# Large Margin Multi-category Discriminant Models and Scale-sensitive Psi-dimensions

Yann Guermeur

INRIA

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

# *Large Margin Multi-category Discriminant Models and Scale-sensitive $\Psi$-dimensions*

Yann Guermeur

**N° 5314**

September 24, 2004

Thème BIO


ISSN 0249-6399 ISRN INRIA/RR--5314--FR+ENG

INRIA
LORRAINE

# Large Margin Multi-category Discriminant Models and Scale-sensitive $\Psi$-dimensions

Yann Guermeur*

Thème BIO — Systèmes biologiques
Projet MODBIO

Rapport de recherche n° 5314 — September 24, 2004 — 47 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 $\Psi$-dimensions are introduced. The covering number of interest can be bounded in terms of these dimensions through extended Sauer's lemmas, as is illustrated in the specific case of the scale-sensitive Natarajan dimension. 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

* CNRS

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 $\Psi$-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 $\Psi$-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. 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 [68] is the theory of bounds, which is at the origin of the structural risk minimization (SRM) inductive principle [66, 58] 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 [69], were dealing with the computation of dichotomies with binary-valued functions. Later on, several studies were devoted to the case of multi-class $\{1, \ldots, Q\}$-valued classifers [9], and large margin classifiers computing dichotomies [8]. 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 [27].

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 [22, 26]. 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 $\{1, \ldots, Q\}$-valued classifers. To that end, we first introduce a new extension of Bartlett's famous theorem on the sample complexity of large margin classifiers [6]. 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 number appearing in the confidence interval of the uniform convergence result. The corresponding generalized Sauer's lemma is established in the particular case of the margin Natarajan dimension. This dimension is then computed for the architecture shared by all the multi-class SVMs (M-SVMs) proposed so far, which makes it possible to compare their objective functions on a theoretical basis.

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 number of interest to the margin Natarajan dimension is established in Section 5. Section 6 is devoted to the computation of the margin Natarajan dimension of the architecture shared by all the M-SVMs. At last, Section 7 deals with the synthesis which can be done of the results derived in the preceding sections. It specifically addresses the question of their utility to select an objective function in the framework of the implementation of the structural risk minimization inductive principle.

# 2 Margin Risk for Multi-category Discriminant Models

In this section, the theoretical framework of the study is introduced. It is based on an extended definition of the notion of margin.

## 2.1 Formalization of the learning problem

We consider the case of a $Q$-category pattern recognition problem, where $Q \geq 3$. Let $\mathcal{X}$ be the space of description and $\mathcal{C} = \{C_1, \ldots, C_k, \ldots, C_Q\}$ the set of categories. We make the assumption, standard in statistical learning theory, that there is a joint probability distribution $F$, fixed but unknown, on $\mathcal{X} \times \mathcal{C}$. Our goal is to find, in a given set $\mathcal{H}$ of functions $h = [h_k]$ from $\mathcal{X}$ into $\mathbb{R}^Q$, a function with the lowest "error rate" on the problem of interest. The "error rate" of a function $h$ is the error rate or *risk* of the corresponding discrimination function, obtained by assigning each pattern $x$ to the category $C_k$ in $\mathcal{C}$ satisfying: $h_k(x) = \max_l h_l(x)$. This discriminant function, hereafter denoted by $f$, 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 [52]), 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 [21], Section 5.3 or [24]). Hereafter, $C(x_i)$ will denote indifferently the category of pattern $x_i$, or the index of this category. Furthermore, $y_i$ will be the canonical coding of this category in $\{-1,1\}^Q$ vector. Precisely, $y_i = [y_{ik}]$, $(1 \leq k \leq Q)$, with $y_{ik} = 1$ if and only if $x_i \in C_k$. To simplify notations, when no confusion is possible, the labels of the categories will be identified with their indexes, i.e. $k$ could be used in place of $C_k$. $\mathcal{Y}$ will be the set of canonical codings of the categories in $\{-1,1\}^Q$ vectors. $\mathcal{S}$ will designate indifferently the two product spaces $\mathcal{X} \times \mathcal{C}$ and $\mathcal{X} \times \mathcal{Y}$.

## 2.2 Multi-class margin

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 $\mathcal{C}$ is the probability that $f(x) \neq C(x)$ for a labelled example $(x, C(x))$ chosen randomly according to $F$, i.e.:*

$$R(f) = \mathbb{P}\{(x,k) : f(x) \neq k\} = \int_{\mathcal{X}\times\mathcal{C}} \mathbb{I}\{f(x) \neq k\}\, dF(x,k) \tag{1}$$

*where $\mathbb{I}$ 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* $s_m = \{(x_1, C(x_1)), \ldots, (x_i, C(x_i)), \ldots, (x_m, C(x_m))\}$ *be a set of* $m$ *elements in* $\mathcal{X} \times \mathcal{C}$. *The* empirical risk *of* $f$ *on* $s_m$ *is defined as:*

$$R_{s_m}(f) = \frac{1}{m} \left| \{(x_i, C(x_i)) \in s_m : f(x_i) \neq C(x_i)\} \right| \tag{2}$$

As stated above, the expected risk (resp. empirical risk) of a function $h$ from $\mathcal{X}$ to $\mathbb{R}^Q$ 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 difference between the output associated with the category of a pattern and the highest output of different index. To take this phenomenon into account, the following operator is introduced:

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

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

*Let* $M(h, x) = \frac{1}{2} \max_k \{h_k(x) - \max_{l \neq k} h_l(x)\}$.

$$\forall k \in \{1, \ldots, Q\}, \quad \Delta^* h_k(x) = \begin{cases} M(h,x) & if \ \frac{1}{2}\{h_k(x) - \max_{l \neq k} h_l(x)\} = M(h,x) \\ -M(h,x) & otherwise \end{cases} \tag{3}$$

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

**Definition 4 (Margin risk)** *Let* $\mathcal{H}$ *be a set of functions from* $\mathcal{X}$ *into* $\mathbb{R}^Q$ *and* $\gamma$ *be a real belonging to the interval* $(0, 1]$. *The* margin *risk with margin* $\gamma$ *of a function* $h$ *of* $\mathcal{H}$ *is defined as:*

$$R_\gamma(h) = \mathbb{P}\{(x, k) : \Delta^* h_k(x) < \gamma\} = \int_{\mathcal{S}} \mathbb{I}\{\Delta^* h_k(x) < \gamma\} \, dF(x, k) \tag{4}$$

The empirical margin is defined accordingly.

**Definition 5 (Empirical margin risk)** *The empirical risk with margin* $\gamma \in (0, 1]$ *of* $h$ *on a set* $s_m$ *of size* $m$ *is*

$$R_{\gamma, s_m}(h) = \frac{1}{m} \left| \left\{(x_i, C(x_i)) \in s_m : \Delta^* h_{C(x_i)}(x_i) < \gamma\right\} \right| \tag{5}$$

*or equivalently*

$$R_{\gamma, s_m}(h) = \frac{1}{m} \left| \left\{ (x_i, y_i) \in s_m : \min_k \Delta^* h_k(x_i) y_{ik} < \gamma \right\} \right| \tag{6}$$

Note that in [16, 17], the authors implicitly make use of the same notion of margin to derive the training algorithm of their "multiclass kernel-based vector machines". The confidence interval that will be added to this empirical risk to bound from above, with high probability, the risk, involves a covering number as capacity measure.

## 2.3 Capacity measure: covering number

The notion of covering number is based on the notion of $\epsilon$-cover[1]

**Definition 6 ($\epsilon$-cover or $\epsilon$-net)** *Let $(E, \rho)$ be a pseudo-metric space, and $B(v, r)$ the open ball of center $v$ and radius $r$ in $E$. Let $H$ be a subset of $E$. An* $\epsilon$-cover *of $H$ is a subset $\bar{H}$ of $E$ such that:*

$$H \subset \bigcup_{v \in \bar{H}} B(v, \epsilon)$$

**Definition 7 (Covering numbers)** *Let $(E, \rho)$ be a pseudo-metric space. If $H \subset E$ has an $\epsilon$-cover of finite cardinality, then its* covering number $\mathcal{N}(\epsilon, H, \rho)$ *is the smallest cardinality of its $\epsilon$-covers. If there is no such finite cover, then the covering number is defined to be $\infty$.*

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

**Definition 8** *Let $\mathcal{H}$ be a set of functions from $\mathcal{X}$ into $\mathbb{R}^Q$. For a set $s$ of points in $\mathcal{X}$ of finite cardinality, define the pseudo-metric $d_{l_\infty, l_\infty(s)}$ on $\mathcal{H}$ as:*

$$\forall (h, \bar{h}) \in \mathcal{H}^2,\ d_{l_\infty, l_\infty(s)}(h, \bar{h}) = \max_{x \in s} \max_{k \in \{1, \ldots, Q\}} \left| h_k(x) - \bar{h}_k(x) \right| \tag{7}$$

For technical reasons, linked in particular to the computation of the upper bound on the covering number, it is useful to bound the values taken by the components of the functions $\Delta^* h$ in the interval $[-\gamma, \gamma]$, where $\gamma$ is the parameter of the margin risk. This is achieved by application of the $\pi_\gamma$ operator [6].

**Definition 9 ($\pi_\gamma$ operator)** *Let $\mathcal{H}$ be a set of functions from $\mathcal{X}$ into $\mathbb{R}^Q$. For $\gamma \in (0, 1]$, let $\pi_\gamma : h = [h_k] \mapsto \pi_\gamma(h) = [\pi_\gamma(h_k)]$ be the piecewise-linear squashing operator defined as:*

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

Note that $\pi_\gamma$ is also a projection operator. In the sequel, $\Delta^*_\gamma$ will designate $\pi_\gamma \circ \Delta^*$, once more a projection operator. Furthermore, $\Delta^*_\gamma \mathcal{H}$ will represent the set of functions $\left\{ \Delta^*_\gamma h : h \in \mathcal{H} \right\}$. For the sake of simplicity, but without loss of generality, the sole covers of $\Delta^*_\gamma \mathcal{H}$ which will be considered in what follows will be subsets of $\Delta^*_\gamma \mathcal{H}$. $\mathcal{N}_{\infty,\infty}(\epsilon, \Delta^*_\gamma \mathcal{H}, m)$ will denote $\max_{s_m \in \mathcal{X}^m} \mathcal{N}(\epsilon, \Delta^*_\gamma \mathcal{H}, d_{l_\infty, l_\infty(s_m)})$.

[1] Introductions to the basic notions of functional analysis used in this document can be found in [14, 18, 65].

# 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 drawn independently from $F$. With probability at least $1-\delta$, for every value of $\gamma$ in $(0,1]$, the risk $R(h)$ of a function $h$ computed by a numerical $Q$-class discriminant model $\mathcal{H}$ is bounded from above by:*

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

This theorem can be seen as an extension of Corollary 9 in [6], Theorem 4.1 in [68], and more generally an extension of the Glivenko-Cantelli theorem (see for instance [50, 18, 68, 64]). Its proof is divided into several steps, following the structure proposed in [19, 50, 56].

## 3.1 First symmetrization

In this first step of the proof, standard techniques are used to replace the problem of matching the empirical measure $R_{\gamma,s_m}$ against the distribution $R$ with the problem of matching $R_{\gamma,s_m}$ 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 [68] (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:

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

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

Applying Fubini's theorem for nonnegative measurable functions [23] to the product measure $\mathbb{P}_{s_m,\tilde{s}_m}$ yields:

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

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

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

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

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

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

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

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

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

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

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

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

As a consequence

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

Furthermore

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

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 [36] (see also Appendix B in [44]). 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 [50, 18]). 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

In order to introduce the main lemma of this subsection, we must first highlight basic properties of the margin risk.

**Lemma 3** *Let $\mathcal{H}$ be a set of functions from $\mathcal{X}$ into $\mathbb{R}^Q$ and $\gamma$ and $\zeta$ be two real values such that $0 < \zeta \leq \gamma \leq 1$. Then*

$$\forall h \in \mathcal{H},\ R_\zeta(h) = R_\zeta(\Delta^* h) = R_\zeta(\Delta^*_\gamma h) \tag{13}$$

**Proof** The left part of the equation directly results from the fact that $\Delta^*$ is a projection operator. The right part of the equation springs from the fact that the use of the operator $\pi_\gamma$ has no incidence on the computation of the margin risk if the margin is inferior or equal to $\gamma$. ∎

This lemma, which justifies *a posteriori* the specification of the operator $\Delta^*$, will prove useful in conjunction with the choice to restrict the search for covers of $\Delta^*_\gamma\mathcal{H}$ to the set of subsests of $\Delta^*_\gamma\mathcal{H}$. Indeed, it will make it possible to associate each such cover with a subset of $\mathcal{H}$.

**Lemma 4** *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 $\gamma/2$-cover of the set $\Delta^*_\gamma\mathcal{H}$ with respect to the pseudo-metric $d_{l_\infty,l_\infty(s_{2m})}$ (satisfying $\overline{\Delta^*_\gamma\mathcal{H}}(s_{2m}) \subset \Delta^*_\gamma\mathcal{H}$). This*

*cover is supposed to be of minimal cardinality, i.e.* $\left|\overline{\Delta_\gamma^* \mathcal{H}}(s_{2m})\right| = \mathcal{N}(\gamma/2, \Delta_\gamma^* \mathcal{H}, d_{l_\infty, l_\infty(s_{2m})})$. *Let* $\overline{\mathcal{H}}(s_{2m})$ *be a subset of* $\mathcal{H}$ *of cardinality* $\left|\overline{\Delta_\gamma^* \mathcal{H}}(s_{2m})\right|$ *the image of which by* $\Delta_\gamma^*$ *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*

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

$$\mathbb{P}_{s_m, \tilde{s}_m}\left(\sup_{\overline{h} \in \overline{\mathcal{H}}(s_{2m})} \left(R_{\gamma/2, \tilde{s}_m}\left(\overline{h}\right) - R_{\gamma/2, s_m}\left(\overline{h}\right)\right) \geq \epsilon - \frac{1}{m}\right) \tag{14}$$

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

$$\left\{\begin{array}{l} \min_k \Delta^* h_k(\tilde{x}_i)\tilde{y}_{ik} < 0 \\ d_{l_\infty, l_\infty(s_{2m})}(\Delta_\gamma^* h, \Delta_\gamma^* \overline{h}) \leq \frac{\gamma}{2} \end{array}\right. \Longrightarrow \left\{\begin{array}{l} \min_k \Delta_\gamma^* h_k(\tilde{x}_i)\tilde{y}_{ik} < 0 \\ d_{l_\infty, l_\infty(s_{2m})}(\Delta_\gamma^* h, \Delta_\gamma^* \overline{h}) \leq \frac{\gamma}{2} \end{array}\right.$$

$$\left\{\begin{array}{l} \min_k \Delta_\gamma^* h_k(\tilde{x}_i)\tilde{y}_{ik} < 0 \\ d_{l_\infty, l_\infty(s_{2m})}(\Delta_\gamma^* h, \Delta_\gamma^* \overline{h}) \leq \frac{\gamma}{2} \end{array}\right. \Longrightarrow \min_k \Delta_\gamma^* \overline{h}_k(\tilde{x}_i)\tilde{y}_{ik} < \frac{\gamma}{2} \Longrightarrow \min_k \Delta^* \overline{h}_k(\tilde{x}_i)\tilde{y}_{ik} < \frac{\gamma}{2} \tag{15}$$

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

$$\left\{\begin{array}{l} \min_k \Delta^* \overline{h}_k(x_i) y_{ik} < \frac{\gamma}{2} \\ d_{l_\infty, l_\infty(s_{2m})}(\Delta_\gamma^* h, \Delta_\gamma^* \overline{h}) \leq \frac{\gamma}{2} \end{array}\right. \Longrightarrow \left\{\begin{array}{l} \min_k \Delta_\gamma^* \overline{h}_k(x_i) y_{ik} < \frac{\gamma}{2} \\ d_{l_\infty, l_\infty(s_{2m})}(\Delta_\gamma^* h, \Delta_\gamma^* \overline{h}) \leq \frac{\gamma}{2} \end{array}\right.$$

$$\left\{\begin{array}{l} \min_k \Delta_\gamma^* \overline{h}_k(x_i) y_{ik} < \frac{\gamma}{2} \\ d_{l_\infty, l_\infty(s_{2m})}(\Delta_\gamma^* h, \Delta_\gamma^* \overline{h}) \leq \frac{\gamma}{2} \end{array}\right. \Longrightarrow \min_k \Delta_\gamma^* h_k(x_i) y_{ik} < \gamma \Longrightarrow \min_k \Delta^* h_k(x_i) y_{ik} < \gamma \tag{16}$$

From (15) it springs that if $d_{l_\infty, l_\infty(s_{2m})}(\Delta_\gamma^* h, \Delta_\gamma^* \overline{h}) \leq \frac{\gamma}{2}$, then

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

Similarly, from (16) it springs that if $d_{l_\infty, l_\infty(s_{2m})}(\Delta_\gamma^* h, \Delta_\gamma^* \overline{h}) \leq \frac{\gamma}{2}$, then

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

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

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

With this last inequality at hand, (14) then directly results from taking the suprema over $\mathcal{H}$ and $\overline{\mathcal{H}}(s_{2m})$ respectively. ∎

Lemma 4 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 $\mathcal{N}_{\infty,\infty}(\gamma/2, \Delta_\gamma^* \mathcal{H}, 2m)$. This will be exploited to apply a standard union bound.

### 3.3 Maximal inequality

To bound from above the right-hand side of (14), we introduce an auxiliary step of randomization. To that end, let us consider a set $\mathfrak{S}$ of permutations $\sigma$ over $\{1,\ldots,2m\}$. For every sample $s_{2m} = (s_m, \tilde{s}_m) \in S^{2m}$, $s^\sigma_{2m} = (s^\sigma_m, \tilde{s}^\sigma_m) = \left\{(x_{\sigma(1)}, y_{\sigma(1)}), \ldots, (x_{\sigma(2m)}, y_{\sigma(2m)})\right\}$ denotes its "range" by $\sigma$. 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 (14) is not affected by a permutation $\sigma$. One thus obtains:

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

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

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

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

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

Using a uniform distribution over $\mathfrak{S}$, this simplifies as follows:

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

$$\int_{\mathcal{S}^{2m}} \mathbb{P}_\sigma\left(\sup_{\overline{h}\in\overline{\mathcal{H}}(s_{2m})}\left(R_{\gamma/2,\tilde{s}^\sigma_m}\left(\overline{h}\right) - R_{\gamma/2,s^\sigma_m}\left(\overline{h}\right)\right) \geq \epsilon - \frac{1}{m}\right) dF(s_{2m}) \tag{17}$$

We now concentrate on the event

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

corresponding to a given sample $s_{2m}$. It is equal to

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

For all $\overline{h}$ in $\overline{\mathcal{H}}(s_{2m})$, let

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

The right-hand side of (17) is equal to

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

By application of the union bound,

$$\mathbb{P}_\sigma \left( \bigcup_{\overline{h} \in \overline{\mathcal{H}}(s_{2m})} E\left(\epsilon, \overline{h}, 2m\right) \right) \leq \sum_{\overline{h} \in \overline{\mathcal{H}}(s_{2m})} \mathbb{P}_\sigma \left( E\left(\epsilon, \overline{h}, 2m\right) \right) \tag{18}$$

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 $\left(\mathbb{I}\left\{\min_k \Delta^* \overline{h}_k(x_i) y_{ik} < \gamma/2\right\}\right)$, $(1 \leq i \leq m)$, (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

$$\mathbb{P}_\sigma \left( E\left(\epsilon, \overline{h}, 2m\right) \right) = \mathbb{P}\left( \frac{1}{m} \sum_{i=1}^{m} \alpha_i \left( \tilde{\xi}_i - \xi_i \right) \geq \epsilon - \frac{1}{m} \right) \tag{19}$$

where the coefficients $\alpha_i$, $(1 \leq i \leq m)$, are chosen independently and uniformly on $\{-1, 1\}$. To bound from above the right-hand side of (19), an exponential bound can be applied.

## 3.4 Exponential bound

Hoeffding's inequality (see for example [34, 50]) is a consequence of Chernoff's inequality [45].

**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 $\eta > 0$,*

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

Applying this bound to the right-hand side of (19) gives:

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

By substitution into the right-hand side of (18) we get:

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

From (17) it then springs that:

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

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

The right-hand side is simply equal to

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

By definition, $\mathcal{N}_{\infty,\infty}(\gamma/2,\Delta_\gamma^*\mathcal{H},2m)\geq\mathbb{E}\left(\mathcal{N}(\gamma/2,\Delta_\gamma^*\mathcal{H},d_{l_\infty,l_\infty(s_{2m})})\right)$. Thus, applying Lemma 4, we get

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

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

and finally, by application of Lemma 1,

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

$$2\mathcal{N}_{\infty,\infty}(\gamma/2,\Delta_\gamma^*\mathcal{H},2m)\exp\left(-\frac{m}{2}\left(\epsilon-\frac{1}{m}\right)^2\right) \tag{20}$$

Setting the right-hand side of (20) to $\delta$ and solving for $\epsilon$ 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*

$$R(h)\leq R_{\gamma,s_m}(h)+\sqrt{\frac{2}{m}\left(\ln(2\mathcal{N}_{\infty,\infty}(\gamma/2,\Delta_\gamma^*\mathcal{H},2m))-\ln(\delta)\right)}+\frac{1}{m} \tag{21}$$

### 3.5 Uniform bound over the margin $\gamma$

Making use of the proposition above requires to specify the quantity $\gamma$ in advance. As pointed out by Bartlett in [6], 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 [58]). 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 [68]. Fortunately, this difficulty can be overcome thanks to the following proposition, proved in [6], and extended in [40], which allows us to give a result that stands uniformly for all values of the margin $\gamma$ in the interval $(0, 1]$.

**Proposition 2 (Bartlett, Proposition 8 in [6])** *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 $\Phi$ as follows:

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

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 $\alpha$ in $(0, 1]$ and all $\delta$ 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 measurables.

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,

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

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

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

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

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

and finally, by definition of $\Phi$,

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

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 3. In earlier works on the generalization capabilities of multi-class discriminant models, we used a slightly different definition of this operator, namely:

**Definition 10 ($\Delta$ operator [25])** *Define $\Delta$ as an operator on $\mathcal{H}$ such that:*

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

$$\forall k \in \{1, \ldots, Q\}\,, \quad \Delta h_k(x) = \frac{1}{2}\left\{ h_k(x) - \max_{l \neq k} h_l(x) \right\} \tag{22}$$

It is easy to check that the proof of Theorem 1 still holds if one substitutes $\Delta$ to $\Delta^*$. 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 number of interest. This question, the nature of which is primarily technical, will be addressed in the following sections. At this point, we can already notice that the $\Delta^*$ operator provides less information on the behaviour of the function on which it is applied than the $\Delta$ operator. This would appear as an advantage to derive a generalized Sauer's lemma, and a drawback to compute an upper bound on the corresponding generalized VC dimension. As a consequence, an efficient approach could result from using the two operators at different steps of the computations. Obviously the difficulty of such a strategy rests in performing the connection between both types of bounds (those involving $\Delta$ and those involving $\Delta^*$).

# 4 Scale-sensitive Ψ-dimensions

Several approaches can be applied to bound from above the covering number of interest for a given family of functions $\mathcal{H}$. In this report, we focus on the standard pathway, in which the covering number is first related to an extended notion of Vapnik-Chervonenkis (VC) dimension [69], 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-Shela lemma [69, 54, 59]. 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-Shela lemmas have been derived (see for instance [33, 3]), which involve extended notions of VC dimension. For large margin bi-class discriminant models, the generalization of the VC dimension which gave birth to the richest set of theoretical results is a scale-sensitive variant called the fat-shattering dimension [37, 38]. In the multi-class case, several alternative solutions were proposed by different authors, such as the graph dimension [20, 48], or the Natarajan dimension [48]. It was proved in [9] 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 [63]. 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 real-valued discriminant models, scale-sensitive extensions of the Ψ-dimensions should make it possible to study the generalization capabilities of discriminant models taking their values in $\mathbb{R}^Q$.

## 4.1 Ψ-dimensions

**Definition 11 (Ψ-shattering [9])** *Let $\mathcal{F}$ be a set of functions on a set $\mathcal{X}$ taking their values in the finite set $\{1,\ldots,Q\}$. Let $\Psi$ be a family of mappings $\psi$ from $\{1,\ldots,Q\}$ into $\{-1,1,*\}$, where $*$ is thought of as a null element. A subset $s_m = \{x_i\}$, $(1 \leq i \leq m)$ of $\mathcal{X}$ is said to be* Ψ-shattered *by $\mathcal{F}$ if there is a mapping $\psi^m = \left(\psi^{(1)},\ldots,\psi^{(i)},\ldots,\psi^{(m)}\right)$ in $\Psi^m$ such that for each vector $v_y$ of $\{-1,1\}^m$, there is a function $f_y$ in $\mathcal{F}$ satisfying*

$$\left[\psi^{(1)} \circ f_y(x_1),\ldots,\psi^{(i)} \circ f_y(x_i),\ldots,\psi^{(m)} \circ f_y(x_m)\right]^T = v_y$$

**Definition 12 (Ψ-dimension [9])** *Let $\mathcal{F}$ and $\Psi$ 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 operators from $\{1,\ldots,Q\}$ 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 choice of one particular dimension (set $\Psi$) utterly rests on the possibility to derive two tight bounds: a generalized Sauer-Shela lemma and a bound on the dimension itself. The most frequently used $\Psi$-dimension is the graph dimension, defined as follows:

**Definition 13 (Graph dimension [20, 48])** *Let $\mathcal{F}$ be a set of functions on a set $\mathcal{X}$ taking their values in a countable set. For any $f \in \mathcal{F}$, the* graph $\mathcal{G}$ *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 $\Psi$-dimension is the following:

**Definition 14 (Graph dimension)** *Let $\mathcal{F}$ be a set of functions on a set $\mathcal{X}$ taking their values in $\{1, \ldots, Q\}$. The* graph dimension *of $\mathcal{F}$ is the $\Psi$-dimension of $\mathcal{F}$ in the specific case where $\Psi$ is the set of $Q$ mappings $\psi_k$, $(1 \leq k \leq Q)$, such that $\psi_k$ 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 $\psi_k$ are the indicator functions of the categories.*

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

**Definition 15 (Natarajan dimension [48])** *Let $\mathcal{F}$ be a set 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 $\Psi$-dimension of $\mathcal{F}$ in the specific case where $\Psi$ is the set of $\binom{Q}{2}$ mappings $\psi_{k,l}$, $(1 \leq k < l \leq Q)$, 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 $\Psi$-dimensions

Our scale-sensitive version of the concept of $\Psi$-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 16 (Vapnik dimension [67])** *Let $\mathcal{H}$ be a set of real-valued functions on a set $\mathcal{X}$. A subset $s_m = \{x_i\}$, $(1 \leq i \leq m)$ 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*

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

*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 17 (Pollard's pseudo-dimension [51, 32])** *Let $\mathcal{H}$ be a set of real-valued functions on a set $\mathcal{X}$. A subset $s_m = \{x_i\}$, $(1 \leq i \leq m)$ 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*

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

*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 $V_\gamma$ dimension is a scale-sensitive variant of Vapnik's dimension.

**Definition 18 ($V_\gamma$ dimension [3, 31])** *Let $\mathcal{H}$ be a set of real-valued functions on a set $\mathcal{X}$. For $\gamma > 0$, a subset $s_m = \{x_i\}$, $(1 \leq i \leq m)$ of $\mathcal{X}$ is said to be* $V_\gamma$-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 $V_\gamma$* dimension *of the set $\mathcal{H}$, $V_\gamma$-dim$(\mathcal{H})$, is the maximal cardinality of a subset of $\mathcal{X}$ $V_\gamma$-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 $V_\gamma$ dimension can be seen as a uniform variant of the fat-shattering dimension.

**Definition 19 (fat-shattering dimension [37, 38])** *Let $\mathcal{H}$ be a set of real-valued functions on a set $\mathcal{X}$. For $\gamma > 0$, a subset $s_m = \{x_i\}$, $(1 \leq i \leq m)$ of $\mathcal{X}$ is said to be* $\gamma$-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 $\gamma$, *or* $P_\gamma$ dimension *of the set $\mathcal{H}$, $P_\gamma$-dim$(\mathcal{H})$, is the maximal cardinality of a subset of $\mathcal{X}$ $\gamma$-shattered by $\mathcal{H}$, if it is finite, or infinity otherwise.*

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

**Definition 20 ($\gamma$-Ψ-shattering)** *Let $\mathcal{H}$ be a set of functions on a set $\mathcal{X}$ taking their values in $\mathbb{R}^Q$. Let $\Psi$ be a family of mappings $\psi$ from $\{1, \ldots, Q\}$ into $\{-1, 1, *\}$. For $\gamma > 0$, a subset $s_m = \{x_i\}$, $(1 \leq i \leq m)$ of $\mathcal{X}$ is said to be* $\gamma$-Ψ-shattered *(Ψ-shattered with margin $\gamma$) by $\mathcal{H}$*

*if there is a mapping* $\psi^m = \left(\psi^{(1)}, \ldots, \psi^{(i)}, \ldots, \psi^{(m)}\right)$ *in* $\Psi^m$ *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*

$$\forall i \in \{1,\ldots,m\}, \left\{ \begin{array}{lll} if\ y_i = & 1 & \exists (k,l) : \left\{ \begin{array}{rcrl} (\psi^{(i)}(k) & = & 1 & \wedge \quad h_{y,k}(x_i) - b_i \geq \gamma) \\ (\psi^{(i)}(l) & = & -1 & \wedge \quad h_{y,l}(x_i) = -h_{y,k}(x_i)) \end{array} \right. \\ if\ y_i = & -1 & \exists (k,l) : \left\{ \begin{array}{rcrl} (\psi^{(i)}(k) & = & 1 & \wedge \quad h_{y,k}(x_i) - b_i \leq -\gamma) \\ (\psi^{(i)}(l) & = & -1 & \wedge \quad h_{y,l}(x_i) = -h_{y,k}(x_i)) \end{array} \right. \end{array} \right.$$

**Definition 21 ($\Psi$-dimension with margin $\gamma$)** *Let* $\mathcal{H}$*,* $\Psi$ *and* $\gamma$ *be defined as above. The* $\Psi$*-dimension of* $\mathcal{H}$ *with margin* $\gamma$*, denoted by* $\Psi\text{-}dim(\mathcal{H},\gamma)$*, is the maximal cardinality of a subset of* $\mathcal{X}$ $\gamma$-$\Psi$*-shattered by* $\mathcal{H}$*, if it is finite, or infinity otherwise.*

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

**Definition 22 (Natarajan dimension with margin $\gamma$)** *Let* $\mathcal{H}$ *be a set of functions on a set* $\mathcal{X}$ *taking their values in* $\mathbb{R}^Q$*. For* $\gamma > 0$*, a subset* $s_m = \{x_i\}$*,* $(1 \leq i \leq m)$ *of* $\mathcal{X}$ *is said to be* $\gamma$-N-shattered *(N-shattered with margin* $\gamma$*) by* $\mathcal{H}$ *if there is a set*

$$I(s_m) = \{(i_1(x_1), i_2(x_1)), \ldots, (i_1(x_i), i_2(x_i)), \ldots, (i_1(x_m), i_2(x_m))\}$$

*of* $m$ *couples of distinct indexes in* $\{1,\ldots,Q\}$ *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*

$$\forall i \in \{1,\ldots,m\}, \left\{ \begin{array}{ll} if\ y_i = 1\ then & \left(h_{y,i_1(x_i)}(x_i) - b_i \geq \gamma \wedge h_{y,i_2(x_i)}(x_i) = -h_{y,i_1(x_i)}(x_i)\right) \\ if\ y_i = -1\ then & \left(h_{y,i_1(x_i)}(x_i) - b_i \leq -\gamma \wedge h_{y,i_2(x_i)}(x_i) = -h_{y,i_1(x_i)}(x_i)\right) \end{array} \right.$$

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

## 4.3 Discussion

Obviously, this definition of the margin $\Psi$-dimensions exhibits all the desirable properties in the case when the vector of biases $v_b$ is equal to the null vector. To highlight this point, one must bear in mind the fact that in the context of this study, one needs to bound them for classes of functions of the form $\Delta_\epsilon^* \mathcal{H}$, with $\epsilon \geq \gamma$. Now, consider the case $y_i = 1$. The fact that there exists an index $k$ such that $\Psi^{(i)}(k) = 1$ and $\Delta_\epsilon^* h_{y,k}(x_i) \geq \gamma$ and an index $l$ satisfying $\Psi^{(i)}(l) = -1$ and $\Delta_\epsilon^* h_{y,l}(x_i) \leq -\gamma$ implies that $M(h_y, x_i) \geq \gamma$. Furthermore, the unique index $k_0$ such that $\Delta^* h_{y,k_0}(x_i) = M(h_y, x_i)$ (or $\Delta_\epsilon^* h_{y,k_0}(x_i) = \min(\epsilon, M(h_y, x_i))$) belongs to the set of indexes $l$ satisfying $\Psi^{(i)}(l) = 1$. The symmetrical properties can be derived in the same way in the case $y_i = -1$.

When applied to the bi-class case, Definition 21 also 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 a 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, $P_\gamma\text{-dim}(\tilde{\mathcal{H}}) = \Psi\text{-dim}(\Delta^*\mathcal{H}, \gamma)$, this result holding irrespective of the choice of the class of mappings $\Psi$ 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-Shela 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 introduced a restriction on the definition of the Natarajan dimension with respect to the one given in [9]. This restriction consists in 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 $\Psi$ considered could be $\binom{Q}{2}$ instead of $Q(Q-1)$. This is useful indeed, since many theorems dealing with Ψ-dimensions involve the cardinality of $\Psi$ (see for instance Theorem 7 in [9]). An equivalent simplification can be performed in the case of the margin Natarajan dimension.

**Proposition 3** *In the definition of the Natarajan dimension with margin $\gamma$, the additional constraint $i_1(x_i) < i_2(x_i)$, $(1 \leq i \leq m)$, can be introduced.*

**Proof** Let $\mathcal{H}_y$ be a subset of $\mathcal{H}$ of cardinality $2^m$ $\gamma$-N-shattering $s_m$ with respect to $I(s_m)$ and $v_b$. Let $I'(s_m)$ be a set of $m$ couples of indexes $(i'_1(x_i), i'_2(x_i))$ deduced from $I(s_m)$ as follows: $\forall i \in \{1,\ldots,m\}$, $(i'_1(x_i), i'_2(x_i)) = (\min(i_1(x_i), i_2(x_i)), \max(i_1(x_i), i_2(x_i)))$. Let $v_{b'} = [b'_i]$ be the vector of $\mathbb{R}^m$ deduced from $v_b$ as follows: $\forall i \in \{1,\ldots,m\}$, $b'_i = b_i$ 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 $\mathcal{H}_y$ still $\gamma$-N-shatters $s_m$ with respect to $I'(s_m)$ 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 $h_{y'}$ "contributes" to the $\gamma$-N-shattering of $s_m$ with respect to $I(s_m)$ 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 22,

$$\forall i \in \{1,\ldots,m\},\ \begin{cases} \text{if } y'_i = 1 \text{ then} & \left(h_{y',i_1(x_i)}(x_i) - b_i \geq \gamma \wedge h_{y',i_2(x_i)}(x_i) = -h_{y',i_1(x_i)}(x_i)\right) \\ \text{if } y'_i = -1 \text{ then} & \left(h_{y',i_1(x_i)}(x_i) - b_i \leq -\gamma \wedge h_{y',i_2(x_i)}(x_i) = -h_{y',i_1(x_i)}(x_i)\right) \end{cases}$$

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

$$\begin{cases} \text{if } y_i = 1 \text{ then} & \left(h_{y',i'_1(x_i)}(x_i) - b'_i \geq \gamma \wedge h_{y',i'_2(x_i)}(x_i) = -h_{y',i'_1(x_i)}(x_i)\right) \\ \text{if } y_i = -1 \text{ then} & \left(h_{y',i'_1(x_i)}(x_i) - b'_i \leq -\gamma \wedge h_{y',i'_2(x_i)}(x_i) = -h_{y',i'_1(x_i)}(x_i)\right) \end{cases} \tag{23}$$

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

$$\begin{cases} \text{if } y_i = -1 \text{ then} & \left(h_{y',i'_2(x_i)}(x_i) + b'_i \geq \gamma \wedge h_{y',i'_1(x_i)}(x_i) = -h_{y',i'_2(x_i)}(x_i)\right) \\ \text{if } y_i = 1 \text{ then} & \left(h_{y',i'_2(x_i)}(x_i) + b'_i \leq -\gamma \wedge h_{y',i'_1(x_i)}(x_i) = -h_{y',i'_2(x_i)}(x_i)\right) \end{cases}$$

The last system can be rewritten as follows:

$$\begin{cases} \text{if } y_i = 1 \text{ then} & \left(-h_{y',i'_1(x_i)}(x_i) + b'_i \leq -\gamma \wedge h_{y',i'_2(x_i)}(x_i) = -h_{y',i'_1(x_i)}(x_i)\right) \\ \text{if } y_i = -1 \text{ then} & \left(-h_{y',i'_1(x_i)}(x_i) + b'_i \geq \gamma \wedge h_{y',i'_2(x_i)}(x_i) = -h_{y',i'_1(x_i)}(x_i)\right) \end{cases}$$

and finally,

$$\begin{cases} \text{if } y_i = 1 \text{ then} & \left(h_{y',i'_1(x_i)}(x_i) - b'_i \geq \gamma \wedge h_{y',i'_2(x_i)}(x_i) = -h_{y',i'_1(x_i)}(x_i)\right) \\ \text{if } y_i = -1 \text{ then} & \left(h_{y',i'_1(x_i)}(x_i) - b'_i \leq -\gamma \wedge h_{y',i'_2(x_i)}(x_i) = -h_{y',i'_1(x_i)}(x_i)\right) \end{cases}$$

This is exactly (23), which thus holds true for all values of $i$ in $\{1, \ldots, m\}$ (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 22, function $h_{y'}$ thus contributes to the $\gamma$-N-shattering of $s_m$ with respect to $I'(s_m)$ 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 $\mathcal{H}_y$ $\gamma$-N-shatters $s_m$ with respect to $I'(s_m)$ and $v_{b'}$, which, by construction of $I'(s_m)$, concludes the proof. ∎

In the sequel, we use the alternative definition of the margin Natarajan dimension resulting from Proposition 3.

# 5 Relating the Covering Number 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 23 (Strong Natarajan dimension)** *Let $n$ be a positive integer and $S$ be equal to $\{-n,\dots,n\}^Q$. Let $\mathcal{F}$ be a set of functions on $\mathcal{X}$ taking their values in $S$. A subset $s_m = \{x_i\}$, $(1 \leq i \leq m)$ of $\mathcal{X}$ is said to be* strongly N-shattered *by $\mathcal{F}$ if there exists a set of couples of indexes*

$$I(s_m) = \{(i_1(x_1), i_2(x_1)), \dots, (i_1(x_i), i_2(x_i)), \dots, (i_1(x_m), i_2(x_m))\}$$

*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 $\{-n+1,\dots,n-1\}^m$ such that, for each binary vector $v_y = [y_i] \in \{-1,1\}^m$, there is a function $f_y = [f_{y,k}]$, $(1 \leq k \leq Q)$, in $\mathcal{F}$ satisfying*

$$\forall i \in \{1,\dots,m\},\ \begin{cases} f_{y,i_1(x_i)}(x_i) - b_i \geq 1 \wedge f_{y,i_2(x_i)}(x_i) = -f_{y,i_1(x_i)}(x_i) & \text{if } y_i = 1 \\ f_{y,i_1(x_i)}(x_i) - b_i \leq -1 \wedge f_{y,i_2(x_i)}(x_i) = -f_{y,i_1(x_i)}(x_i) & \text{if } y_i = -1 \end{cases}$$

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

**Definition 24 (Packing numbers)** *Let $(E,\rho)$ be a pseudo-metric space. A set $H \subset E$ is* $\epsilon$-separated *if, for any distinct points $v_1$ and $v_2$ in $H$, $\rho(v_1,v_2) \geq \epsilon$. The* $\epsilon$-packing number *of $H$, $\mathcal{M}(\epsilon,H,\rho)$, is the maximal size of an $\epsilon$-separated subset of $H$.*

**Definition 25 (Separation)** *Let $\mathcal{F}$ be a set of functions on $\mathcal{X}$ taking their values in $S$, and let $s_m = \{x_i\}$, $(1 \leq i \leq m)$, be a subset of $\mathcal{X}$. Two functions $f^{(1)}$ and $f^{(2)}$ in the class $\mathcal{F}$ are* separated *if they are 2-separated with respect to the pseudo-metric $d_{l_\infty, l_\infty(s_m)}$, i.e. if*

$$\max_{x_i \in s_m} \max_{k \in \{1,\dots,Q\}} \left| f_k^{(1)}(x_i) - f_k^{(2)}(x_i) \right| \geq 2$$

**Definition 26 (Pairwise separated set of functions)** *Let $\mathcal{X}$ be any set and let $S = \{-n,\dots,n\}^Q$. A set $\mathcal{F}$ of functions from $\mathcal{X}$ into $S$ is* pairwise separated *if any two distinct functions of $\mathcal{F}$ are separated.*

**Definition 27 ($\eta$-discretization)** *Let $h = [h_k]$ be a function from $\mathcal{X}$ into $\mathbb{R}^Q$ and $\eta > 0$. The* $\eta$-discretization *of $h$, denoted by $h^{(\eta)} = \left[h_k^{(\eta)}\right]$, is the function from $\mathcal{X}$ into $\mathbb{Z}^Q$ such that*

$$\forall k \in \{1, \ldots, Q\}\,,\ h_k^{(\eta)}(x) = \begin{cases} \lfloor \frac{h_k(x)}{\eta} \rfloor & \text{if } h_k(x) \geq 0 \\ \lceil \frac{h_k(x)}{\eta} \rceil & \text{otherwise} \end{cases}$$

*or equivalently, for all $k$ in $\{1, \ldots, Q\}$, $h_k^{(\eta)}(x) = \max\{j \in \mathbb{Z}_+ :\ j\eta \leq h_k(x)\}$ if $h_k(x) \geq 0$, $h_k^{(\eta)}(x) = \min\{j \in \mathbb{Z}_- :\ j\eta \geq h_k(x)\}$ otherwise. For a set $\mathcal{H}$ of vector-valued functions, let*

$$\mathcal{H}^{(\eta)} = \left\{h^{(\eta)} :\ h \in \mathcal{H}\right\}$$

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 the hypothesis of nonnegativity has been relaxed. Indeed, this hypothesis is here useless. Furthermore, it must be borne in mind that we are ultimately interested in the functions $\Delta_\gamma^* h$ which take their values in $[-\gamma, \gamma]^Q$.

## 5.2 Lemmas

There is a close connection between covering and packing properties of bounded subsets in metric spaces. The following lemma, a proof of which can for instance be found in [39, 5], will prove useful in what follows.

**Lemma 5** *For every pseudo-metric space $(E, \rho)$, every totally bounded subset $H$ of $E$ and $\epsilon > 0$,*

$$\mathcal{M}(2\epsilon, H, \rho) \leq \mathcal{N}(\epsilon, H, \rho) \leq \mathcal{M}(\epsilon, H, \rho)$$

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

**Lemma 6** *For any class $\mathcal{H}$ of functions on $\mathcal{X}$ taking their values in $\mathbb{R}^Q$ and for any $\gamma \in (0, 1]$ and $\eta > 0$:*

1. *for every couple $(\epsilon, \eta)$ satisfying $0 < \eta \leq \gamma$ and $0 < \epsilon \leq \eta/2$,*

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

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

$$\mathcal{M}(\epsilon, \Delta_\gamma^* \mathcal{H}, d_{l_\infty, l_\infty(s_m)}) \leq \mathcal{M}(2, \left(\Delta_\gamma^* \mathcal{H}\right)^{(\eta)}, d_{l_\infty, l_\infty(s_m)}).$$

**Proof** To prove the first proposition, it is enough to establish that any set strongly N-shattered by $\left(\Delta_\gamma^*\mathcal{H}\right)^{(\eta)}$ is also N-shattered with margin $\eta/2$ by $\Delta_\gamma^*\mathcal{H}$. If $s_m$, a subset of $\mathcal{X}$ of cardinality $m$, is strongly N-shattered by $\left(\Delta_\gamma^*\mathcal{H}\right)^{(\eta)}$, then according to Definition 23, there exists a set of couples of indexes $I(s_m)$ and a vector $v_b$ in $\{\lceil -\gamma/\eta\rceil + 1, \ldots, \lfloor\gamma/\eta\rfloor - 1\}^m$ such that for every vector $v_y = [y_i] \in \{-1,1\}^m$, there is a function $\left(\Delta_\gamma^* h_y\right)^{(\eta)} = \left[\left(\Delta_\gamma^* h_{y,k}\right)^{(\eta)}\right]$ in $\left(\Delta_\gamma^*\mathcal{H}\right)^{(\eta)}$, i.e. a function $h_y$ in $\mathcal{H}$ satisfying $\forall i \in \{1,\ldots,m\}$:

$$\begin{cases} \left(\Delta_\gamma^* h_{y,i_1(x_i)}\right)^{(\eta)}(x_i) - b_i \geq 1 \wedge \left(\Delta_\gamma^* h_{y,i_2(x_i)}\right)^{(\eta)}(x_i) = -\left(\Delta_\gamma^* h_{y,i_1(x_i)}\right)^{(\eta)}(x_i) & \text{if } y_i = 1 \\ \left(\Delta_\gamma^* h_{y,i_1(x_i)}\right)^{(\eta)}(x_i) - b_i \leq -1 \wedge \left(\Delta_\gamma^* h_{y,i_2(x_i)}\right)^{(\eta)}(x_i) = -\left(\Delta_\gamma^* h_{y,i_1(x_i)}\right)^{(\eta)}(x_i) & \text{if } y_i = -1 \end{cases}$$

From $\left(\Delta_\gamma^* h_{y,i_2(x_i)}\right)^{(\eta)}(x_i) = -\left(\Delta_\gamma^* h_{y,i_1(x_i)}\right)^{(\eta)}(x_i)$, whether $y_i = 1$ or $y_i = -1$, it springs that $\Delta_\gamma^* h_{y,i_1(x_i)}(x_i)$ and $\Delta_\gamma^* h_{y,i_2(x_i)}(x_i)$ have different signs and thus, given the definition of the $\Delta^*$ operator, $\Delta_\gamma^* h_{y,i_1(x_i)}(x_i) = -\Delta_\gamma^* h_{y,i_2(x_i)}(x_i)$. Thus, it remains to exhibit a scalar $\tilde{b}_i$ such that $\left(\Delta_\gamma^* h_{y,i_1(x_i)}\right)^{(\eta)}(x_i) - b_i \geq 1 \Longrightarrow \Delta_\gamma^* h_{y,i_1(x_i)}(x_i) - \tilde{b}_i \geq \eta/2$ and $\left(\Delta_\gamma^* h_{y,i_1(x_i)}\right)^{(\eta)}(x_i) - b_i \leq 1 \Longrightarrow \Delta_\gamma^* h_{y,i_1(x_i)}(x_i) - \tilde{b}_i \leq -\eta/2$. 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 $y_i = -1$

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

or equivalently

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

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

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

or equivalently

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

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

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

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

or equivalently

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

To sum up, a satisfactory solution consists in setting $\tilde{b}_i = \eta(b_i + 1/2)$ if $b_i \geq 0$ and $\tilde{b}_i = \eta(b_i - 1/2)$ otherwise. By definition, the set of functions $\Delta_\gamma^* h_y$, for $v_y$ in $\{-1,1\}^m$, N-shatters $s_m$ with margin $\eta/2$, for a set of couples of indexes and a vector of biases respectively equal to $I(s_m)$ and $\tilde{v}_b = \left[\tilde{b}_i\right]$. As a consequence, any set strongly N-shattered by $\left(\Delta_\gamma^* \mathcal{H}\right)^{(\eta)}$ 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:

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

$$\left|\Delta_\gamma^* h_k^{(1)}(x) - \Delta_\gamma^* h_k^{(2)}(x)\right| \geq 3\eta \Longrightarrow \left|\left(\Delta_\gamma^* h_k^{(1)}\right)^{(\eta)}(x) - \left(\Delta_\gamma^* h_k^{(2)}\right)^{(\eta)}(x)\right| \geq 2$$

Indeed, without loss of generality, we can make the hypothesis that $\Delta_\gamma^* h_k^{(1)}(x) > \Delta_\gamma^* h_k^{(2)}(x)$. Then,

$$\left(\left(\Delta_\gamma^* h_k^{(2)}\right)^{(\eta)}(x) - 1\right)\eta < \Delta_\gamma^* h_k^{(2)}(x) < \Delta_\gamma^* h_k^{(1)}(x) < \left(\left(\Delta_\gamma^* h_k^{(1)}\right)^{(\eta)}(x) + 1\right)\eta$$

Thus

$$\left(\left(\Delta_\gamma^* h_k^{(1)}\right)^{(\eta)}(x) + 1\right)\eta - \left(\left(\Delta_\gamma^* h_k^{(2)}\right)^{(\eta)}(x) - 1\right)\eta > 3\eta$$

and finally

$$\left(\Delta_\gamma^* h_k^{(1)}\right)^{(\eta)}(x) - \left(\Delta_\gamma^* h_k^{(2)}\right)^{(\eta)}(x) > 1$$

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

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

$$\forall\ \left(\Delta_\gamma^* h^{(1)}, \Delta_\gamma^* h^{(2)}\right) \in s^2_{\Delta_\gamma^* \mathcal{H}},\ d_{l_\infty, l_\infty(s_m)}\left(\Delta_\gamma^* h^{(1)}, \Delta_\gamma^* h^{(2)}\right) \geq 3\eta \Longrightarrow$$

$$\max_{x\in s_m}\max_k \left|\Delta_\gamma^* h_k^{(1)}(x) - \Delta_\gamma^* h_k^{(2)}(x)\right| \geq 3\eta \Longrightarrow$$
$$\max_{x\in s_m}\max_k \left|\left(\Delta_\gamma^* h_k^{(1)}\right)^{(\eta)}(x) - \left(\Delta_\gamma^* h_k^{(2)}\right)^{(\eta)}(x)\right| \geq 2 \Longrightarrow$$
$$d_{l_\infty,l_\infty(s_m)}\left(\left(\Delta_\gamma^* h_k^{(1)}\right)^{(\eta)}, \left(\Delta_\gamma^* h_k^{(2)}\right)^{(\eta)}\right) \geq 2$$

We have thus proved the second proposition. ∎

Note that a more interesting second proposition could have resulted from using a different definition of the $\eta$-discretization. Indeed, setting $h_k^{(\eta)}(x) = \lfloor \frac{h_k(x)}{\eta} \rfloor$ irrespective of the sign of $h_k(x)$, one can easily establish that the following proposition (with a dependence between $\epsilon$ and $\eta$ identical to the one of [3]) holds true: for every $\epsilon \geq 2\eta$ and every $s_m \in \mathcal{X}^m$, $\mathcal{M}(\epsilon, \Delta_\gamma^*\mathcal{H}, d_{l_\infty,l_\infty(s_m)}) \leq \mathcal{M}(2, \left(\Delta_\gamma^*\mathcal{H}\right)^{(\eta)}, d_{l_\infty,l_\infty(s_m)})$. The reason for our choice is to get an additional property, namely:

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

Indeed, this property will prove very useful in what follows, actually no later than in the next subsection.

## 5.3 Classes of $\delta$ functions

Our generalized Sauer-Shela lemma will make use of specific properties of the discretizations of the $\Delta_\gamma^* h$ functions, which leads us to introduce the notion of $\delta$ functions.

**Definition 28 ($\delta$ functions)** *Let $\mathcal{F}$ be a class of vector-valued functions $f = [f_k]$, $(1 \leq k \leq Q)$, from a domain $\mathcal{X}$ into a finite range $S = \{-n,\ldots,n\}^Q$. $\mathcal{F}$ is a class of* $\delta$ functions *if: $\forall f \in \mathcal{F}$, $\forall x \in \mathcal{X}$, $\exists\,(k(f,x), i(f,x))$ in $\{1,\ldots,Q\} \times \{0,\ldots,n\}$ such that:*

$$\forall l \in \{1,\ldots,Q\},\ \begin{cases} f_l(x) = i(f,x) & \text{if } l = k(f,x) \\ f_l(x) = -i(f,x) & \text{otherwise} \end{cases} \tag{24}$$

**Lemma 7** *Let $\mathcal{H}$ be a set of functions on a set $\mathcal{X}$ taking their values in $\mathbb{R}^Q$ and $\gamma$ and $\eta$ two positive real values. Then $\left(\Delta_\gamma^*\mathcal{H}\right)^{(\eta)}$ is a set of $\delta$ functions.*

**Proof** This lemma directly springs from Definitions 3, 9, 27 and 28. ∎

**Lemma 8** *Let $\mathcal{F}$ be a class of $\delta$ functions from $\mathcal{X}$ into $\{-n,\ldots,n\}^Q$, and let $s_m = \{x_i\}$, $(1 \leq i \leq m)$. If two functions $f^{(1)}$ and $f^{(2)}$ in $\mathcal{F}$ are separated on $s_m$, then there exists an element $x_i$ of $s_m$ such that $\{f^{(1)}, f^{(2)}\}$ strongly N-shatters the singleton $\{x_i\}$.*

**Proof** Let $k_0$ be an index in $\{1, \ldots, Q\}$ such that $\left| f_{k_0}^{(1)}(x_i) - f_{k_0}^{(2)}(x_i) \right| \geq 2$. Without loss of generality, we can make the assumption that $f_{k_0}^{(1)}(x_i) - f_{k_0}^{(2)}(x_i) \geq 2$. To prove the assertion, four cases must be considered.

1) $k_0 = k(f^{(1)}, x_i) \wedge k_0 = k(f^{(2)}, x_i)$
Let $b_i = f_{k_0}^{(1)}(x_i) - 1$. Then $f_{k_0}^{(1)}(x_i) - b_i \geq 1$ and $f_{k_0}^{(2)}(x_i) - b_i \leq -1$. Furthermore, for any index $l_0$ different from $k_0$, $f_{l_0}^{(1)}(x_i) = -f_{k_0}^{(1)}(x_i)$ and $f_{l_0}^{(2)}(x_i) = -f_{k_0}^{(2)}(x_i)$.
2) $k_0 = k(f^{(1)}, x_i) \wedge k_0 \neq k(f^{(2)}, x_i)$
Let $b_i = f_{k_0}^{(1)}(x_i) - 1$. Then $f_{k_0}^{(1)}(x_i) - b_i \geq 1$ and $f_{k_0}^{(2)}(x_i) - b_i \leq -1$. Furthermore, let $l_0 = k(f^{(2)}, x_i)$. $f_{l_0}^{(1)}(x_i) = -f_{k_0}^{(1)}(x_i)$ and $f_{l_0}^{(2)}(x_i) = -f_{k_0}^{(2)}(x_i)$.
3) $k_0 \neq k(f^{(1)}, x_i) \wedge k_0 = k(f^{(2)}, x_i)$
This case leads to a contradiction. Indeed, $k_0 \neq k(f^{(1)}, x_i) \Longrightarrow f_{k_0}^{(1)}(x_i) \leq 0$, whereas $k_0 = k(f^{(2)}, x_i) \Longrightarrow f_{k_0}^{(2)}(x_i) \geq 0$. Thus, $f_{k_0}^{(2)}(x_i) \geq f_{k_0}^{(1)}(x_i)$, which is in contradiction with the hypothesis $f_{k_0}^{(1)}(x_i) - f_{k_0}^{(2)}(x_i) \geq 2$.
4) $k_0 \neq k(f^{(1)}, x_i) \wedge k_0 \neq k(f^{(2)}, x_i)$
Let $l_0 = k(f^{(1)}, x_i)$. If $l_0 = k(f^{(2)}, x_i)$, let $b_i = f_{k_0}^{(1)}(x_i) - 1$. Then $f_{k_0}^{(1)}(x_i) - b_i \geq 1$ and $f_{k_0}^{(2)}(x_i) - b_i \leq -1$. Furthermore, $f_{l_0}^{(1)}(x_i) = -f_{k_0}^{(1)}(x_i)$ and $f_{l_0}^{(2)}(x_i) = -f_{k_0}^{(2)}(x_i)$. If $l_0 \neq k(f^{(2)}, x_i)$, let $m_0 = k(f^{(2)}, x_i)$ and $b_i = f_{m_0}^{(2)}(x_i) - 1$. Then $f_{m_0}^{(2)}(x_i) - b_i \geq 1$. Furthermore, $f_{k_0}^{(1)}(x_i) - f_{k_0}^{(2)}(x_i) \geq 2 \Longrightarrow f_{l_0}^{(1)}(x_i) - f_{m_0}^{(2)}(x_i) \leq -2$. Since $f_{m_0}^{(1)}(x_i) \leq f_{l_0}^{(1)}(x_i)$, $f_{m_0}^{(1)}(x_i) - f_{m_0}^{(2)}(x_i) \leq -2$ and thus $f_{m_0}^{(1)}(x_i) - b_i \leq -1$. At last, $f_{l_0}^{(1)}(x_i) = -f_{m_0}^{(1)}(x_i)$ and $f_{l_0}^{(2)}(x_i) = -f_{m_0}^{(2)}(x_i)$. ∎

Lemma 8 will appear of central importance in the sequel of this section. Now, if $\left(\Delta_\gamma^* \mathcal{H}\right)^{(\eta)}$ is a set of $\delta$ functions, the same is not true of $(\Delta_\gamma \mathcal{H})^{(\eta)}$. What is more important, one can exhibit two functions $h^{(1)}$ and $h^{(2)}$ and a point $x$ such that $\left(\Delta_\gamma h^{(1)}\right)^{(\eta)}$ and $\left(\Delta_\gamma h^{(2)}\right)^{(\eta)}$ are separated on $s_1 = \{x\}$ whereas they do not shatter $s_1$. A simple example is $h^{(1)}(x) = [0.5, -0.5, -0.9]^T$, $h^{(2)}(x) = [0.5, -0.5, -0.5]^T$, with $\gamma = 1$ and $\eta = 0.1$. As a consequence, Lemma 8 does not hold anymore for the class of functions $(\Delta_\gamma \mathcal{H})^{(\eta)}$ when no additional hypothesis is made regarding the class of functions $\mathcal{H}$ (the problem is considered in its full generality). This is the main reason why the results of this section are derived for the $\Delta^*$ margin operator and not the $\Delta$ operator.

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

## 5.4 Generalized Sauer-Shela lemma

**Lemma 9** *Let $\mathcal{F}$ be a class of $\delta$ functions $f = [f_k]$, $(1 \leq k \leq Q)$, from a finite domain $\mathcal{X}$ of cardinality $|\mathcal{X}|$ to a finite range $S = \{-n, \ldots, n\}^Q$. Let SN-dim$(\mathcal{F}) = d$. Then*

$$\mathcal{M}(2, \mathcal{F}, d_{l_\infty, l_\infty(\mathcal{X})}) < 2\left(|\mathcal{X}|(Qn+1)Qn\right)^{\lceil \log_2(\phi(d,|\mathcal{X}|)) \rceil} \tag{25}$$

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

**Proof** Let us say that a class $\mathcal{F}$ as above strongly N-shatters a triplet $(s_m, I(s_m), v_b)$ (for a nonempty subset $s_m$ of $\mathcal{X}$ of cardinality $m$, a set of couples of indexes $I(s_m)$ and a vector of biases $v_b$ if $\mathcal{F}$ strongly N-shatters $s_m$ according to $I(s_m)$ and $v_b$. For all integers $l \geq 2$ and $|\mathcal{X}| \geq 1$, let $t(l, |\mathcal{X}|)$ denote the maximum number $t$ such that, for every set $\mathcal{F}_l$ of $l$ pairwise separated functions from $\mathcal{F}$, $\mathcal{F}_l$ strongly N-shatters at least $t$ triplets $(s, I(s), v_b)$. If no such $\mathcal{F}_l$ exists, then $t(l, |\mathcal{X}|)$ is infinite.

The number of triplets $(s, I(s), v_b)$ that could be shattered and for which the cardinality of $s$ does not exceed $d \geq 1$ is less than $\sum_{i=1}^{d} \binom{|\mathcal{X}|}{i} \left(\binom{Q}{2}(2n-1)\right)^i$, since for $s$ of size $i > 0$, there are strictly less than $\left(\binom{Q}{2}(2n-1)\right)^i$ possibilities to choose the couple $(I(s), v_b)$. It follows that, given a set of functions $\mathcal{F}$ from $\mathcal{X}$ into $S$, $t(l, |\mathcal{X}|) \geq \phi(d, |\mathcal{X}|)$ for some $l$ and SN-dim$(\mathcal{F}) \leq d$ implies $t(l, |\mathcal{X}|) = \infty$. As a consequence, by definition of $t(l, |\mathcal{X}|)$, there is no set $\mathcal{F}_l$ of $l$ pairwise separated functions in $\mathcal{F}$ (otherwise $t(l, |\mathcal{X}|)$ would be finite) and finally, by definition of $\mathcal{M}(2, \mathcal{F}, d_{l_\infty, l_\infty(\mathcal{X})})$, $\mathcal{M}(2, \mathcal{F}, d_{l_\infty, l_\infty(\mathcal{X})}) < l$. Therefore, to finish the proof, it suffices to show that, for all $d \geq 1$ and $|\mathcal{X}| \geq 1$,

$$t\left(2\left(|\mathcal{X}|(Qn+1)Qn\right)^{\lceil \log_2(\phi(d,|\mathcal{X}|)) \rceil}, |\mathcal{X}|\right) \geq \phi(d, |\mathcal{X}|) \tag{26}$$

We claim that

$$t(2, |\mathcal{X}|) \geq 1 \tag{27}$$

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

$$t\left(2m|\mathcal{X}|(Qn+1)Qn, |\mathcal{X}|\right) \geq 2t(2m, |\mathcal{X}| - 1) \tag{28}$$

for all $m \geq 1$ and $|\mathcal{X}| \geq 2$.

The first part of the claim is a direct consequence of Lemma 8.

For the second part, first note that if no set of $2m|\mathcal{X}|(Qn+1)Qn$ pairwise separated functions from $\mathcal{X}$ to $S$ exists, then by definition $t\left(2m|\mathcal{X}|(Qn+1)Qn, |\mathcal{X}|\right) = \infty$ and hence the claim holds. Assume then that there is a set $\mathcal{F}_0$ of $2m|\mathcal{X}|(Qn+1)Qn$ pairwise separated functions from $\mathcal{X}$ to $S$. Split it arbitrarily into $m|\mathcal{X}|(Qn+1)Qn$ pairs. For each pair $\left(f^{(1)}, f^{(2)}\right)$,

there exists a point $x \in \mathcal{X}$ strongly N-shattered by $\left(f^{(1)}, f^{(2)}\right)$. Once more, this is a direct consequence of Lemma 8. The number of different values that a vector $f(x)$ can take is equal to $Qn+1$. Thus, by the pigeonhole principle, switching the indexes in the couples of functions if needed, for each procedure of this type, at least $(m|\mathcal{X}|(Qn+1)Qn) \,/\, \left(|\mathcal{X}|\binom{Qn+1}{2}\right) = 2m$ of the resulting couples of functions are such that they all shatter the same point $x_0$ and the triplet $\left(x_0, f^{(1)}(x_0), f^{(2)}(x_0)\right)$ is the same. 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 $x_0 \in \mathcal{X}$, $(k_0, l_0) \in \{1, \ldots, Q\}^2$ with $k_0 < l_0$, two scalars $K_{0,+}$ and $K_{0,-}$ in $\{-n, \ldots, n\}$ and a scalar $b_0$ in $\{-n+1, \ldots, n-1\}$ such that:

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

Since the members of $\mathcal{F}_+$ are pairwise separated on $\mathcal{X}$ but are all equal on $x_0$, they are pairwise separated on $\mathcal{X} \setminus \{x_0\}$. The same holds for the members of $\mathcal{F}_-$. Hence, by definition of the function $t$, $\mathcal{F}_+$ strongly N-shatters at least $t\,(2m, |\mathcal{X}| - 1)$ triplets $(s, I(s), v_b)$ with $s \subseteq \mathcal{X} \setminus \{x_0\}$, and the same holds for $\mathcal{F}_-$. Clearly, $\mathcal{F}_0$ strongly N-shatters all triplets strongly N-shattered either by $\mathcal{F}_+$ or by $\mathcal{F}_-$. Moreover, if the same triplet $(s, I(s), v_b)$ is strongly N-shattered both by $\mathcal{F}_+$ and by $\mathcal{F}_-$, then $\mathcal{F}_0$ also strongly N-shatters the triplet $(\{x_0\} \bigcup s, \{(k_0, l_0)\} \bigcup I(s), \bar{v}_b,)$, 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 $(s, I(s), v_b)$ be a triplet strongly N-shattered both by $\mathcal{F}_+$ and by $\mathcal{F}_-$. Then, for any vector $v_y = [y_i]$ in $\{-1, 1\}^{|s|}$, there exists (at least) one function $f_{+,y}$ in $\mathcal{F}_+$ such that

$$\forall i \in \{1, \ldots, |s|\},\ \begin{cases} f_{+,y,i_1(x_i)}(x_i) - b_i \geq 1 \wedge f_{+,y,i_2(x_i)}(x_i) = -f_{+,y,i_1(x_i)}(x_i) & \text{if } y_i = 1 \\ f_{+,y,i_1(x_i)}(x_i) - b_i \leq -1 \wedge f_{+,y,i_2(x_i)}(x_i) = -f_{+,y,i_1(x_i)}(x_i) & \text{if } y_i = -1 \end{cases}$$

and

$$f_{+,y,k_0}(x_0) - b_0 \geq 1 \wedge f_{+,y,l_0}(x_0) = -f_{+,y,k_0}(x_0)$$

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

$$\forall i \in \{1, \ldots, |s|\},\ \begin{cases} f_{-,y,i_1(x_i)}(x_i) - b_i \geq 1 \wedge f_{-,y,i_2(x_i)}(x_i) = -f_{-,y,i_1(x_i)}(x_i) & \text{if } y_i = 1 \\ f_{-,y,i_1(x_i)}(x_i) - b_i \leq -1 \wedge f_{-,y,i_2(x_i)}(x_i) = -f_{-,y,i_1(x_i)}(x_i) & \text{if } y_i = -1 \end{cases}$$

and

$$f_{-,y,k_0}(x_0) - b_0 \leq -1 \wedge f_{-,y,l_0}(x_0) = -f_{-,y,k_0}(x_0)$$

Since, once more by construction, neither $\mathcal{F}_+$ nor $\mathcal{F}_-$ strongly N-shatters $\{x_0\} \bigcup s$ (whatever the couple $(I(\{x_0\} \bigcup s), \bar{v}_b)$ may be), it follows that $t\,(2m|\mathcal{X}|(Qn+1)Qn, |\mathcal{X}|) \geq 2t(2m, |\mathcal{X}| -$

1) which is precisely (28).

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

$$l = 2\left((Qn+1)Qn\right)^r \ \Pi_{u=0}^{r-1}(|\mathcal{X}| - u)$$

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

$$t\left(2\left(|\mathcal{X}|(Qn+1)Qn\right)^r, |\mathcal{X}|\right) \geq 2^r$$

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

$$t\left(2\left(|\mathcal{X}|(Qn+1)Qn\right)^{\lceil \log_2(\phi(d,|\mathcal{X}|)) \rceil}, |\mathcal{X}|\right) \geq 2^{\lceil \log_2(\phi(d,|\mathcal{X}|)) \rceil}$$

and consequently

$$t\left(2\left(|\mathcal{X}|(Qn+1)Qn\right)^{\lceil \log_2(\phi(d,|\mathcal{X}|)) \rceil}, |\mathcal{X}|\right) \geq 2^{\log_2(\phi(d,|\mathcal{X}|))} = \phi(d, |\mathcal{X}|)$$

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

$$2\left(|\mathcal{X}|(Qn+1)Qn\right)^{\lceil \log_2(\phi(d,|\mathcal{X}|)) \rceil} > (Qn+1)^{|\mathcal{X}|}$$

Since the total number of $\delta$ functions from $\mathcal{X}$ into $S$ is precisely $(Qn+1)^{|\mathcal{X}|}$, there is no set of pairwise separated functions of cardinality larger than this included in $\mathcal{F}$ and hence, by definition of $t$,

$$t\left(2\left(|\mathcal{X}|(Qn+1)Qn\right)^{\lceil \log_2(\phi(d,|\mathcal{X}|)) \rceil}, |\mathcal{X}|\right) = \infty$$

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

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

In order to apply the lemmas derived in this section to compute an upper bound on the covering number appearing in the confidence interval of our uniform convergence result, $\mathcal{N}_{\infty,\infty}(\gamma/4, \Delta_\gamma^* \mathcal{H}, 2m)$, we must first remember that to prove Theorem 1, we have specifically considered $\epsilon$-covers of $\Delta_\gamma^* \mathcal{H}$ included in $\Delta_\gamma^* \mathcal{H}$. However, Lemma 5 is not based on this hypothesis. Fortunately, this rises no difficulty, since making use of the triangle inequality,

it is easy to check that from any $\epsilon/2$-covers of $\Delta_\gamma^* \mathcal{H}$ not included in $\Delta_\gamma^* \mathcal{H}$, it is possible to build an $\epsilon$-cover of $\Delta_\gamma^* \mathcal{H}$ of equal cardinality included in $\Delta_\gamma^* \mathcal{H}$. For all $0 < \epsilon, \eta \leq \gamma \leq 1$, let

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

and

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

Bearing in mind the specificity of the covering number considered, applying Lemma 5 to $\Delta_\gamma^* \mathcal{H}$ gives:

$$\mathcal{N}_{\infty,\infty}(\gamma/4, \Delta_\gamma^* \mathcal{H}, 2m) \leq \mathcal{M}_{\infty,\infty}(\gamma/8, \Delta_\gamma^* \mathcal{H}, 2m)$$

Setting $\epsilon = \gamma/8$ ($\eta = \gamma/24$) in Proposition 2 of Lemma 6, one establishes that:

$$\mathcal{M}_{\infty,\infty}(\gamma/8, \Delta_\gamma^* \mathcal{H}, 2m) \leq \mathcal{M}_{\infty,\infty}(2, \left(\Delta_\gamma^* \mathcal{H}\right)^{(\gamma/24)}, 2m)$$

Similarly, the packing numbers of the discretized set of functions can be bounded thanks to Lemma 9, by setting $\mathcal{F} = \left(\Delta_\gamma^* \mathcal{H}\right)^{(\gamma/24)}$ and $|\mathcal{X}| = 2m$. To make use of this lemma, the nature of the range $S$, and more precisely the value of the parameter $n$, must first be established. By definition, each component $\Delta_\gamma^* h_k$ of a function $\Delta_\gamma^* h$ in $\Delta_\gamma^* \mathcal{H}$ takes its values in $[-\gamma, \gamma]$. As a consequence, its $\gamma/24$-discretization takes its values in $\{-24, \ldots, 24\}$, i.e. in a set of cardinality 49. Thus, $n = 24$, $2n + 1 = 49$ and we get:

$$\mathcal{M}_{\infty,\infty}(2, \left(\Delta_\gamma^* \mathcal{H}\right)^{(\gamma/24)}, 2m) < 2\left(48m(24Q+1)Q\right)^{\lceil \log_2(\phi(d,2m)) \rceil} \tag{29}$$

In the right-hand side of (29), $\phi(d, 2m) = \sum_{i=1}^{d} \binom{2m}{i} \left(47\binom{Q}{2}\right)^i$, where $d$ is the strong Natarajan dimension of $\left(\Delta_\gamma^* \mathcal{H}\right)^{(\gamma/24)}$. Since we are interested in upperbounding the capacity measure, one can also make use of Proposition 1 in Lemma 6 to replace $d$ with the Natarajan dimension with margin $\gamma/48$ of $\Delta_\gamma^* \mathcal{H}$, N-dim $\left(\Delta_\gamma^* \mathcal{H}, \gamma/48\right)$. 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 $\gamma$ in $(0, 1]$ and every integer value of $m$ satisfying $2m \geq$ N-dim $\left(\Delta_\gamma^* \mathcal{H}, \gamma/48\right)$, the following bound is true:*

$$\mathcal{N}_{\infty,\infty}(\gamma/4, \Delta_\gamma^* \mathcal{H}, 2m) < 2\left(48m(24Q+1)Q\right)^{\lceil \log_2(\phi(d,2m)) \rceil} \tag{30}$$

*where $d =$ N-dim $\left(\Delta_\gamma^* \mathcal{H}, \gamma/48\right)$ and $\phi(d, 2m) = \sum_{i=1}^{d} \binom{2m}{i} \left(47\binom{Q}{2}\right)^i$.*

## 5.6 Standard bound on function $\phi$

To find an upper bound of $\phi(d, |\mathcal{X}|)$, we take our inspiration from the version of "Sauer's lemma" due to Vapnik and Chervonenkis (see for instance [69]). For all triplets $(d, |\mathcal{X}|, K)$ of integers satisfying $1 \leq d \leq |\mathcal{X}|$, $K \geq 1$ let

$$\Phi(d, |\mathcal{X}|, K) = \sum_{i=0}^{d} \binom{|\mathcal{X}|}{i} K^i$$

i.e. $\Phi\left(d, |\mathcal{X}|, 47\binom{Q}{2}\right) = \phi(d, |\mathcal{X}|) + 1$ for $1 \leq d \leq |\mathcal{X}|$. Function $\Phi$ satisfies the following recurrence formula:

$$\forall d \geq 1,\ \forall |\mathcal{X}| \geq d,\ \forall K \geq 1,\ \Phi(d+1, |\mathcal{X}|+1, K) = \Phi(d+1, |\mathcal{X}|, K) + K\Phi(d, |\mathcal{X}|, K) \quad (31)$$

We will now prove the following lemma, which extends Lemma 4.5 in [68] (see also the appendix of Chapter 6 in [66]).

**Lemma 10** *For all triplet $(d, |\mathcal{X}|, K)$ of positive integers such that $d \leq |\mathcal{X}|$, and $K \geq 1$, the following bound is true:*

$$\Phi(d, |\mathcal{X}|, K) < \left(\frac{Ke|\mathcal{X}|}{d}\right)^d \quad (32)$$

**Proof** First, note that the proof in the case $d = |\mathcal{X}|$ is trivial. Furthermore, $1.5\,(K|\mathcal{X}|)^d / d!$ is a lower bound of $(Ke|\mathcal{X}|/d)^d$. Proving this last bound is equivalent to proving that

$$\frac{1.5}{d!} < \left(\frac{e}{d}\right)^d$$

for $d \geq 1$. This can be done by recurrence. Let $u_d = 1.5/d!$ and $v_d = (e/d)^d$. The property is obviously true for $d = 1$ since $u_1 = 1.5 < v_1 = e$. Furthermore, for all $d \geq 1$,

$$\frac{v_{d+1}u_d}{v_d u_{d+1}} = e\left(\frac{d}{d+1}\right)^d$$

The sequence of general term $(d/(d+1))^d$ is well known to be decreasing and have a limit equal to $1/e$. As a consequence, $(v_{d+1}u_d)/(v_d u_{d+1})$ is always greater than 1 for $d \geq 1$. Consequently, $v_d/u_d$ increases when $d$ increases, and thus $u_d < v_d \Longrightarrow u_{d+1} < v_{d+1}$. Given the recurrence formula (31), proving (32) can also be performed by recurrence. This amounts to proving three separate results: that (32) stands for $d = 1$, irrespective of the value of $|\mathcal{X}| \geq 1$, that it stands for $|\mathcal{X}| = d+1$, irrespective of the value of $d \geq 1$, and that if it stands for both couples $(d, |\mathcal{X}|)$ and $(d+1, |\mathcal{X}|)$, then it also stands for the couple $(d+1, |\mathcal{X}|+1)$. The case $d = 1$ is trivial. Indeed,

$$\Phi(1, |\mathcal{X}|, K) = 1 + K|\mathcal{X}| < 1.5K|\mathcal{X}|$$

As for the "general case", making use of (31) gives:

$$\left(\Phi(d,|\mathcal{X}|,K) < 1.5\frac{(K|\mathcal{X}|)^d}{d!}\right) \wedge \left(\Phi(d+1,|\mathcal{X}|,K) < 1.5\frac{(K|\mathcal{X}|)^{d+1}}{(d+1)!}\right) \Longrightarrow$$

$$K\Phi(d,|\mathcal{X}|,K) + \Phi(d+1,|\mathcal{X}|,K) < K\,(d+1+|\mathcal{X}|)\,1.5\frac{(K|\mathcal{X}|)^d}{(d+1)!} \Longrightarrow$$

$$\Phi(d+1,|\mathcal{X}|+1,K) < K\,(d+1+|\mathcal{X}|)\,1.5\frac{(K|\mathcal{X}|)^d}{(d+1)!} \tag{33}$$

Newton's binomial expansion gives:

$$(|\mathcal{X}|+1)^{d+1} = \sum_{k=0}^{d+1}\binom{(d+1)}{k}|\mathcal{X}|^k$$

and consequently, restricting the expansion to the terms corresponding to $k = d$ and $k = d+1$,

$$(|\mathcal{X}|+1)^{d+1} > (d+1+|\mathcal{X}|)|\mathcal{X}|^d$$

By substitution into (33), this leads to

$$\Phi(d+1,|\mathcal{X}|+1,K) < 1.5K^{d+1}\frac{(|\mathcal{X}|+1)^{d+1}}{(d+1)!}$$

which is precisely our claim. For the last part of the proof, corresponding to the case $|\mathcal{X}| = d+1$, we have:

$$\Phi(d,d+1,K) = (K+1)^{d+1} - K^{d+1} < (d+1)(K+1)^d$$

$\forall d \in \mathbb{N}^*$, $2 \leq ((d+1)/d)^d \leq e$. As a consequence,

$$2(Ke)^d \leq \left(\frac{Ke(d+1)}{d}\right)^d$$

It is obvious that the relation

$$(d+1)(K+1)^d < 2(Ke)^d$$

stands for all couple $(d,K)$ of positive integers. By transitivity, we thus get

$$\Phi(d,d+1,K) < \left(\frac{Ke(d+1)}{d}\right)^d$$

which is exactly (32) in the case $|\mathcal{X}| = d+1$ and consequently concludes the proof. ∎

Note that this proof could be derived without making use of Stirling's approximation, and is consequently simpler than the one from which it is inspired.

## 5.7 Main theorem and discussion

To derive the main theorem relating the covering number of interest to the margin Natarajan dimension of $\Delta_\gamma^* \mathcal{H}$, it suffices to make use of Lemma 10 with $|\mathcal{X}| = 2m$ and $K = 47\binom{Q}{2}$. This implies that

$$\phi(d, 2m) < \Phi\left(d, 2m, 47\binom{Q}{2}\right) < (47emQ(Q-1)/d)^d$$

and consequently

$$\log_2\left(\phi(d, 2m)\right) < d \log_2\left(47emQ(Q-1)/d\right)$$

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 $\gamma$ in $(0, 1]$ and every integer value of $m$ satisfying $2m \geq$ N-dim$\left(\Delta_\gamma^* \mathcal{H}, \gamma/48\right)$, the following bound is true:*

$$\mathcal{N}_{\infty,\infty}(\gamma/4, \Delta_\gamma^* \mathcal{H}, 2m) < 2\left(48m(24Q+1)Q\right)^{\lceil d \log_2(47emQ(Q-1)/d)\rceil} \tag{34}$$

*where $d =$ N-dim$\left(\Delta_\gamma^* \mathcal{H}, \gamma/48\right)$.*

To sum up, in this section, we have derived a bound on the covering number of interest in terms of a scale-sensitive extension of one of the Ψ-dimensions, the Natarajan dimension. Obviously, such a generalized Sauer-Shela 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 [29], the computations leading to Theorem 4 have highlighted the difficulty to upperbound the margin graph dimension in the case of a multivariate affine model (architecture of the multi-class SVMs). In the following section, it will appear clearly that switching to the margin Natarajan dimension makes it possible to follow the standard pathway used to bound 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 [10, 15] as a nonlinear extension of the maximal margin hyperplane [66]. 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 $\Phi$. 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 [55, 47, 68, 49, 2]. The most recent development of this approach is presented in [4], where the authors introduce a new machine which represents a mixture of the SVM for pattern recognition and the SVM for real-valued function estimation. The multi-class SVMs are globally more recent. They are all obtained by combining a multivariate affine model with the nonlinear mapping $\Phi$ into the feature space [68, 72, 13, 28, 16, 17, 43, 25, 35, 42]. Formally, the family $\mathcal{H}$ of functions $h = [h_k]$ computed by these machines is defined by

$$\forall k \in \{1, \ldots, Q\}\,,\ h_k(x) = \langle w_k, \Phi(x) \rangle + b_k \tag{35}$$

As usual, the mapping $\Phi$ does not appear explicitely in the computations. Thanks to the "kernel trick", it is replaced with the *reproducing kernel function* $\kappa$, which computes the $l_2$ dot product in the feature space, i.e.:

$$\forall (x^{(1)}, x^{(2)}) \in \mathcal{X}^2,\ \kappa(x^{(1)}, x^{(2)}) = \langle \Phi(x^{(1)}), \Phi(x^{(2)}) \rangle \tag{36}$$

Hence, the "linear part" of each component of the model is a function of $x$ belonging to a Reproducing Kernel Hilbert Space (RKHS) (see for instance [53, 70, 71]). The kernel satisfies Mercer's conditions [1]. Hereafter, the feature space is denoted by $E_{\Phi(\mathcal{X})}$.

## 6.2 Upperbounding the margin Natarajan dimension of $\Delta_\gamma \mathcal{H}$

Since all the M-SVMs proposed so far only differ in their learning algorithm and not their architecture, their margin Natarajan dimension, or more precisely the margin Natarajan dimension of $\Delta_\gamma \mathcal{H}$, where $\mathcal{H}$ is the class of functions computed by a multivariate affine model, is the same. Hereafter, the strategy implemented to bound it from above is of the "divide and conquer" type. It consists in deriving progressively a bound on N-dim $(\Delta_\gamma \mathcal{H}, \epsilon)$ in terms of the dimensions of simpler classes of functions. A first simplification results from the following trivial bound:

$$\forall (\gamma, \epsilon) : 0 < \epsilon \leq \gamma \leq 1,\ \ \text{N-dim}\,(\Delta_\gamma \mathcal{H}, \epsilon) \leq \text{N-dim}\,(\Delta \mathcal{H}, \epsilon) \tag{37}$$

Furthermore, in the computation of an upper bound on the covering number of interest, it is possible to deal with the biases $b_k$ in (35) thanks to a simple extension of a result appearing in the conclusion of [73] (see for instance Theorem 5 in [25]). As a consequence, for the sake of simplicity, in what follows, we only consider M-SVMs based on a multivariate linear (and not affine) architecture. Anyway, even without making use of the aforementioned results, this simplification is not very restrictive, especially in the case of a high/infinite-dimensional feature space.

**Lemma 11** *Let $\mathcal{H}$ be the class of functions computed by a $Q$-class linear (without bias) SVM on a domain $\mathcal{X}$. Let $\Delta\mathcal{H}$ be given by Definition 10. Suppose that $\Phi(\mathcal{X})$ is included in the ball of radius $\Lambda_{\Phi(\mathcal{X})}$ in $E_{\Phi(\mathcal{X})}$ and the vectors $w_k$ defining the hyperplanes satisfy $1/2 \max_{1\leq k<l\leq Q} \|w_k - w_l\|_{E_{\Phi(\mathcal{X})}} \leq \Lambda_w$. If a subset $s_m = \{x_i\}$, $(1 \leq i \leq m)$, of $\mathcal{X}$ is N-shattered with margin $\epsilon$ by $\Delta\mathcal{H}$, then there exists a subset $s'_m$ of $s_m$ of cardinality $m'$ at least equal to $\left\lceil \frac{m}{\binom{Q}{2}} \right\rceil$ such that for every partition of $s'_m$ into two subsets $s^{(1)}_{m'}$ and $s^{(2)}_{m'}$, the following bound holds true:*

$$\left\| \sum_{x_i \in s^{(1)}_{m'}} \Phi(x_i) - \sum_{x_i \in s^{(2)}_{m'}} \Phi(x_i) \right\|_{E_{\Phi(\mathcal{X})}} \geq \frac{\left\lceil \frac{m}{\binom{Q}{2}} \right\rceil \epsilon}{\Lambda_w} \tag{38}$$

**Proof** Suppose that $s_m = \{x_i\}$, $(1 \leq i \leq m)$, is a subset of $\mathcal{X}$ N-shattered with margin $\epsilon$ by $\Delta\mathcal{H}$. Let $(I(s_m), v_b)$ witness this shattering. According to the pigeonhole principle, there is at least one couple of indexes $(k, l)$ with $1 \leq k < l \leq Q$ such that there are at least $m' = \left\lceil \frac{m}{\binom{Q}{2}} \right\rceil$ points in $s_m$ for which the couple $(i_1(x_i), i_2(x_i))$ is $(k, l)$. For the sake of simplicity, the points in $s_m$ are reordered in such a way that the $m'$ first of them exhibit this property. The corresponding subset of $s_m$ is denoted $s_{m'}$. This means that for all vector $v_y$ in $\{-1, 1\}^m$, there is a function $h_y$ in $\mathcal{H}$ characterized by the vectors $w_{y,n}$, $(1 \leq n \leq Q)$, such that, for all $i$ in $\{1, \ldots, m'\}$:

$$\begin{cases} \text{if } y_i = 1 \text{ then} & (\Delta h_{y,k}(x_i) - b_i \geq \epsilon \wedge \Delta h_{y,l}(x_i) = -\Delta h_{y,k}(x_i)) \\ \text{if } y_i = -1 \text{ then} & (\Delta h_{y,k}(x_i) - b_i \leq -\epsilon \wedge \Delta h_{y,l}(x_i) = -\Delta h_{y,k}(x_i)) \end{cases}$$

But $\Delta h_{y,l}(x_i) = -\Delta h_{y,k}(x_i)$ implies that $h_{y,k}(x) = \max_n h_{y,n}(x)$ and $h_{y,l}(x) = \max_{n \neq k} h_{y,n}(x)$ or the contrary, and by way of consequence, $\Delta h_{y,k}(x_i) = 1/2\langle w_{y,k} - w_{y,l}, \Phi(x_i)\rangle$ and $\Delta h_{y,l}(x_i) = 1/2\langle w_{y,l} - w_{y,k}, \Phi(x_i)\rangle$. Note that this step of the proof does not hold anymore if one uses the $\Delta^*$ operator in place of the $\Delta$ operator. This is the reason why it is specifically the $\Delta$ 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). Thus, for all $i$ in $\{1, \ldots, m'\}$,

$$\begin{cases} \text{if } y_i = 1 \text{ then} & 1/2\langle w_{y,k} - w_{y,l}, \Phi(x_i)\rangle - b_i \geq \epsilon \\ \text{if } y_i = -1 \text{ then} & 1/2\langle w_{y,l} - w_{y,k}, \Phi(x_i)\rangle + b_i \geq \epsilon \end{cases}$$

Consider now any partition of $s_{m'}$ into to subsets $s_{m'}^{(1)}$ and $s_{m'}^{(2)}$. Consider any vector $v_y$ in $\{-1,1\}^m$ such that $y_i = 1$ if $x_i \in s_{m'}^{(1)}$ and $y_i = -1$ if $x_i \in s_{m'}^{(2)}$. We have thus:

$$1/2\langle w_{y,k} - w_{y,l}, \sum_{x_i \in s_{m'}^{(1)}} \Phi(x_i)\rangle - \sum_{x_i \in s_{m'}^{(1)}} b_i + 1/2\langle w_{y,l} - w_{y,k}, \sum_{x_i \in s_{m'}^{(2)}} \Phi(x_i)\rangle + \sum_{x_i \in s_{m'}^{(2)}} b_i \geq |s_{m'}|\,\epsilon$$

which simplifies into

$$1/2\langle w_{y,k} - w_{y,l}, \sum_{x_i \in s_{m'}^{(1)}} \Phi(x_i) - \sum_{x_i \in s_{m'}^{(2)}} \Phi(x_i)\rangle - \sum_{x_i \in s_{m'}^{(1)}} b_i + \sum_{x_i \in s_{m'}^{(2)}} b_i \geq |s_{m'}|\,\epsilon$$

Conversely, consider any vector $v_y$ such that $y_i = -1$ if $x_i \in s_{m'}^{(1)}$ and $y_i = 1$ if $x_i \in s_{m'}^{(2)}$. We have:

$$1/2\langle w_{y,l} - w_{y,k}, \sum_{x_i \in s_{m'}^{(1)}} \Phi(x_i) - \sum_{x_i \in s_{m'}^{(2)}} \Phi(x_i)\rangle + \sum_{x_i \in s_{m'}^{(1)}} b_i - \sum_{x_i \in s_{m'}^{(2)}} b_i \geq |s_{m'}|\,\epsilon$$

As a consequence, if $\sum_{x_i \in s_{m'}^{(1)}} b_i - \sum_{x_i \in s_{m'}^{(2)}} b_i \geq 0$, there is a function $h_y$ in $\mathcal{H}$ such that

$$1/2\langle w_{y,k} - w_{y,l}, \sum_{x_i \in s_{m'}^{(1)}} \Phi(x_i) - \sum_{x_i \in s_{m'}^{(2)}} \Phi(x_i)\rangle \geq \left\lceil \frac{m}{\binom{Q}{2}} \right\rceil \epsilon \tag{39}$$

whereas if $\sum_{x_i \in s_{m'}^{(1)}} b_i - \sum_{x_i \in s_{m'}^{(2)}} b_i < 0$, there is another function $h_y$ in $\mathcal{H}$ such that

$$1/2\langle w_{y,l} - w_{y,k}, \sum_{x_i \in s_{m'}^{(1)}} \Phi(x_i) - \sum_{x_i \in s_{m'}^{(2)}} \Phi(x_i)\rangle \geq \left\lceil \frac{m}{\binom{Q}{2}} \right\rceil \epsilon \tag{40}$$

Finally, applying the Cauchy-Schwartz inequality to both (39) and (40), it springs that whatever the sign of $\sum_{x_i \in s_{m'}^{(1)}} b_i - \sum_{x_i \in s_{m'}^{(2)}} b_i$ is,

$$\frac{1}{2}\,\|w_k - w_l\|_{E_{\Phi(\mathcal{X})}} \left\| \sum_{x_i \in s_{m'}^{(1)}} \Phi(x_i) - \sum_{x_i \in s_{m'}^{(2)}} \Phi(x_i) \right\|_{E_{\Phi(\mathcal{X})}} \geq \left\lceil \frac{m}{\binom{Q}{2}} \right\rceil \epsilon$$

from which (38) directly springs, as a consequence of the constraint on $\|w_k - w_l\|_{E_{\Phi(\mathcal{X})}}$. ∎

**Lemma 12 (Lemma 4.3 in [8])** *For all subset $s_m = \{x_i\}$, $(1 \leq i \leq m)$, of $\mathcal{X}$ such that $\max_{x_i \in s_m} \|\Phi(x_i)\|_{E_{\Phi(\mathcal{X})}} \leq \Lambda_{\Phi(s_m)}$, $s_m$ can be split into two non overlapping subsets $s_{m,+}$ and $s_{m,-}$ satisfying*

$$\left\| \sum_{x_i \in s_{m,+}} \Phi(x_i) - \sum_{x_i \in s_{m,-}} \Phi(x_i) \right\|_{E_{\Phi(\mathcal{X})}} \leq \sqrt{m}\Lambda_{\Phi(s_m)} \tag{41}$$

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

**Theorem 5** *Let $\mathcal{H}$ be the class of functions computed by a $Q$-class linear (without bias) SVM on a domain $\mathcal{X}$. Let $\Delta\mathcal{H}$ be given by Definition 10. Suppose that $\Phi(\mathcal{X})$ is included in the ball of radius $\Lambda_{\Phi(\mathcal{X})}$ in $E_{\Phi(\mathcal{X})}$ and the vectors $w_k$ defining the hyperplanes satisfy $1/2 \max_{1\leq k<l\leq Q} \|w_k - w_l\|_{E_{\Phi(\mathcal{X})}} \leq \Lambda_w$. Then, for any positive real value $\epsilon$, the following bound holds true:*

$$N\text{-}dim(\Delta\mathcal{H}, \epsilon) \leq \binom{Q}{2} \left(\frac{\Lambda_w \Lambda_{\Phi(\mathcal{X})}}{\epsilon}\right)^2 \tag{42}$$

**Proof** Let $s_m$ be a subset of $\mathcal{X}$ of cardinality $m$ N-shattered with margin $\epsilon$ by $\Delta\mathcal{H}$. According to Lemma 11, there is at least a subset $s_{m'}$ of $s_m$ of cardinality $m' = \left\lceil \frac{m}{\binom{Q}{2}} \right\rceil$ satisfying (38) for all its partitions into two subsets $s_{m'}^{(1)}$ and $s_{m'}^{(2)}$. Since, according to Lemma 12, there is at least one of these partitions for which (41) holds true,

$$\frac{\left\lceil \frac{m}{\binom{Q}{2}} \right\rceil \epsilon}{\Lambda_w} \leq \sqrt{\left\lceil \frac{m}{\binom{Q}{2}} \right\rceil} \Lambda_{\Phi(s_{m'})}$$

and thus

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

which concludes the proof. ∎

Note that in the bi-class case, using the notations of Subsection 4.3, (42) becomes

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

with is precisely the bound provided by Theorem 4.6 in [8] (see also Remark 1 in [31]). 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

So far, the results of this report have involved two distinct margin operators, $\Delta$ and $\Delta^*$. 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, the generalized Sauer-Shela lemma, only stands for $\Delta^*$. On the contrary, we have seen in Subsection 6.2 that the proof of Theorem 5, the bound on the margin Natarajan dimension, was making use of a specific property of $\Delta$. These observations highlight the qualities and shortcomings of Definition 10 and Definition 3. Loosely speaking, the advantage of $\Delta$ over $\Delta^*$ rests in the fact that it keeps available the index of the second highest output. The drawback springs from the fact that too much "useless" information is kept (namely the differences between the lowest outputs and the highest one), which introduces some "noise" in the computations. Now, it is difficult to imagine a way to bound the margin Natarajan dimension of $\Delta^*\mathcal{H}$, where $\mathcal{H}$ is a multivariate linear model, as a function of the constraints on the vectors $w_k$. This is due to the fact that $\Delta^* h_l$ cannot easily be expressed in terms of the vectors $w_k$ (just notice that it does not necessarily involve vector $w_l$). On the contrary, one could expect that the choice of an appropriate pseudo-metric could reestablish the connection between separation and shattering capacity, for the functions of the form $\Delta h$. This should provide us with a bound similar to (26), and, by way of consequence, with a generalization of the Sauer-Shela lemma applying to the classes $\Delta\mathcal{H}$. Such a result, which would complete our theoretical framework, could also prove interesting in its own right.

## 7.1 On the theoretical grounding of the M-SVMs of the literature

Theorem 5 highlights the fact that the functional $\max_{k<l} \|w_k - w_l\|^2$, or alternatively $\sum_{k<l}^{Q} \|w_k - w_l\|^2$, plays for M-SVMs a role similar to the one played by $\|w\|^2$ for the standard binary SVM. This is satisfactory indeed, since both functions are convex. Their use as control term in the objective function of the training procedure, as was done in [28, 25], is thus once more justified. In [72, 68, 17, 43], the functional selected to perform the capacity control is slightly different, since it is $\sum_{k=1}^{Q} \|w_k\|^2$, whereas in [13], the authors used instead $\sum_{k<l}^{Q} \|w_k - w_l\|^2 + \sum_{k=1}^{Q} \|w_k\|^2$. Can the theorems derived here justify these choices as well ? This is the case indeed. For instance, it was proved in [25] (see also [35]) that the machines introduced in [72, 68, 13, 28], in spite of their different formulations, are utterly equivalent, since they all generate the same optimal solution, provided that the value of their soft margin parameter $C$ is selected appropriately. Furthermore, variants of Theorem 5 can easily be derived, to fit more precisely a given training algorithm (penalty term). We have thus here the basis to endow all the M-SVMs published so far with a well founded theoretical justification, which should make it possible to compare their performance on a sound basis.

# 8 Conclusions and Future Work

This paper has introduced a new uniform convergence result for the empirical risk of large margin multi-category discriminant models. The measure of capacity it involves, a covering number, can be bounded in terms of different scale-sensitive $\Psi$-dimensions, thanks to generalized Sauer-Shela 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. This study thus paves the way for new strides in our attempt to endow all the training algorithms (objective functions) proposed so far for these machines with a unifying theory.

Indeed, the main contribution of this study is a new characterization of the variation of the capacity of a multivariate affine model as a function of the constraints on its parameters. In that sense, it completes previous works on the same subject [25], which had followed another path, namely the computation of a bound on the entropy numbers of a linear operator [73, 30]. However, sharper bounds should obviously result from using concentration inequalities derived in the framework of the empirical process theory [60, 61, 41, 45]. This makes it possible, for instance, to work with data dependent capacity measures such as the empirical VC entropy. In this field, the most promising studies are probably those reported in [11, 12]. The latest developments on the analysis of the learning rate of an empirical risk minimizer can be found in [62, 46] (see also [57] for the specific case of pattern recognition SVMs). More generally, the study of model selection based on penalized empirical loss minimization, as presented for instance in [7], should also prove particularly fruitful. In parallel with the derivation of a generalized Sauer-Shela lemma devoted to $\Delta\mathcal{H}$, the extension of these works to the case we are interested in, and the subsequent comparisons, are the subject of an ongoing work.

**Acknowledgements**
It is a pleasure to thank S. Kroon and R. Vert for discussions and bibliographical help. Thanks are also due to E. Domenjoud and A. Elisseeff for carefully reading 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(2):113–141, 2001.

[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. *Artificial Neural Network Learning: Theoretical Foundations*. Cambridge University Press, 1999.

[6] 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.

[7] P.L. Bartlett, S. Boucheron, and G. Lugosi. Model selection and error estimation. *Machine Learning*, 48:85–113, 2002.

[8] 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.

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

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

[11] S. Boucheron, G. Lugosi, and P. Massart. A sharp concentration inequality with applications. Technical Report NC2-TR-1999-057, NeuroCOLT2, 1999.

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

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

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

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

[16] K. Crammer and Y. Singer. On the learnability and design of output codes for multiclass problems. In *Proceedings of the Thirteen Annual Conference on Computational Learning Theory (COLT)*, pages 35–46, 2000.

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

[18] L. Devroye, L. Györfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition.* Springer, 1996.

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

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

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

[22] 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).

[23] J. Gapaillard. *Intégration pour la licence.* Masson, 1997. (in French).

[24] 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.

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

[26] Y. Guermeur. A simple unifying theory of multi-class support vector machines. Technical Report RR-4669, INRIA, 2002.

[27] 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.

[28] 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.

[29] Y. Guermeur, A. Elisseeff, and D. Zelus. A comparative study of multi-class support vector machines in the unifying framework of large margin classifiers. *Appl. Stochastic Models Bus. Ind.*, 20, 2004. (in press).

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

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

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

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

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

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

[36] 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.

[37] 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.

[38] 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.

[39] A.N. Kolmogorov and V.M. Tihomirov. $\epsilon$-entropy and $\epsilon$-capacity of sets in functional spaces. *Amer. Math. Soc. Translations (2)*, 17:277–364, 1961.

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

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

[42] 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.

[43] Y. Lee, Y. Lin, and G. Wahba. Multicategory support vector machines. Technical Report 1043, University of Wisconsin, Madison, Department of Statistics, 2001.

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

[45] G. Lugosi. Concentration-of-measure inequalities. Lecture notes, Summer School on Machine Learning at the Australian National University, Canberra, 2003.

[46] P. Massart and E. Nédélec. Risk bounds for statistical learning. Preprint of the "Laboratoire de Mathématiques", Université Paris-Sud, 2003.

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

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

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

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

[51] 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.

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

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

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

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

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

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

[58] 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.

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

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

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

[62] A.B. Tsybakov. Optimal aggregation of classifiers in statistical learning. *Annals of Statistics*, 32(1), 2004.

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

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

[65] 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.

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

[67] 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.

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

[69] 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.

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

[71] 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.

[72] 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.

[73] 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 Trans. on Information Theory*, 47(6):2516–2532, 2001.

# Contents

**1 Introduction** **3**

**2 Margin Risk for Multi-category Discriminant Models** **4**
2.1 Formalization of the learning problem . . . . . 4
2.2 Multi-class margin . . . . . 4
2.3 Capacity measure: covering number . . . . . 6

**3 Uniform Convergence of the Empirical Margin Risk** **7**
3.1 First symmetrization . . . . . 7
3.2 Second symmetrization . . . . . 9
3.3 Maximal inequality . . . . . 11
3.4 Exponential bound . . . . . 12
3.5 Uniform bound over the margin $\gamma$ . . . . . 14
3.6 Choice of the "margin" operator . . . . . 15

**4 Scale-sensitive Ψ-dimensions** **17**
4.1 Ψ-dimensions . . . . . 17
4.2 Margin Ψ-dimensions . . . . . 18
4.3 Discussion . . . . . 20

**5 Relating the Covering Number and the Margin Natarajan Dimension** **23**
5.1 Definitions . . . . . 23
5.2 Lemmas . . . . . 24
5.3 Classes of $\delta$ functions . . . . . 27
5.4 Generalized Sauer-Shela lemma . . . . . 29
5.5 First upper bound on the covering number of $\Delta_\gamma^* \mathcal{H}$ . . . . . 31
5.6 Standard bound on function $\phi$ . . . . . 33
5.7 Main theorem and discussion . . . . . 35

**6 Margin Natarajan Dimension of the Multi-class SVMs** **36**
6.1 Architecture and training of the M-SVMs . . . . . 36
6.2 Upperbounding the margin Natarajan dimension of $\Delta_\gamma \mathcal{H}$ . . . . . 36

**7 Guaranteed Risk and Implementation of the SRM Inductive Principle** **40**
7.1 On the theoretical grounding of the M-SVMs of the literature . . . . . 40

**8 Conclusions and Future Work** **41**

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