this space. We thus have:

$$\forall (x, x') \in \mathcal{X}^2,\ \kappa(x, x') = \langle \Phi(x), \Phi(x') \rangle_{H_\kappa}. \tag{34}$$

Hence, the "linear part" of each component function of the model is a function of x belonging to the Reproducing Kernel Hilbert Space (RKHS) (see for instance [6, 59, 75, 76, 12]) $\left(H_\kappa, \langle ., . \rangle_{H_\kappa}\right)$ and $\mathcal{H} \subset \left(\left(H_\kappa, \langle ., . \rangle_{H_\kappa}\right) + \{1\}\right)^Q$. Furthermore, κ is supposed to satisfy Mercer's conditions [1].

We now introduce the standard hypotheses on $\mathcal{X}$ and $\mathcal{H}$ which will allow us to formulate the upper bound on the margin Natarajan dimension of interest. We suppose that $\Phi(\mathcal{X})$ is included in the ball of radius $\Lambda_{\Phi(\mathcal{X})}$ in $\left(H_\kappa, \langle ., . \rangle_{H_\kappa}\right)$ and the vectors w_k defining the separating hyperplanes satisfy $1/2 \max_{1 \leq k < l \leq Q} \|w_k - w_l\|_{H_\kappa} \leq \Lambda_w$. Furthermore, the vector of "biases" $b = (b_k)_{1 \leq k \leq Q}$ is supposed to belong to $[-\beta, \beta]^Q$.

6.2 Switching from $\Delta_\gamma^* \mathcal{H}$ to $\Delta^* \bar{\mathcal{H}}$

When $\mathcal{H}$ is the class of functions implemented by a M-SVM, The classes of functions of interest to compute the value of the guaranteed risk, $\Delta_\gamma \mathcal{H}$ and $\Delta_\gamma^* \mathcal{H}$, are difficult to handle due to the presence of two components: the π_γ operator and the vector b. The aim of this section is to get rid of those components, paving the way for the expression of a final result the proof of which will appear as a straightforward extension of the one of its bi-class counterpart. Such transitions are easier to do when working with covering numbers rather than with generalized VC dimensions.

Lemma 9 *Let $\mathcal{H}$ be a class of functions from a domain $\mathcal{X}$ into $\mathbb{R}^Q$, let (γ, ϵ) be a couple of reals satisfying $0 < \epsilon \leq \gamma \leq 1$ and let $m \in \mathbb{N}^*$. Then,*

$$\mathcal{N}^{(p)}_{\infty,\infty}(\epsilon, \Delta^*_\gamma \mathcal{H}, m) \leq \mathcal{N}^{(p)}_{\infty,\infty}(\epsilon, \Delta^* \mathcal{H}, m). \tag{35}$$

Proof This property directly springs from the fact that π_γ satisfies the Lipschitz condition with constant 1. Thus, $\forall\,(h, h') \in \mathcal{H}^2$, $\forall x \in \mathcal{X}$, $\forall (\gamma, \epsilon) \in \mathbb{R}^2 : 0 < \epsilon \leq \gamma \leq 1$,

$$\left\| \Delta^* h(x) - \Delta^* h'(x) \right\|_\infty \leq \epsilon \Longrightarrow \left\| \Delta^*_\gamma h(x) - \Delta^*_\gamma h'(x) \right\|_\infty \leq \epsilon.$$

∎

Lemma 10 *Let $\mathcal{H}$ be the class of functions that a Q-category M-SVM can implement under the hypotheses exposed in Section 6.1. Let $\bar{\mathcal{H}}$ be the subset of $\mathcal{H}$ corresponding to the functions satisfying $b = 0_Q$. Let $\epsilon \in \mathbb{R}^*$ and $m \in \mathbb{N}^*$. Then*

$$\mathcal{N}^{(p)}_{\infty,\infty}(\epsilon, \Delta^* \mathcal{H}, m) \leq \left(2 \left\lceil \frac{\beta}{\epsilon} \right\rceil + 1 \right)^Q \mathcal{N}^{(p)}_{\infty,\infty}(\epsilon/2, \Delta^* \bar{\mathcal{H}}, m). \tag{36}$$

Proof Let $B =$

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

By construction, B^Q is a proper $\epsilon/2$-cover of $[-\beta, \beta]^Q$ with respect to the ℓ_∞ norm. For $s_{\mathcal{X}^m} \in \mathcal{X}^m$, let $\overline{\Delta^* \bar{\mathcal{H}}}$ be a proper $\epsilon/2$-cover of $\Delta^* \bar{\mathcal{H}}$ with respect to the $d_{\ell_\infty, \ell_\infty(s_{\mathcal{X}^m})}$ pseudo-metric. We make the assumption that $\overline{\Delta^* \bar{\mathcal{H}}}$ is of minimal cardinality, that is to say $\# \overline{\Delta^* \bar{\mathcal{H}}} = \mathcal{N}^{(p)}(\epsilon/2, \Delta^* \bar{\mathcal{H}}, d_{\ell_\infty, \ell_\infty(s_{\mathcal{X}^m})})$. Then, due to the triangle inequality, $\overline{\Delta^* \bar{\mathcal{H}}} \times B^Q$ is a proper ϵ-cover of $\Delta^* \mathcal{H}$ with respect to the $d_{\ell_\infty, \ell_\infty(s_{\mathcal{X}^m})}$ pseudo-metric. Since the cardinality of B^Q is $\left(2 \left\lceil \frac{\beta}{\epsilon} \right\rceil + 1 \right)^Q$, this ϵ-cover is of cardinality $\left(2 \left\lceil \frac{\beta}{\epsilon} \right\rceil + 1 \right)^Q \mathcal{N}^{(p)}(\epsilon/2, \Delta^* \bar{\mathcal{H}}, d_{\ell_\infty, \ell_\infty(s_{\mathcal{X}^m})})$. As a consequence, $\mathcal{N}^{(p)}(\epsilon, \Delta^* \mathcal{H}, d_{\ell_\infty, \ell_\infty(s_{\mathcal{X}^m})}) \leq \left(2 \left\lceil \frac{\beta}{\epsilon} \right\rceil + 1 \right)^Q \mathcal{N}^{(p)}(\epsilon/2, \Delta^* \bar{\mathcal{H}}, d_{\ell_\infty, \ell_\infty(s_{\mathcal{X}^m})})$. Taking the supremum of both sides of this inequality over all the possible sets $s_{\mathcal{X}^m}$ thus concludes the proof. ∎

In our will to simplify the computation, we have skipped a difficulty. With the elemination of the operator π_γ, Theorem 4, no longer applies. Note however that the main contribution of the operator π_γ in the derivation of this theorem is to ensure that the classes of functions $\mathcal{F}$ and $\mathcal{F}^*$ involved in the Sauer-Shelah lemma (Lemma 7) have a finite range, and thus contain a finite number of different functions, since their domain is of finite cardinality. In the case of a M-SVM, the hypotheses introduced in Section 6.1, which imply that both $\Delta\mathcal{H}$ and $\Delta^*\mathcal{H}$ have a bounded range, make it possible to derive an equivalent of Lemma 7 involving directly $\left.(\Delta\mathcal{H})^{(\eta)}\right|_{\mathcal{D}}$ and $\left.(\Delta^*\mathcal{H})^{(\eta)}\right|_{\mathcal{D}}$. Thus, with a little work, Theorem 4 can be adapted so as to become compatible with the simplifications introduced in the current section. This raises no difficulty whatsoever.

6.3 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 [10].

Lemma 11 *Let $\bar{\mathcal{H}}$ be the class of functions that a Q-category M-SVM can implement under the hypotheses exposed in Section 6.1 and the additional constraint $b = 0_Q$. Let $\epsilon \in \mathbb{R}_+^*$. If a subset $s_{\mathcal{X}^m} = \{x_i : 1 \leq i \leq m\}$ of $\mathcal{X}$ is N-shattered with margin ϵ by $\Delta\bar{\mathcal{H}}$, then there exists a subset $s_{\mathcal{X}^n}$ of $s_{\mathcal{X}^m}$ of cardinality n at least equal to $\left\lceil \frac{m}{\binom{Q}{2}} \right\rceil$ such that for every partition of $s_{\mathcal{X}^n}$ into two subsets s_1 and s_2, the following bound holds true:*

$$\left\| \sum_{x_i \in s_1} \Phi(x_i) - \sum_{x_i \in s_2} \Phi(x_i) \right\|_{H_\kappa} \geq \frac{\left\lceil \frac{m}{\binom{Q}{2}} \right\rceil}{\Lambda_w} \epsilon. \tag{37}$$

Proof Suppose that $s_{\mathcal{X}^m} = \{x_i : 1 \leq i \leq m\}$ is a subset of $\mathcal{X}$ N-shattered with margin ϵ by $\Delta\bar{\mathcal{H}}$. Let $(I(s_{\mathcal{X}^m}), v_b)$ witness this shattering. 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 $n = \left\lceil \frac{m}{\binom{Q}{2}} \right\rceil$ points in $s_{\mathcal{X}^m}$ 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 $s_{\mathcal{X}^m}$ are reordered in such a way that the n first of them exhibit this property. The corresponding subset of $s_{\mathcal{X}^m}$ is denoted $s_{\mathcal{X}^n}$. This means that for all vector v_y in $\{-1, 1\}^m$, there is a function $\bar{h}_y$ in $\bar{\mathcal{H}}$ characterized by the vectors $w_{y,k}$, $(1 \leq k \leq Q)$, such that:

$$\forall i \in \{1, \dots, n\}, \begin{cases} if\ y_i = \ \ 1, & \Delta\bar{h}_{y,k_0}(x_i) - b_i \geq \epsilon \\ if\ y_i = -1, & \Delta\bar{h}_{y,l_0}(x_i) + b_i \geq \epsilon \end{cases}.$$

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

$$\forall i \in \{1, \dots, n\}, \begin{cases} if\ y_i = \ \ 1, & \frac{1}{2}\langle w_{y,k_0} - \max_{k \neq k_0} w_{y,k}, \Phi(x_i)\rangle - b_i \geq \epsilon \\ if\ y_i = -1, & \frac{1}{2}\langle w_{y,l_0} - \max_{k \neq l_0} w_{y,k}, \Phi(x_i)\rangle + b_i \geq \epsilon \end{cases} \tag{38}$$

and thus implies

$$\forall i \in \{1, \dots, n\}, \begin{cases} if\ y_i = \ \ 1, & \frac{1}{2}\langle w_{y,k_0} - w_{y,l_0}, \Phi(x_i)\rangle - b_i \geq \epsilon \\ if\ y_i = -1, & \frac{1}{2}\langle w_{y,l_0} - w_{y,k_0}, \Phi(x_i)\rangle + b_i \geq \epsilon \end{cases}. \tag{39}$$

Note that this step of the proof does not hold any more if one uses the Δ^* operator in place of the Δ operator. Indeed, reformulating (38) with Δ^* in place of Δ, one cannot derive (39) any more. This is precisely the reason why it is specifically the Δ operator which appears in the hypotheses of Lemma 11 and, by way of consequence, the hypotheses of the final bound on the margin Natarajan dimension (see Theorem 5 below). Consider now any partition of $s_{\mathcal{X}^n}$ into to subsets s_1 and s_2. Consider any vector v_y in $\{-1,1\}^m$ such that $y_i = 1$ if $x_i \in s_1$ and $y_i = -1$ if $x_i \in s_2$. We have thus:

$$\frac{1}{2}\langle w_{y,k_0} - w_{y,l_0}, \sum_{x_i \in s_1} \Phi(x_i)\rangle - \sum_{i:x_i \in s_1} b_i + \frac{1}{2}\langle w_{y,l_0} - w_{y,k_0}, \sum_{x_i \in s_2} \Phi(x_i)\rangle + \sum_{i:x_i \in s_2} b_i \geq \# s_{\mathcal{X}^n}\epsilon$$

which simplifies into

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

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:

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

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

$$\frac{1}{2}\langle w_{y,k_0} - w_{y,l_0}, \sum_{x_i \in s_1} \Phi(x_i) - \sum_{x_i \in s_2} \Phi(x_i)\rangle \geq \left\lceil \frac{m}{\binom{Q}{2}} \right\rceil \epsilon \tag{40}$$

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

$$\frac{1}{2}\langle w_{y,l_0} - w_{y,k_0}, \sum_{x_i \in s_1} \Phi(x_i) - \sum_{x_i \in s_2} \Phi(x_i)\rangle \geq \left\lceil \frac{m}{\binom{Q}{2}} \right\rceil \epsilon. \tag{41}$$

Finally, applying the Cauchy-Schwartz inequality to both (40) and (41), it springs that whatever the sign of $\sum_{i:x_i \in s_1} b_i - \sum_{i:x_i \in s_2} b_i$ is,

$$\frac{1}{2}\left\|w_{k_0} - w_{l_0}\right\|_{H_\kappa} \left\| \sum_{x_i \in s_1} \Phi(x_i) - \sum_{x_i \in s_2} \Phi(x_i) \right\|_{H_\kappa} \geq \left\lceil \frac{m}{\binom{Q}{2}} \right\rceil \epsilon$$

from which (37) directly springs, as a consequence of the constraint $1/2 \max_{1 \leq k < l \leq Q} \|w_k - w_l\|_{H_\kappa} \leq \Lambda_w$. ∎

Lemma 12 (Lemma 4.3 in [10]) *All subset $s_{\mathcal{X}^m} = \{x_i : 1 \leq i \leq m\}$ of $\mathcal{X}$ can be partitioned into two subsets s_1 and s_2 satisfying*

$$\left\| \sum_{x_i \in s_1} \Phi(x_i) - \sum_{x_i \in s_2} \Phi(x_i) \right\|_{H_\kappa} \leq \sqrt{m} \Lambda_{\Phi(\mathcal{X})}. \tag{42}$$

The following theorem is a direct consequence of Lemma 11 and Lemma 12.

Theorem 5 *Let $\bar{\mathcal{H}}$ be the class of functions that a Q-category M-SVM can implement under the hypotheses exposed in Section 6.1 and the additional constraint $b = 0_Q$. Then, for any positive real value ϵ, the following bound holds true:*

$$N\text{-}dim\left(\Delta\bar{\mathcal{H}}, \epsilon\right) \leq \binom{Q}{2} \left(\frac{\Lambda_w \Lambda_{\Phi(\mathcal{X})}}{\epsilon} \right)^2. \tag{43}$$

Proof Let $s_{\mathcal{X}^m}$ be a subset of $\mathcal{X}$ of cardinality m N-shattered with margin ϵ by $\Delta\bar{\mathcal{H}}$. According to Lemma 11, there is at least a subset $s_{\mathcal{X}^n}$ of $s_{\mathcal{X}^m}$ of cardinality $n = \left\lceil \frac{m}{\binom{Q}{2}} \right\rceil$ satisfying (37) for all its partitions into two subsets s_1 and s_2. Since, according to Lemma 12, there is at least one of these partitions for which (42) holds true,

$$\frac{n}{\Lambda_w} \epsilon \leq \sqrt{n} \Lambda_{\Phi(\mathcal{X})}$$

which implies that

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

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

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

which concludes the proof. ∎

Note that in the bi-class case, using the notations of Subsection 4.3, (43) becomes

$$P_\epsilon\text{-dim}\left(\tilde{\mathcal{H}}\right) \leq \left(\frac{\Lambda_w \Lambda_{\Phi(\mathcal{X})}}{\epsilon} \right)^2$$

which is precisely the bound provided by Theorem 4.6 in [10] (see also Remark 1 in [35]). This bound is the tightest bound on the fat-shattering dimension of a linear classifier currently available.

7 Guaranteed Risk and Implementation of the SRM Inductive Principle

In this section, we discuss the significance of the results derived in the two previous sections, with the aim to highlight the similarities and differences existing between the multi-class case and the bi-class one.

7.1 Characterization of relevant information

The main results of this report have involved two distinct margin operators, Δ and Δ^*. Theorem 1, the basic uniform convergence result on which all this study is based, holds true for both of them. However, we pointed out in Subsection 5.3 the reason why Theorem 4 (or Lemma 7), the generalized Sauer-Shelah lemma, requires specifically the use of Δ^*. On the contrary, we have seen in Subsection 6.3 that the proof of Theorem 5, the bound on the margin Natarajan dimension of the M-SVMs, was making use of a specific property of Δ. Fortunately, the connection between the capacities of $\Delta_\gamma \mathcal{H}$ and $\Delta^*_\gamma \mathcal{H}$ is provided by Lemma 6 and Lemma 7. 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 that simply the index of the highest output and the difference between the two highest outputs, i.e. what is relevant to determine both the classification performed and the confidence one can have in the correctness of this classification. This suggests that some improvement could be made regarding either the definition of the scale-sensitive Ψ-dimensions or the choice of the pseudo-metric. However, it is difficult to figure out how these changes could remain compatible with the reasoning followed in Section 6 to derive the bound on the margin Natarajan dimension. Indeed, the choices we made to extend the VC theory to the case of large margin multi-category discriminant models were primarily governed by one concern: allowing a natural extension of the proof of Lemma 3.3 in [3] and the proof of Theorem 4.6 in [10] 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?

7.2 On the theoretical grounding of the M-SVMs of the literature

All the M-SVMs proposed so far can be classified into three categories according to the nature of the control term appearing in their objective function (training criterion). This control term is either $\sum_{k=1}^{Q} \|w_k\|_{H_k}^2$, in [77, 73, 21, 22, 48], $\sum_{k<l}^{Q} \|w_k - w_l\|\|_{H_k}^2$ (with a variant, $\max_{k<l} \|w_k - w_l\|_{H_k}^2$), in [32, 30] or a combination of the two, $\sum_{k<l}^{Q} \|w_k - w_l\|_{H_k}^2 + \sum_{k=1}^{Q} \|w_k\|_{H_k}^2$ in [17]. However, we proved in [30] (see also [39]) that these choices are equivalent, provided that the value of the soft margin parameter C is selected appropriately,

for the following reason. When using $\sum_{k<l}^{Q} \|w_k - w_l\|\|_{H_k}^2$ as control term, the formulation of the primal problem is such that the vectors w_k are only defined up to an additive constant. A natural way to overcome this underdetermination consists in setting $\sum_{k=1}^{Q} w_k = 0$. In that case, $\sum_{k<l}^{Q} \|w_k - w_l\|\|_{H_k}^2 = Q \sum_{k=1}^{Q} \|w_k\|_{H_k}^2$. Thus, Theorem 5 provides a posteriori all the choices listed above with a theoretical justification. They can be derived in the context of the implementation of the SRM inductive principle in exactly the same way as the choice of the term $\|w\|\|_{H_k}^2$ for the bi-class case. This result is interesting in its own right since, strangely enough, such a basic justification was lacking so far. The arguments put forward to justify the specification of the different M-SVMs were primarily dealing with the choice of the loss function characterizing the contribution of the empirical risk. For instance, the advantage of the model introduced in [21, 22] rests in the fact that the constraints of correct classification are expressed in such a way that the formulation of the Wolfe dual problem requires only one variable per example (instead of $Q - 1$ for the other M-SVMs). As for the model of Lee and her co-authors, the loss function is tailored to target the coded class with the maximum conditional probability, thus ensuring the consistency of the learning procedure. From now on, comparing M-SVMs boils down to studying the incidence of the choice of specific loss functions.

8 Conclusions and Future Work

This paper has introduced a new uniform convergence result for the empirical (margin) risk of large margin multi-category discriminant models. The measure of capacity it involves, a covering number, can be upper bounded in terms of different scale-sensitive Ψ-dimensions, thanks to generalized Sauer-Shelah lemmas. It is thus possible to choose the most appropriate of these extended notions of VC dimensions as a function of the model of interest. 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. This contribution to the VC theory of large margin classifiers has thus made it possible to endow all the training algorithms (objective functions) of the M-SVMs proposed so far with a unifying theory. Indeed, the main practical interest of this study should regard the implementation of the SRM inductive principle, through a new characterization of the variation of the capacity of a multivariate affine model as a function of the constraints on its domain and parameters. In that sense, it completes previous works on the same subject [30, 33], which had followed another path, namely the computation of a bound on the entropy numbers of a linear operator [78, 79, 34].

Readers primarily interested in computing sample complexities should be aware of the fact that sharper bounds should result from using different (more recent) sources of inspiration, also even in that case, some results exposed here could still prove useful. An obvious possibility is represented by new tools of concentration theory and empirical processes [66, 67, 46, 51, 50]. 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 [15]. Regarding more specifically pattern recognition SVMs, the results the extension of which appears most promising are those reported in [16, 63, 13]. Performing these extensions is the subject of an ongoing work. We also intend to study the connection between the finiteness of the margin Ψ-dimensions (for all strictly positive values of their parameter γ), and the property of universal consistency [73].

Acknowledgements
It is a pleasure to thank S. Kroon and R. Vert for instructive discussions and bibliographical help. Thanks are also due to E. Domenjoud, A. Elisseeff, E. Monfrini and F. Sur for carefully reading either the initial or the revised version of this manuscript.

References

[1] M. Aizerman, E. Braverman, and L. Rozonoer. Theoretical foundations of the potential function method in pattern recognition learning. *Automation and Remote Control*, 25:821–837, 1964.

[2] 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.

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

[4] C. Angulo, X. Parra, and A. Català. K-SVCR. A support vector machine for multi-class classification. *Neurocomputing*, 55(1–2):57–77, 2003.

[5] M. Anthony and P.L. Bartlett. *Neural Network Learning: Theoretical Foundations*. Cambridge University Press, New York, 1999.

[6] N. Aronszajn. Theory of reproducing kernels. *Transactions of the American Mathematical Society*, 68:337–404, 1950.

[7] P. Barbe and M. Ledoux. *Probabilité*. De la licence à l'agrégation. Belin, 1998.

[8] 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.

[9] 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.

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

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

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

[13] G. Blanchard, O. Bousquet, and P. Massart. Statistical performance of support vector machines. *The Annals of Statistics*, 2004. (submitted).

[14] B. Boser, I. Guyon, and V. Vapnik. A training algorithm for optimal margin classifiers. In *COLT'92*, pages 144–152, 1992.

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

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

[17] E.J. Bredensteiner and K.P. Bennett. Multicategory Classification by Support Vector Machines. *Computational Optimization and Applications*, 12(1/3):53–79, 1999.

[18] H. Brezis. *Analyse fonctionnelle, Théorie et applications*. MASSON, 1993.

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

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

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

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

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

[24] R.M. Dudley. Central limit theorems for empirical measures. *Annals of Probability*, 6:899–929, 1978.

[25] R.M. Dudley. Universal Donsker classes and metric entropy. *Ann. Probab.*, 15(4):1306–1326, 1987.

[26] R.M. Dudley. *Uniform Central Limit Theorems*. Cambridge University Press, Cambridge, UK, 1999.

[27] 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).

[28] J. Gapaillard. *Intégration pour la licence*. Masson, 1997.

[29] E. Giné and A. Guillou. Rates of strong uniform consistency for multivariate kernel density estimators. *Annales de l'Institut Henri Poincaré (B) Probability and Statistics*, 38(6):907–921, 2002.

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

[31] Y. Guermeur, A. Elisseeff, and H. Paugam-Moisy. Estimating the sample complexity of a multi-class discriminant model. In *ICANN'99*, pages 310–315. IEE, 1999.

[32] Y. Guermeur, A. Elisseeff, and H. Paugam-Moisy. A new multi-class SVM based on a uniform convergence result. In *IJCNN'00*, volume IV, pages 183–188, 2000.

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

[34] Y. Guo, P.L. Bartlett, J. Shawe-Taylor, and R.C. Williamson. Covering numbers for support vector machines. *IEEE Transactions on Information Theory*, 48(1):239–250, 2002.

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

[36] D. Haussler. Decision theoretic generalizations of the PAC model for neural net and other learning applications. *Information and Computation*, 100:78–150, 1992.

[37] D. Haussler and P.M. Long. A Generalization of Sauer's Lemma. *Journal of Combinatorial Theory, Series A*, 71:219–240, 1995.

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

[39] C.-W. Hsu and C.-J. Lin. A comparison of methods for multi-class support vector machines. *IEEE Transactions on Neural Networks*, 13(2):415–425, 2002.

[40] K. Jogdeo and S.M. Samuels. Monotone Convergence of Binomial Probabilities and a Generalization of Ramanujan's equation. *Ann. Math. Statist.*, 39(4):1191–1195, 1968.

[41] M.J. Kearns and R.E. Schapire. Efficient distribution-free learning of probabilistic concepts. In *Proceedings of the 31st Annual Symposium on Foundations of Computer Science*, volume 1, pages 382–391. IEEE Computer Society Press, 1990.

[42] 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.

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

[44] U. Kreßel. Pairwise classification and support vector machines. In B. Schölkopf, C.J.C. Burges, and A. Smola, editors, *Advances in Kernel Methods, Support Vector Learning*, pages 255–268. The MIT Press, Cambridge, 1999.

[45] R.S. Kroon. Support vector machines, generalization bounds, and transduction. Master's thesis, University of Stellenbosch, South Africa, December 2003.

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

[47] Y. Lee. Multicategory support vector machines, theory, and application to the classification of microarray data and satellite radiance data. Technical Report 1063, University of Wisconsin, Madison, Department of Statistics, 2002.

[48] 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.

[49] T. Leighton and C.G. Plaxton. Hypercubic sorting networks. *SIAM J. Comput*, 27(1):1–47, 1998.

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

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

[52] E. Mayoraz and E. Alpaydin. Support Vector Machines for Multi-Class Classification. Technical Report 98-06, IDIAP, 1998.

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

[54] J.C. Platt, N. Cristianini, and J. Shawe-Taylor. Large margin DAGs for multiclass classification. In *NIPS'12*, pages 547–553, 2000.

[55] D. Pollard. *Convergence of stochastic processes*. Springer-Verlag, N.Y., 1984.

[56] D. Pollard. Empirical processes: Theory and applications. In *NFS-CBMS Regional Conference Series in Probability and Statistics*, volume 2. Institute of Math. Stat. and Am. Stat. Assoc., 1990.

[57] M.D. Richard and R.P. Lippmann. Neural network classifiers estimate bayesian a posteriori probabilities. *Neural Computation*, 3:461–483, 1991.

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

[59] S. Saitoh. *Theory of Reproducing Kernels and its Applications*. Longman Scientific & Technical, Harlow, England, 1988.

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

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

[62] B. Schölkopf and A.J. Smola. *Learning with Kernels, Support Vector Machines, Regularization, Optimization and Beyond.* The MIT Press, 2002.

[63] C. Scovel and I. Steinwart. Fast rates for support vector machines. Technical Report LA-UR 03-9117, Los Alamos National Laboratory, 2004.

[64] J. Shawe-Taylor, P.L. Bartlett, R.C. Williamson, and M. Anthony. Structural Risk Minimization over Data-Dependent Hierarchies. Technical Report NC-TR-96-053, NeuroCOLT, 1996.

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

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

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

[68] L. Valiant. A theory of the learnable. *Comm. ACM*, 27(11):1134–1142, 1984.

[69] A.W. van der Vaart. *Asymptotic Statistics.* Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 1998.

[70] 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, Inc., 1996.

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

[72] V.N. Vapnik. Inductive principles of the search for empirical dependencies. In *Proceedings of the 2nd Annual Workshop on Computational Learning Theory*, pages 3–21, 1989.

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

[74] 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.

[75] G. Wahba. Spline models for observational data. In *SIAM*, volume 59 of *CBMS-NSF Regional Conference Series in Applied Mathematics*, 1990.

[76] 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*, pages 69–88. The MIT Press, 1999.

[77] 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.

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

[79] R.C. Williamson, A.J. Smola, and B. Schölkopf. Generalization Performance of Regularization Networks and Support Vector Machines *via* Entropy Numbers of Compact Operators. *IEEE Transactions on Information Theory*, 47(6):2516–2532, 2001.

Contents

Unité de recherche INRIA Lorraine
LORIA, Technopôle de Nancy-Brabois - Campus scientifique
615, rue du Jardin Botanique - BP 101 - 54602 Villers-lès-Nancy Cedex (France)

Unité de recherche INRIA Futurs : Parc Club Orsay Université - ZAC des Vignes
4, rue Jacques Monod - 91893 ORSAY Cedex (France)
Unité de recherche INRIA Rennes : IRISA, Campus universitaire de Beaulieu - 35042 Rennes Cedex (France)
Unité de recherche INRIA Rhône-Alpes : 655, avenue de l'Europe - 38334 Montbonnot Saint-Ismier (France)
Unité de recherche INRIA Rocquencourt : Domaine de Voluceau - Rocquencourt - BP 105 - 78153 Le Chesnay Cedex (France)
Unité de recherche INRIA Sophia Antipolis : 2004, route des Lucioles - BP 93 - 06902 Sophia Antipolis Cedex (France)

Éditeur
INRIA - Domaine de Voluceau - Rocquencourt, BP 105 - 78153 Le Chesnay Cedex (France)
http://www.inria.fr
ISSN 0249-6399