A Simple Unifying Theory of Multi-Class Support Vector Machines

Yann Guermeur

INRIA

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

A Simple Unifying Theory of Multi-Class Support Vector Machines

Yann Guermeur

N° 4669

December 2002

THÈME 2

ISSN 0249-6399 ISRN INRIA/RR--4669--FR+ENG

A Simple Unifying Theory of Multi-Class Support Vector Machines

Yann Guermeur*

Thème 2 — Génie logiciel
et calcul symbolique
Projet MODBIO

Rapport de recherche n° 4669 — December 2002 — 35 pages

Abstract: Vapnik's statistical learning theory has mainly been developed for two types of problems: pattern recognition (computation of dichotomies) and regression (estimation of real-valued functions). Only in recent years has multi-class discriminant analysis been studied independently. Extending several standard results, among which a famous theorem by Bartlett, we have derived distribution-free uniform strong laws of large numbers devoted to multi-class large margin discriminant models. This technical report deals with the computation of the capacity measures involved in these bounds on the expected risk. Straightforward extensions of results regarding large margin classifiers highlight the central role played by a new generalized VC dimension, which can be seen either as an extension of the fat-shattering dimension to the multivariate case, or as a scale-sensitive version of the graph dimension. The theorems derived are applied to the architecture shared by all the multi-class SVMs proposed so far, which provides us with a simple theoretical framework to study them, compare their performance and design new machines.

Key-words: Multi-class support vector machines, uniform strong laws of large numbers, structural risk minimization inductive principle

* Université Henri Poincaré - Nancy I

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

Une théorie unificatrice simple des machines à vecteurs support multi-classes

Résumé : La théorie statistique de l'apprentissage de Vapnik a principalement été développée pour deux types de problèmes : la reconnaissance des formes (calcul de dichotomies) et la régression (estimation de fonctions à valeurs réelles). Ce n'est que récemment que l'analyse discriminante à catégories multiples a fait l'objet d'études spécifiques. Etendant un ensemble de résultats standard, parmi lesquels un célèbre théorème de Bartlett, nous avons établi des lois fortes des grands nombres uniformes indépendantes de la distribution sous-jacente et dédiées aux systèmes de discrimination multi-classe à grande marge. L'objet de ce rapport est la calcul de la mesure de capacité intervenant dans ces bornes sur le risque. L'extension naturelle de résultats portant sur les classifieurs à grande marge permet de mettre en évidence le rôle central joué par une nouvelle dimension de Vapnik-Chervonenkis étendue, qui représente à la fois une extension de la "fat-shattering dimension" au cas multi-classe et une extension de la dimension graphique incorporant la notion de marge. Les théorèmes établis sont appliqués à l'architecture commune à l'ensemble des M-SVM publiées à ce jour. Ceci permet de les munir d'un cadre théorique unificateur, d'étudier leur comportement d'un point de vue à la fois théorique et pratique, et de proposer de nouvelles machines.

Mots-clés : Machines à vecteurs support multi-classes, lois fortes des grands nombres uniformes, principe inductif de minimisation structurelle du risque

1 Introduction

One of the pioneering contributions to the study of the generalization capabilities of infinite sets of functions is the work of Vapnik and Chervonenkis [50] relating the consistency of the empirical risk minimization (ERM) inductive principle to the finiteness of a simple combinatorial quantity called the Vapnik-Chervonenkis dimension. Since then, the consistency of the ERM principle has been analysed in various contexts [8, 2]. Concomitantly, many studies have been devoted to deriving bounds on the expected risk (computing sample complexities), or implementing the structural risk minimization (SRM) inductive principle [46]. Among the main contributions are [43, 4]. However, the case of multi-class discrimination has seldom been studied independently [42], although we pointed out in [20] the fact that it is inappropriate to tackle it in the straightforward manner, by plugging a generalized VC dimension as capacity measure in a standard uniform convergence bound. In [17], we have extended to the multi-class case a lemma by Bartlett [4] expressing the sample complexity of pattern classification models in terms of a margin-based covering number. Deriving bounds on such covering numbers is the subject of this report. The organization of the paper is as follows. Section 2 briefly summarizes our initial uniform convergence result. Section 3 is devoted to improving this theorem, by changing the pseudo-metric. Building upon the main lemma of [2], we derive in Section 4 a bound on the covering numbers of interest, in terms of a new capacity measure. This measure can be seen either as an extension of the fat-shattering dimension to the multivariate case, or a scale-sensitive version of the graph dimension. In Section 5, this bound is applied to the architecture shared by all the multi-class support vector machines (M-SVMs) described in literature. This provides us with a new justification of the principles of their training algorithms, which can thus be simply casted in the framework of the SRM inductive principle. At last, Section 6 deals with alternative possibilities to compute sample complexities.

2 Guaranteed Risk for Multi-Class Discriminant Models

We consider the case of a Q-category pattern recognition problem, where $Q \geq 3$. Let $\mathcal{X}$ be the space of description and $\mathcal{C}$ the set of categories. We make the assumption, standard in statistical learning theory, that there is a joint probability distribution P, 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 lowest "error rate". The "error rate" of a function h is the error rate 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 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 [37]), applying this decision function simply amounts to implementing Bayes' estimated decision rule. Hereafter, $C(x_i)$ will denote indifferently the category of pattern x_i, or the index of this category. $\mathcal{Y} = \{y\}$

will be the set of canonical codings of the categories in $\{-1,1\}^Q$ vectors. The uniform convergence result we established is based on the following definitions.

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 P, i.e.:*

$$R(f) = \int_{\mathcal{X}\times\mathcal{C}} \mathbf{I}_{\{f(x)\neq C\}} dP(x, C) \tag{1}$$

where the indicator function $\mathbf{I}_{\{f(x)\neq C\}}$ is defined as follows: $\mathbf{I}_{\{f(x)\neq C\}} = 1$ if $f(x) \neq C$, $\mathbf{I}_{\{f(x)\neq C\}} = 0$ otherwise.

Definition 2 (Empirical risk) *Let $s_m = \{(x_i, C(x_i))\} \in (\mathcal{X}\times\mathcal{C})^m$. The* empirical risk *of f on s_m is defined as:*

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

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.

Definition 3 (ϵ-cover or ϵ-net) *Let (E, ρ) be a pseudo-metric space, and $B(v, r)$ the closed ball of center v and radius r in E. Let H be a subset of E. An* ϵ-cover *of H is a subset $\bar{H}$ of E such that:*

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

Definition 4 (Covering numbers) *Let (E, ρ) be a pseudo-metric space. If $H \subset E$ has an ϵ-cover of finite cardinality, then its* covering number $\mathcal{N}(\epsilon, H, \rho)$ *is the smallest cardinality of its ϵ-covers. If there is no such finite cover, then the covering number is defined to be ∞.*

Definition 5 *Let $\mathcal{H}$ be a set of functions from $\mathcal{X}$ into $\mathbb{R}^Q$. For a set s of points in $\mathcal{X}$, define the pseudo-metric $d_{l_\infty, l_1(s)}$ on $\mathcal{H}$ as:*

$$\forall (h, \bar{h}) \in \mathcal{H}^2, \ d_{l_\infty, l_1(s)}(h, \bar{h}) = \max_{x\in s} \sum_{k=1}^{Q} \left|h_k(x) - \bar{h}_k(x)\right|$$

Definition 6 (Canonical function) *For all $h \in \mathcal{H}$ and all $x \in \mathcal{X}$, let $M_1(h, x)$ be the smallest index l such that $h_l(x) = \max_k h_k(x)$ and $M_2(h, x)$ the smallest index $l \neq M_1(h, x)$ such that $h_l(x) = \max_{k\neq M_1(h,x)} h_k(x)$. Define $\Delta h = [\Delta h_k]$, $(1 \leq k \leq Q)$, as the function from $\mathcal{X}$ into $\mathbb{R}^Q$, satisfying*

$$\Delta h_k(x) = \begin{cases} \frac{1}{2}\left(h_k(x) - h_{M_2(h,x)}(x)\right) & if\ k = M_1(h, x) \\ \frac{1}{2}\left(h_k(x) - h_{M_1(h,x)}(x)\right) & otherwise \end{cases}$$

Extending a definition from Bartlett [4], we introduced the following definition:

Definition 7 (Empirical margin risk) *The empirical risk with margin $\gamma \in (0,1]$ of h on a set s_m of size m is*

$$R^{\gamma}_{s_m}(h) = \frac{1}{m} \left| \{ (x_i, C(x_i)) \in s_m \ / \ \Delta h_{C(x_i)}(x_i) < \gamma \} \right|$$

For $\gamma \in (0,1]$, let $\pi_\gamma : \mathbb{R} \to [-\gamma, \gamma]$ be the piecewise-linear squashing function defined as

$$\pi_\gamma(x) = \begin{cases} \gamma.sign(x) & if \ |x| \geq \gamma \\ x & otherwise \end{cases}$$

$\forall h \in \mathcal{H}$, $\Delta h^\gamma = [\Delta h^\gamma_k] = [\pi_\gamma \circ \Delta h_k]$, $(1 \leq k \leq Q)$. $\Delta\mathcal{H}^\gamma = \{\Delta h^\gamma \ / \ h \in \mathcal{H}\}$. Let $\mathcal{N}_{\infty,1}(\gamma/2, \Delta\mathcal{H}^\gamma, m) = \max_{s_m \in \mathcal{X}^m} \mathcal{N}(\gamma/2, \Delta\mathcal{H}^\gamma, d_{l_\infty, l_1(s_m)})$. With these hypotheses and definitions at hand, extending Lemma 4 and Corollary 9 from [4], as well as the basic lemma of Theorem 4.1 in [48], we established in [17] the following theorem:

Theorem 1 *Let s_m be a m-sample of examples drawn independently from P. With probability at least $1-\delta$, for every value of γ in $(0,1]$, the risk $R(h)$ of a function h computed by a numerical Q-class discriminant model $\mathcal{H}$ is bounded above by:*

$$R(h) \leq R^{\gamma}_{s_m}(h) + \sqrt{\frac{1}{2m}\left(\ln\left(2\mathcal{N}_{\infty,1}(\gamma/2, \Delta\mathcal{H}^\gamma, 2m)\right) + \ln\left(\frac{2}{\gamma\delta}\right)\right)} + \frac{1}{m} \qquad (3)$$

3 Improved Convergence Result

In this section, we establish that Theorem 1 can be simply improved by changing the pseudo-metric for 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 $\mathcal{X}$, define the pseudo-metric $d_{l_\infty, l_\infty(s)}$ on $\mathcal{H}$ as:*

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

Let $\mathcal{N}_{\infty,\infty}(\gamma/2, \Delta\mathcal{H}^\gamma, m) = \max_{s_m \in \mathcal{X}^m} \mathcal{N}(\gamma/2, \Delta\mathcal{H}^\gamma, d_{l_\infty, l_\infty(s_m)})$. We then get:

Theorem 2 *Let s_m be a m-sample of examples drawn independently from P. With probability at least $1-\delta$, for every value of γ in $(0,1]$, the risk $R(h)$ of a function h computed by a numerical Q-class discriminant model $\mathcal{H}$ is bounded above by:*

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

The sketch of the proof is unchanged. We detail it here for the sake of completeness.
Proof Let θ denotes the step-function

$$\theta(x) = \begin{cases} 1 \; if \; x \geq 0 \\ 0 \; otherwise \end{cases}$$

and let $\hat{\theta}$ be defined as:

$$\hat{\theta}(x) = \begin{cases} 1 \; if \; x > 0 \\ 0 \; otherwise \end{cases}$$

The empirical margin risk has been introduced in the previous section. The *margin risk* is the corresponding extension to the whole product space $S = \mathcal{X} \times \mathcal{C}$ (in what follows, S will indifferently designate $\mathcal{X} \times \mathcal{C}$ or $\mathcal{X} \times \mathcal{Y}$).

Definition 9 (Margin risk)

$$\begin{aligned} R^{\gamma}(h) &= P\left(\exists k \in \{1,..,Q\} \; / \; \Delta h_k(x) y_k < \gamma\right) \\ &= \int_S \hat{\theta}\left[Q - \sum_{k=1}^{Q} \theta\left(\Delta h_k(x) y_k - \gamma\right)\right] dP(z) \end{aligned}$$

where we have set $z = (x, y) \in S$. Note that we cannot simplify the expression as in the case of the empirical margin risk, since we have made the assumption of a joint probability distribution on $\mathcal{X}$ and $\mathcal{C}$, not a functional dependence, with the consequence that the descriptions are not associated with a single category. The following risks will also prove useful:

$$\begin{aligned} \mathcal{R}^{\gamma}(h) &= P\left(\exists k \in \{1,..,Q\} \; / \; |\pi_{\gamma}(\Delta h_k(x)) - \gamma y_k| \geq \gamma\right) \\ &= \int_S \hat{\theta}\left[\sum_{k=1}^{Q} \theta\left(|\pi_{\gamma}\left(\Delta h_k(x)\right) - \gamma y_k| - \gamma\right)\right] dP(z) \end{aligned}$$

$$\mathcal{R}^{\gamma}_{s_m}(h) = \frac{1}{m} \left| \left\{ (x_i, y_i) \in s_m \; / \; \sum_{k=1}^{Q} \theta\left(|\pi_{\gamma}\left(\Delta h_k(x_i)\right) - \gamma y_{ik}| - \gamma\right) \neq 0 \right\} \right|$$

We have

$$\forall \gamma > 0, \;\; \Delta h_k(x) y_k \leq 0 \Longrightarrow \Delta h_k(x) y_k - \gamma < 0$$

Thus

$$\forall \gamma > 0, \; R(h) \leq R^{\gamma}(h)$$

Furthermore

$$(\gamma > 0 \; \wedge \; \Delta h_k(x) y_k \leq 0) \Longrightarrow |\Delta h_k^{\gamma}(x) - \gamma y_k| = |\Delta h_k^{\gamma}(x)| + \gamma\,|y_k| = |\Delta h_k^{\gamma}(x)| + \gamma \geq \gamma$$

In short

$$\forall \gamma > 0, \; R(h) \leq \mathcal{R}^{\gamma}(h)$$

First symmetrization

Taking our inspiration from the proof of Vapnik's basic lemma in [48] (Section 4.5), we prove the following result:

Lemma 1

$$P_{s_m}\left(\sup_{h\in\mathcal{H}}\left(\mathcal{R}^\gamma(h)-R^\gamma_{s_m}(h)\right)\geq\epsilon\right)\leq 2P_{s_m,\tilde{s}_m}\left(\sup_{h\in\mathcal{H}}\left(\mathcal{R}^\gamma_{\tilde{s}_m}(h)-R^\gamma_{s_m}(h)\right)\geq\epsilon-\frac{1}{m}\right)$$

where P_{s_m} is a probability over the training set s_m and $P_{s_m,\tilde{s}_m}$ is a probability over a learning set $s_m\cup\tilde{s}_m$ of size $2m$. Indeed, by definition:

$$P_{s_m,\tilde{s}_m}\left(\sup_{h\in\mathcal{H}}\left(\mathcal{R}^\gamma_{\tilde{s}_m}(h)-R^\gamma_{s_m}(h)\right)\geq\epsilon-\frac{1}{m}\right)=$$

$$\int_{S^{2m}}\theta\left[\sup_{h\in\mathcal{H}}\left(\mathcal{R}^\gamma_{\tilde{s}_m}(h)-R^\gamma_{s_m}(h)\right)-\epsilon+\frac{1}{m}\right]dP(s_m,\tilde{s}_m)$$

Applying Fubini's theorem for nonnegative measurable functions [18] to the product measure $P_{s_m,\tilde{s}_m}$ yields:

$$P_{s_m,\tilde{s}_m}\left(\sup_{h\in\mathcal{H}}\left(\mathcal{R}^\gamma_{\tilde{s}_m}(h)-R^\gamma_{s_m}(h)\right)\geq\epsilon-\frac{1}{m}\right)=$$

$$\int_{S^m}dP(s_m)\int_{S^m}\theta\left[\sup_{h\in\mathcal{H}}\left(\mathcal{R}^\gamma_{\tilde{s}_m}(h)-R^\gamma_{s_m}(h)\right)-\epsilon+\frac{1}{m}\right]dP(\tilde{s}_m)$$

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

$$\mathcal{Q}=\left\{s_m\in S^m\ /\ \sup_{h\in\mathcal{H}}\left(\mathcal{R}^\gamma(h)-R^\gamma_{s_m}(h)\right)\geq\epsilon\right\}$$

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

$$P_{s_m,\tilde{s}_m}\left(\sup_{h\in\mathcal{H}}\left(\mathcal{R}^\gamma_{\tilde{s}_m}(h)-R^\gamma_{s_m}(h)\right)\geq\epsilon-\frac{1}{m}\right)\geq$$

$$\int_{\mathcal{Q}}dP(s_m)\underbrace{\int_{S^m}\theta\left[\sup_{h\in\mathcal{H}}\left(\mathcal{R}^\gamma_{\tilde{s}_m}(h)-R^\gamma_{s_m}(h)\right)-\epsilon+\frac{1}{m}\right]dP(\tilde{s}_m)}_{I}\qquad(5)$$

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

$$\sup_{h\in\mathcal{H}}\left(\mathcal{R}^\gamma(h)-R^\gamma_{s_m}(h)\right)\geq\epsilon$$

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

$$\mathcal{R}^\gamma(h^*)-R^\gamma_{s_m}(h^*)\geq\epsilon$$

By definition of h^*, the following inequality holds

$$I \geq \int_{S^m} \theta \left[\mathcal{R}^\gamma_{\tilde{s}_m}(h^*) - R^\gamma_{s_m}(h^*) - \epsilon + \frac{1}{m} \right] dP(\tilde{s}_m)$$

Since $\mathcal{R}^\gamma_{\tilde{s}_m}(h^*) - \mathcal{R}^\gamma(h^*) + \frac{1}{m} \leq \mathcal{R}^\gamma_{\tilde{s}_m}(h^*) - \mathcal{R}^\gamma(h^*) + \frac{1}{m} + \mathcal{R}^\gamma(h^*) - R^\gamma_{s_m}(h^*) - \epsilon$, or equivalently, $\mathcal{R}^\gamma_{\tilde{s}_m}(h^*) - \mathcal{R}^\gamma(h^*) + \frac{1}{m} \leq \mathcal{R}^\gamma_{\tilde{s}_m}(h^*) - R^\gamma_{s_m}(h^*) - \epsilon + \frac{1}{m}$, this implies that

$$I \geq \int_{S^m} \theta \left[\mathcal{R}^\gamma_{\tilde{s}_m}(h^*) - \mathcal{R}^\gamma(h^*) + \frac{1}{m} \right] dP(\tilde{s}_m)$$

Furthermore

$$\int_{S^m} \theta \left[\mathcal{R}^\gamma_{\tilde{s}_m}(h^*) - \mathcal{R}^\gamma(h^*) + \frac{1}{m} \right] dP(\tilde{s}_m) = P_{\tilde{s}_m} \left(\mathcal{R}^\gamma_{\tilde{s}_m}(h^*) \geq \mathcal{R}^\gamma(h^*) - \frac{1}{m} \right)$$

By definition,

$$P_{\tilde{s}_m} \left(\mathcal{R}^\gamma_{\tilde{s}_m}(h^*) \geq \mathcal{R}^\gamma(h^*) - \frac{1}{m} \right) = \sum_{l \geq m\mathcal{R}^\gamma(h^*)-1} \binom{m}{l} \mathcal{R}^\gamma(h^*)^l \left(1 - \mathcal{R}^\gamma(h^*)\right)^{m-l}$$

Let X be a random variable described by a binomial distribution with parameters m and $\mathcal{R}^\gamma(h^*)$ ($X \hookrightarrow \mathcal{B}(m, \mathcal{R}^\gamma(h^*))$). Then $\mathbb{E}(X) = m\mathcal{R}^\gamma(h^*)$. By definition, the right-hand side of the equation above is thus greater than $P(X \geq \mathbb{E}(X))$, i.e. greater than $1/2$. Substituting this lower bound of I into (5) yields

$$P_{s_m, \tilde{s}_m} \left(\sup_{h \in \mathcal{H}} \left(\mathcal{R}^\gamma_{\tilde{s}_m}(h) - R^\gamma_{s_m}(h) \right) \geq \epsilon - \frac{1}{m} \right) \geq \frac{1}{2} \int_{\Omega} dP(s_m)$$

or equivalently, by definition of Ω:

$$P_{s_m, \tilde{s}_m} \left(\sup_{h \in \mathcal{H}} \left(\mathcal{R}^\gamma_{\tilde{s}_m}(h) - R^\gamma_{s_m}(h) \right) \geq \epsilon - \frac{1}{m} \right) \geq \frac{1}{2} P_{s_m} \left(\sup_{h \in \mathcal{H}} \left(\mathcal{R}^\gamma(h) - R^\gamma_{s_m}(h) \right) \geq \epsilon \right)$$

which is the result announced. From this lemma, the bound:

$$P_{s_m} \left(\sup_{h \in \mathcal{H}} \left(R(h) - R^\gamma_{s_m}(h) \right) \geq \epsilon \right) \leq 2 P_{s_m, \tilde{s}_m} \left(\sup_{h \in \mathcal{H}} \left(\mathcal{R}^\gamma_{\tilde{s}_m}(h) - R^\gamma_{s_m}(h) \right) \geq \epsilon - \frac{1}{m} \right) \qquad (6)$$

springs then directly from the fact that $R(h) \leq \mathcal{R}^\gamma(h)$. Note that this first symmetrization is the same as the one in [17]. Indeed, this part of the computation does not involve the pseudo-metric used on $\Delta\mathcal{H}^\gamma$.

Second symmetrization

Let us now consider the set S_t of all permutations σ over $\{1,\ldots,2m\}$ that realize transpositions between elements of same ranking in the first and second half of this set. Let $\mathcal{U}$ be the uniform distribution over the elements of S_t. For every sample $s_{2m} = (s_m, \tilde{s}_m) \in S^{2m}$, $s^\sigma_{2m} = (s^\sigma_m, \tilde{s}^\sigma_m) = \left\{z_{\sigma(1)},\ldots,z_{\sigma(2m)}\right\}$ denotes its "range" by σ. Since the set $(s_m, \tilde{s}_m)$ is chosen according to the product probability measure $P_{s_m,\tilde{s}_m}$ over S^{2m}, the right-hand side of (6) is not affected by a permutation σ. Consequently

$$P_{s_m,\tilde{s}_m}\left(\sup_{h\in\mathcal{H}}\left(\mathcal{R}^\gamma_{\tilde{s}_m}(h) - R^\gamma_{s_m}(h)\right) \geq \epsilon - \frac{1}{m}\right) \leq$$

$$\sup_{s_m,\tilde{s}_m} \mathcal{U}\left(\left\{\sigma \in S_t \;/\; \sup_{h\in\mathcal{H}}\left(\mathcal{R}^\gamma_{\tilde{s}^\sigma_m}(h) - R^\gamma_{s^\sigma_m}(h)\right) \geq \epsilon - \frac{1}{m}\right\}\right) \tag{7}$$

In short, to bound from above the probability of the event of interest, we consider specifically the "worst" possible choice for the samples s_m and $\tilde{s}_m$ over S^{2m}.

The set $(s_m, \tilde{s}_m)$ being fixed, $\Delta\bar{\mathcal{H}}^\gamma$ will refer to a $\gamma/2$-cover of the set $\Delta\mathcal{H}^\gamma = \{\pi_\gamma(\Delta h) \;/\; h \in \mathcal{H}\}$ with respect to the pseudo-metric $d_{l_\infty,l_\infty(s_{2m})}$. This set will be of minimal cardinality, i.e. $\left|\Delta\bar{\mathcal{H}}^\gamma\right| = \mathcal{N}_{\infty,\infty}(\gamma/2, \Delta\mathcal{H}^\gamma, 2m)$. By definition

$$\forall \Delta h^\gamma \in \Delta\mathcal{H}^\gamma,\ \exists \Delta\bar{h}^\gamma \in \Delta\bar{\mathcal{H}}^\gamma \;/$$

$$\max\left(\max_{(x,y)\in s_m}\max_k |\Delta h^\gamma_k(x) - \Delta\bar{h}^\gamma_k(x)|,\ \max_{(\tilde{x},\tilde{y})\in\tilde{s}_m}\max_k |\Delta h^\gamma_k(\tilde{x}) - \Delta\bar{h}^\gamma_k(\tilde{x})|\right) \leq \frac{\gamma}{2}$$

For all $\left(\Delta h^\gamma, \Delta\bar{h}^\gamma\right) \in \Delta\mathcal{H}^\gamma \times \Delta\bar{\mathcal{H}}^\gamma$,

$$\left\{\begin{array}{l} \max_k |\Delta h^\gamma_k(\tilde{x}^\sigma) - \gamma\tilde{y}^\sigma_k| \geq \gamma \\ d_{l_\infty,l_\infty(s_{2m})}(\Delta h^\gamma, \Delta\bar{h}^\gamma) \leq \frac{\gamma}{2} \end{array}\right. \implies$$

$$\exists k_0 \in \{1,\ldots,Q\} \;/\; \left\{\begin{array}{l} |\Delta h^\gamma_{k_0}(\tilde{x}^\sigma) - \gamma\tilde{y}^\sigma_{k_0}| \geq \gamma \\ |\Delta h^\gamma_{k_0}(\tilde{x}^\sigma) - \Delta\bar{h}^\gamma_{k_0}(\tilde{x}^\sigma)| \leq \frac{\gamma}{2} \end{array}\right. \implies$$

$$\exists k_0 \in \{1,\ldots,Q\} \;/\; |\Delta\bar{h}^\gamma_{k_0}(\tilde{x}^\sigma) - \gamma\tilde{y}^\sigma_{k_0}| \geq \frac{\gamma}{2} \implies \max_k |\Delta\bar{h}^\gamma_k(\tilde{x}^\sigma) - \gamma\tilde{y}^\sigma_k| \geq \frac{\gamma}{2}$$

In short,

$$\left\{\begin{array}{l} \max_k |\Delta h^\gamma_k(\tilde{x}^\sigma) - \gamma\tilde{y}^\sigma_k| \geq \gamma \\ d_{l_\infty,l_\infty(s_{2m})}(\Delta h^\gamma, \Delta\bar{h}^\gamma) \leq \frac{\gamma}{2} \end{array}\right. \implies \max_k |\Delta\bar{h}^\gamma_k(\tilde{x}^\sigma) - \gamma\tilde{y}^\sigma_k| \geq \frac{\gamma}{2} \tag{8}$$

Similarly, the triangle inequality implies:

$$\left\{\begin{array}{l} \max_k |\Delta h^\gamma_k(x^\sigma) - \gamma y^\sigma_k| = 0 \\ d_{l_\infty,l_\infty(s_{2m})}(\Delta h^\gamma, \Delta\bar{h}^\gamma) \leq \frac{\gamma}{2} \end{array}\right. \implies \max_k |\Delta\bar{h}^\gamma_k(x^\sigma) - \gamma y^\sigma_k| \leq \frac{\gamma}{2}$$

From this, by considering the converse, we deduce that

$$\left\{ \begin{array}{l} \max_k \left|\Delta\bar{h}_k^\gamma(x^\sigma) - \gamma y_k^\sigma\right| > \frac{\gamma}{2} \\ d_{l_\infty, l_\infty(s_{2m})}(\Delta h^\gamma, \Delta\bar{h}^\gamma) \leq \frac{\gamma}{2} \end{array} \right. \Longrightarrow \max_k \left|\Delta h_k^\gamma(x^\sigma) - \gamma y_k^\sigma\right| \neq 0 \tag{9}$$

From (8), it springs that if $d_{l_\infty, l_\infty(s_{2m})}(\Delta h^\gamma, \Delta\bar{h}^\gamma) \leq \frac{\gamma}{2}$, then

$$\mathcal{R}^\gamma_{\tilde{s}^\sigma_m}(h) \leq \frac{1}{m}\left|\left\{(\tilde{x}_i^\sigma, \tilde{y}_i^\sigma) \in \tilde{s}_m^\sigma \;/\; \max_k \left|\Delta\bar{h}_k^\gamma(\tilde{x}_i^\sigma) - \gamma\tilde{y}_{ik}^\sigma\right| \geq \frac{\gamma}{2}\right\}\right|$$

Similarly, from (9), it springs that if $d_{l_\infty, l_\infty(s_{2m})}(\Delta h^\gamma, \Delta\bar{h}^\gamma) \leq \frac{\gamma}{2}$, then

$$\frac{1}{m}\left|\left\{(x_i^\sigma, y_i^\sigma) \in s_m^\sigma \;/\; \max_k \left|\Delta\bar{h}_k^\gamma(x_i^\sigma) - \gamma y_{ik}^\sigma\right| \geq \frac{\gamma}{2}\right\}\right| \leq R^\gamma_{s^\sigma_m}(h)$$

By substitution in (7), one thus obtains:

$$P_{s_m, \tilde{s}_m}\left(\sup_{h\in\mathcal{H}}\left(\mathcal{R}^\gamma_{\tilde{s}_m}(h) - R^\gamma_{s_m}(h)\right) \geq \epsilon - \frac{1}{m}\right) \leq$$

$$\sup_{s_m, \tilde{s}_m} \mathcal{U}\left(\left\{\sigma \in S_t \;/\; \sup_{\Delta\bar{h}^\gamma \in \Delta\bar{\mathcal{H}}^\gamma}\left(\frac{1}{m}\left|\left\{(\tilde{x}_i^\sigma, \tilde{y}_i^\sigma) \in \tilde{s}_m^\sigma \;/\; \max_k \left|\Delta\bar{h}_k^\gamma(\tilde{x}_i^\sigma) - \gamma\tilde{y}_{ik}^\sigma\right| \geq \frac{\gamma}{2}\right\}\right| - \right.\right.\right.$$
$$\left.\left.\left.\frac{1}{m}\left|\left\{(x_i^\sigma, y_i^\sigma) \in s_m^\sigma \;/\; \max_k \left|\Delta\bar{h}_k^\gamma(x_i^\sigma) - \gamma y_{ik}^\sigma\right| \geq \frac{\gamma}{2}\right\}\right|\right) \geq \epsilon - \frac{1}{m}\right\}\right)$$
$$\leq \mathcal{N}_{\infty,\infty}(\gamma/2, \Delta\mathcal{H}^\gamma, 2m)$$
$$\sup_{s_m, \tilde{s}_m} \sup_{\Delta\bar{h}^\gamma \in \Delta\bar{\mathcal{H}}^\gamma} \mathcal{U}\left(\left\{\sigma \in S_t \;/\; \left(\frac{1}{m}\left|\left\{(\tilde{x}_i^\sigma, \tilde{y}_i^\sigma) \in \tilde{s}_m^\sigma \;/\; \max_k \left|\Delta\bar{h}_k^\gamma(\tilde{x}_i^\sigma) - \gamma\tilde{y}_{ik}^\sigma\right| \geq \frac{\gamma}{2}\right\}\right| - \right.\right.\right.$$
$$\left.\left.\left.\frac{1}{m}\left|\left\{(x_i^\sigma, y_i^\sigma) \in s_m^\sigma \;/\; \max_k \left|\Delta\bar{h}_k^\gamma(x_i^\sigma) - \gamma y_{ik}^\sigma\right| \geq \frac{\gamma}{2}\right\}\right|\right) \geq \epsilon - \frac{1}{m}\right\}\right)$$
$$\leq \mathcal{N}_{\infty,\infty}(\gamma/2, \Delta\mathcal{H}^\gamma, 2m) \sup_{\{a_i, b_i\}} Pr\left(\frac{1}{m}\sum_{i=1}^m (a_i - b_i)\beta_i \geq \epsilon - \frac{1}{m}\right) \tag{10}$$

where a_i is equal to 1 if $\max_k \left|\Delta\bar{h}_k^\gamma(\tilde{x}_i^\sigma) - \gamma\tilde{y}_{ik}^\sigma\right| \geq \frac{\gamma}{2}$ and is equal to 0 otherwise. Similarly, b_i is equal to 1 if $\max_k \left|\Delta\bar{h}_k^\gamma(x_i^\sigma) - \gamma y_{ik}^\sigma\right| \geq \frac{\gamma}{2}$ and is equal to 0 otherwise. The probability is over the β_i chosen independently and uniformly on $\{-1, 1\}$.

Exponential bound

The right-hand side of (10) is bounded by Hoeffding's inequality (see for example [25, 35]).

Theorem 3 (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 $\eta > 0$,*

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

Applying this bound in (10) and substituting the result to the right-hand side of (6) leads to

$$P_{s_m}\left(\sup_{h\in\mathcal{H}}\left(R(h) - R^{\gamma}_{s_m}(h)\right) \geq \epsilon\right) \leq$$

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

Setting the right-hand side to δ and solving for ϵ finally gives:

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

$$R(h) \leq R^{\gamma}_{s_m}(h) + \sqrt{\frac{1}{2m}\left(\ln(2\mathcal{N}_{\infty,\infty}(\gamma/2, \Delta\mathcal{H}^{\gamma}, 2m)) - \ln(\delta)\right)} + \frac{1}{m} \tag{11}$$

Making use of the proposition above requires to specify the quantity γ in advance. As pointed out by Bartlett in [4], this seems unnatural, since γ will be observed after the examples are seen. This difficulty can be overcome thanks to the following proposition, proved in [4], which allows us to give a result that stands uniformly for all values of the margin γ between 0 and 1.

Proposition 2 (Bartlett, Proposition 8 in [4]) *Let $(\Omega, \mathcal{B}, P)$ 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 $0 < \delta \leq 1$, $P(E(\alpha, \alpha, \delta)) \leq \delta$;*
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 $0 < a, \delta < 1$

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

Let $\alpha = \gamma$ and $a = 1/2$. Define $E(\gamma_1, \gamma_2, \delta)$ as the set of all the sequences $s_m \in S^m$ for which there is some h in $\mathcal{H}$ satisfying:

$$R(h) > R^{\gamma}_{s_m}(h) + \sqrt{\frac{1}{2m}\left(\ln(2\mathcal{N}_{\infty,\infty}(\gamma/2, \Delta\mathcal{H}^{\gamma}, 2m)) - \ln(\delta)\right)} + \frac{1}{m}$$

Using this parameterization in Proposition 2, Theorem 2 then directly springs from (11). ∎

4 Covering Numbers and Extended Fat-shattering/Graph Dimension

In this section, the covering numbers of interest are bounded using the strategy advocated in [4]. To that end, extensions of several lemmas in [2] to the case of vector-valued functions are derived. These extensions rest on an original capacity measure, the M-fat-shattering dimension. This measure is concomitantly an extension of the fat-shattering dimension to the multivariate case and a scale-sensitive variant of the graph dimension.

4.1 Definitions

In order to establish the lemmas, and the resulting bound, we must first introduce some definitions.

Definition 10 (Growth function [50]) *Let $\mathcal{F}$ be a set of indicator (binary-valued) functions on a set $\mathcal{X}$. Let $\Pi_{\mathcal{F}}$ be the function which maps any set s of points in $\mathcal{X}$ to the number of dichotomies $\Pi_{\mathcal{F}}(s)$ computed on it by the functions in $\mathcal{F}$. Then, the* growth function *of $\mathcal{F}$ is the function from the nonnegative integers to the nonnegative integers given by:*

$$\Pi_{\mathcal{F}}(m) = \max_{s_m \in \mathcal{X}^m} \Pi_{\mathcal{F}}(s_m)$$

Definition 11 (Vapnik-Chervonenkis (VC) dimension [50]) *Let $\mathcal{F}$ be a set of indicator functions on a set $\mathcal{X}$. A subset s_m of $\mathcal{X}^m$ is said to be* shattered *by $\mathcal{F}$ if $\Pi_{\mathcal{F}}(s_m) = 2^m$, i.e. if each dichotomy on s_m is computed by a function of $\mathcal{F}$. The* VC dimension *of $\mathcal{F}$, VC-dim($\mathcal{F}$), is the largest value of m such that $\Pi_{\mathcal{F}}(m) = 2^m$, if this value is finite, of infinity otherwise. If the VC dimension is finite, it is thus the size of the largest set of points shattered by $\mathcal{F}$.*

Pollard's pseudo-dimension extends the notion of (VC) dimension to the case of sets of real-valued functions.

Definition 12 (Pollard's pseudo-dimension [36, 24]) *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*

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

The P-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 fat-shattering dimension of Kearns and Schapire is a scale-sensitive version of the pseudo-dimension.

Definition 13 (Fat-shattering dimension [27, 28]) *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* γ-shattered *by* $\mathcal{H}$ *if there is a vector* $v_b = [b_i] \in \mathbb{R}^m$ *such that, for each binary vector* $v_y = [y_i] \in \{-1, 1\}^m$, *there is a function* $h_y \in \mathcal{H}$ *satisfying*

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

The vector v_b *is then said to* witness *the* γ-*shattering of* s_m *by* $\mathcal{H}$. *The* fat-shattering dimension *of the set* $\mathcal{H}$, $fat_{\mathcal{H}}$, *is a function from the positive real numbers to the integers which maps a value* γ *to the size of the largest set* γ-*shattered by functions of* $\mathcal{H}$, *if this size is finite, or to infinity otherwise.*

We propose to extend this definition to the case of vector-valued functions in the following manner.

Definition 14 (M-fat-shattering dimension) *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* M-γ-shattered *by* $\mathcal{H}$ *if there is a vector* $v_b = [b_i] \in \mathbb{R}^m$ *and a vector* $v_c = [c_i] \in \{1, \ldots, Q\}^m$ *such that, for each binary vector* $v_y = [y_i] \in \{-1, 1\}^m$, *there is a function* $h_y = [h_{yk}]$, $(1 \leq k \leq Q) \in \mathcal{H}$ *satisfying*

$$(h_{yc_i}(x_i) - b_i)\, y_i \geq \gamma,\ (1 \leq i \leq m)$$

The couple (v_b, v_c) *is then said to* witness *the* M-γ-*shattering of* s_m *by* $\mathcal{H}$. *The* M-fat-shattering dimension *of the set* $\mathcal{H}$, M-$fat_{\mathcal{H}}$, *is a function from the positive real numbers to the integers which maps a value* γ *to the size of the largest set* M-γ-*shattered by functions of* $\mathcal{H}$, *if this size is finite, or to infinity otherwise.*

As stated in introduction, this measure can be seen alternatively as a straightforward scale-sensitive extension of the graph dimension, introduced independently in [16, 33] (see also [42]).

Definition 15 (Graph dimension) *Let* $\mathcal{H}$ *be a set of functions on a set* $\mathcal{X}$ *taking their values in a countable set. For any* $h \in \mathcal{H}$, *the* graph $\mathcal{G}$ *of* h *is* $\mathcal{G}(h) = \{(x, h(x)) \ / \ x \in \mathcal{X}\}$ *and the* graph space *of* $\mathcal{H}$ *is* $\mathcal{G}(\mathcal{H}) = \{\mathcal{G}(h) \ / \ h \in \mathcal{H}\}$. *Then the* graph dimension *of* $\mathcal{H}$, $\mathcal{G}$-*dim*$(\mathcal{H})$, *is defined to be the VC dimension of the space* $\mathcal{G}(\mathcal{H})$.

In the context of our study, the most natural way to handle this dimension consists in making use of the general scheme developed in [8], which rests on the notion of Ψ-dimension.

Definition 16 (Ψ-shattering) *Let $\mathcal{F}$ be a set of functions on a set $\mathcal{X}$ taking their values in the finite set $\{1, \ldots, Q\}$. Let Ψ be a family of mappings $\psi = [\psi_i]$, such that each ψ_i is a mapping 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 ψ in Ψ such that for each vector v_y of $\{-1, 1\}^m$, there is a function f in $\mathcal{F}$ satisfying $[\psi_1 \circ f(x_1), \ldots, \psi_i \circ f(x_i), \ldots, \psi_m \circ f(x_m)] = v_y$.*

Definition 17 (Ψ-dimension) *Let $\mathcal{F}$ and Ψ be defined as above. The* Ψ-dimension *of $\mathcal{F}$, denoted by Ψ-dim$(\mathcal{F})$, is the maximal cardinality of a subset of $\mathcal{X}$ Ψ-shattered by $\mathcal{F}$, if it is finite, or infinity otherwise.*

When the functions in $\mathcal{F}$ have a finite range, the graph dimension appears as a particular case of Ψ-dimension, as can be seen with the following alternative definition.

Definition 18 (Graph dimension) *Let $\mathcal{F}$ be defined as above. The* graph dimension *of $\mathcal{F}$ is the Ψ-dimension of $\mathcal{F}$ in the restricted case where each mapping ψ_i is required to take the value 1 for one and only one value in $\{1, \ldots, Q\}$, and be equal to -1 otherwise.*

To understand the way the M-fat-shattering dimension can be seen as a scale-sensitive extension of the graph dimension, suffice it to notice two things. First, the functions f involved in the definition of the graph dimension can be seen as the discriminant functions associated with the multivariate functions h, by application of the "max" rule defined in the beginning of Section 2. Second, the choice of the components of the vector v_c plays in the case of the M-fat-shattering dimension the role played by the choice of the mappings ψ_i in the case of the graph dimension. To sum up, the M-fat-shattering dimension is related to the fat-shattering dimension through the parameters γ and v_b, which deal with the margin, whereas it is related to the graph dimension through the vector v_c which privileges, for each of the points considered, a specific category.

Definition 19 (M-strong dimension) *Let $\mathcal{X}$ be any set and let $\mathcal{S}$ be equal to $\{-n, \ldots, n\}^Q$. Let $\mathcal{F}$ be a set of functions on $\mathcal{X}$ taking their values in $\mathcal{S}$. $\mathcal{F}$* M-strongly shatters *a subset $s_m = \{x_i\}$, $(1 \leq i \leq m)$ of $\mathcal{X}$ if there exists a vector $v_b \in \{-n, \ldots, n\}^m$ and a vector $v_c = [c_i] \in \{1, \ldots, Q\}^m$ such that (v_b, v_c) witnesses the M-1-shattering of s_m by $\mathcal{F}$. The M-*strong dimension *of $\mathcal{F}$, M-strong$(\mathcal{F})$, is the maximal cardinality of a subset of $\mathcal{X}$ M-strongly shattered by $\mathcal{F}$, if it is finite, or infinity otherwise.*

Definition 20 (Packing numbers) *Let (E, ρ) be a pseudo-metric space. A set $H \subset E$ is* ϵ-separated *if, for any distinct points v_1 and v_2 in H, $\rho(v_1, v_2) \geq \epsilon$. The* ϵ-packing number *of H, $\mathcal{M}(\epsilon, H, \rho)$, is the maximal size of an ϵ-separated subset of H.*

As in the case of the covering numbers, the functional pseudo-metric which will be used here is $d_{l_\infty, l_\infty(s)}$. Two functions $f^{(1)}$ and $f^{(2)}$ from $\mathcal{X}$ into $\{-n, \ldots, n\}^Q$ are *M-separated* if they are M-2-separated with respect to the metric of interest, i.e. if there exists some $x \in \mathcal{X}$ and $k \in \{1, \ldots, Q\}$ such that $|f_k^{(1)}(x) - f_k^{(2)}(x)| \geq 2$.

Definition 21 (Pairwise M-separated set of functions) *Let $\mathcal{X}$ be any set and let $\mathcal{S} = \{-n, \ldots, n\}^Q$. A set $\mathcal{F}$ of functions from $\mathcal{X}$ into $\mathcal{S}$ is* pairwise M-separated *if any two distinct functions of $\mathcal{F}$ are M-separated.*

Definition 22 (η-discretization) *Let $h = [h_k]$ be a function from $\mathcal{X}$ into $\mathbb{R}^Q$ and $\eta > 0$. The* η-discretization *of h, denoted by $h^{(\eta)} = \left[h_k^{(\eta)}\right]$, is the function from $\mathcal{X}$ into $\mathbb{Z}^Q$ such that*

$$\forall k \in \{1, \ldots, Q\},\ h_k^{(\eta)}(x) = \lfloor \frac{h_k(x)}{\eta} \rfloor$$

or equivalently, $h_k^{(\eta)}(x) = \max\{j \in \mathbb{Z} \ / \ j\eta \leq h_k(x)\}$. 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 [2], 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 Δh^γ which take their values in $[-\gamma, \gamma]^Q$.

4.2 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 [29, 3], will prove useful in what follows.

Lemma 2 *For every pseudo-metric space (E, ρ), 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 3.1 in [2]:

Lemma 3 *For any class $\mathcal{H}$ of functions on $\mathcal{X}$ taking their values in $[-1, 1]^Q$ and for any $\eta > 0$:*

(1) for every $0 < \gamma \leq \eta/2$, M-strong$\left(\mathcal{H}^{(\eta)}\right) \leq M$-fat$_{\mathcal{H}}(\gamma)$;

(2) for every $\epsilon \geq 2\eta$ and every $s_m \in \mathcal{X}^m$,

$$\mathcal{M}(\epsilon, \mathcal{H}, d_{l_\infty, l_\infty(s_m)}) \leq \mathcal{M}(2, \mathcal{H}^{(\eta)}, d_{l_\infty, l_\infty(s_m)})$$

Proof To prove (1), it is enough to establish that any set M-strongly shattered by $\mathcal{H}^{(\eta)}$ is also M-$\eta/2$-shattered by $\mathcal{H}$. If s_m, a subset of $\mathcal{X}$ of cardinality m, is M-strongly shattered by $\mathcal{H}^{(\eta)}$, then according to Definition 19, there exists a couple of vectors $v_b \in$

$\{\lfloor -1/\eta \rfloor, \ldots, \lfloor 1/\eta \rfloor\}^m$ and $v_c \in \{1, \ldots, Q\}^m$ such that for every vector $v_y = [y_i] \in \{-1, 1\}^m$, there is a function $h_y^{(\eta)} = \left[h_{yk}^{(\eta)}\right] \in \mathcal{H}^{(\eta)}$, i.e. a function $h_y = [h_{yk}] \in \mathcal{H}$, satisfying for all x_i in s_m:

$$y_i \left(h_{yc_i}^{(\eta)}(x_i) - b_i\right) \geq 1$$

If $y_i = 1$ then

$$\left(h_{yc_i}^{(\eta)}(x_i) - b_i\right) \geq 1 \Longrightarrow \eta \left(h_{yc_i}^{(\eta)}(x_i) - b_i\right) \geq \eta$$

Since by definition of the η-discretization, $\eta h_{yc_i}^{(\eta)}(x_i) \leq h_{yc_i}(x_i)$, this implies

$$h_{yc_i}(x_i) - \eta b_i \geq \eta$$

or equivalently:

$$y_i \left(h_{yc_i}(x_i) - \eta b_i - \eta/2\right) \geq \eta/2$$

If $y_i = -1$ then

$$h_{yc_i}^{(\eta)}(x_i) - b_i \leq -1 \Longrightarrow \eta \left(h_{yc_i}^{(\eta)}(x_i) + 1\right) - \eta b_i \leq 0$$

Since by definition of the η-discretization, $\eta \left(h_{yc_i}^{(\eta)}(x_i) + 1\right) \geq h_{yc_i}(x_i)$, this implies that:

$$h_{yc_i}(x_i) - \eta b_i \leq 0$$

or equivalently:

$$y_i \left(h_{yc_i}(x_i) - \eta b_i - \eta/2\right) \geq \eta/2$$

Let $\tilde{v}_b = \left[\tilde{b}_i\right] \in \mathbb{R}^m$ be the vector deduced from v_b as follows:

$$\forall i \in \{1, \ldots, m\}\,,\ \tilde{b}_i = \eta b_i + \eta/2$$

$\tilde{v}_b$ has been constructed so that the couple $(\tilde{v}_b, v_c)$ witnesses the M-$\eta/2$-shattering of s_m by $\mathcal{H}$ (or more precisely the set of functions h_y). As a consequence, any set M-strongly shattered by $\mathcal{H}^{(\eta)}$ is also M-$\eta/2$-shattered by $\mathcal{H}$, which is precisely our claim.

To prove (2), let us first notice that:

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

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

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

$$h_k^{(2)(\eta)}(x)\eta \leq h_k^{(2)}(x) < h_k^{(1)}(x) < (h_k^{(1)(\eta)}(x) + 1)\eta$$

Thus,

$$(h_k^{(1)(\eta)}(x)+1)\eta - h_k^{(2)(\eta)}(x)\eta > 2\eta$$

and finally

$$h_k^{(1)(\eta)}(x) - h_k^{(2)(\eta)}(x) > 1$$

from which the desired result springs directly, keeping in mind that the η-discretizations are integers.

Let $s_{\mathcal{H}}$ be a M-2η-separated substet of $\mathcal{H}$ with respect to the pseudo-metric $d_{l_\infty,l_\infty(\mathcal{X})}$. It results from Definition 8 that:

$$\forall (h^{(1)},h^{(2)}) \in s_{\mathcal{H}}^2,\ d_{l_\infty,l_\infty(\mathcal{X})}(h^{(1)},h^{(2)}) \geq 2\eta \Longrightarrow$$

$$\max_{x\in\mathcal{X}}\max_k \left| h_k^{(1)}(x) - h_k^{(2)}(x)\right| \geq 2\eta$$

$$\Longrightarrow \max_{x\in\mathcal{X}}\max_k \left| h_k^{(1)(\eta)}(x) - h_k^{(2)(\eta)}(x)\right| \geq 2 \Longrightarrow d_{l_\infty,l_\infty(\mathcal{X})}(h^{(1)(\eta)},h^{(2)(\eta)}) \geq 2$$

We have thus proved (2). ∎

We now prove our main combinatorial result, an extension of Lemma 3.2 in [2], which gives a new generalization of Sauer's lemma [50, 39, 44].

Lemma 4 *Let $\mathcal{F}$ be a set of vector-valued functions from a finite domain $\mathcal{X}$ of cardinality $|\mathcal{X}|$ to a finite range $\mathcal{S} = \{-n,\ldots,n\}^Q$. Let M-strong$(\mathcal{F}) = d$, with $d < |\mathcal{X}|$, then*

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

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

Proof Let us say that a class $\mathcal{F}$ as above M-strongly shatters a triplet (s_m, v_b, v_c) (for a nonempty subset s_m of $\mathcal{X}$ of cardinality m, a vector $v_b \in \{-n,\ldots,n\}^m$ and a vector $v_c \in \{1,\ldots,Q\}^m$) if $\mathcal{F}$ M-strongly shatters s_m according to (v_b, v_c). 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 M-separated functions from $\mathcal{F}$, $\mathcal{F}_l$ M-strongly shatters at least t triplets (s, v_b, v_c). If no such $\mathcal{F}_l$ exists, then $t(l,|\mathcal{X}|)$ is infinite.

The number of triplets (s, v_b, v_c) that could be shattered and for which the cardinality of s does not exceed $d \geq 1$ is strictly less than $\sum_{i=1}^{d}\binom{|\mathcal{X}|}{i}(Q(2n+1))^i$, since for s of size $i > 0$, there are strictly less than $(Q(2n+1))^i$ possibilities to choose the couple (v_b, v_c). It follows that, given a set of functions $\mathcal{F}$ from $\mathcal{X}$ into $\mathcal{S}$, $t(l,|\mathcal{X}|) \geq \phi(d,|\mathcal{X}|)$ for some l and M-strong$(\mathcal{F}) \leq d$ 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 M-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 $|\mathcal{X}| \geq 1$,

$$t(2\left(|\mathcal{X}|Q(2n+1)^Q\right)^{\lceil \log_2(\phi(d,|\mathcal{X}|)) \rceil}, |\mathcal{X}|) \geq \phi(d, |\mathcal{X}|) \tag{12}$$

We claim that

$$t(2, |\mathcal{X}|) = 1 \tag{13}$$

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

$$t(2m|\mathcal{X}|Q(2n+1)^Q, |\mathcal{X}|) \geq 2t(2m, |\mathcal{X}| - 1) \tag{14}$$

for all $m \geq 1$ and $|\mathcal{X}| \geq 2$. The first part of the claim is obvious. For the second part, first note that if no set of $2m|\mathcal{X}|Q(2n+1)^Q$ pairwise M-separated functions from $\mathcal{X}$ to $\mathcal{S}$ exists, then by definition $t(2m|\mathcal{X}|Q(2n+1)^Q, |\mathcal{X}|) = \infty$ and hence the claim holds. Assume then that there is a set $\mathcal{F}_0$ of $2m|\mathcal{X}|Q(2n+1)^Q$ pairwise M-separated functions from $\mathcal{X}$ to $\mathcal{S}$. Split it arbitrarily into $m|\mathcal{X}|Q(2n+1)^Q$ pairs. For each pair $\left(f^{(1)}, f^{(2)}\right)$, find a point $x \in \mathcal{X}$ and a component $k \in \{1, \ldots, Q\}$ such that $\left|f_k^{(1)}(x) - f_k^{(2)}(x)\right| \geq 2$. We know that such a couple always exists since by hypothesis $\mathcal{F}_0$ is pairwise M-separated. By the pigeonhole principle, each procedure of this type will result in (at least) one particular couple (x_0, k_0) being selected at least $(m|\mathcal{X}|Q(2n+1)^Q)/(|\mathcal{X}|Q) = m(2n+1)^Q$ times. According to the same principle, there is at least one couple $(i, j) \in \{-n, \ldots, n\}^2$, satisfying $|j - i| \geq 2$ which will be associated with this couple (x_0, k_0) by at least $m(2n+1)^Q / \binom{2n+1}{2} > 2m$ of the corresponding pairs of functions. This means that there are two sub-classes of $\mathcal{F}_0$ of cardinality at least $2m$, call them $\mathcal{F}_1$ and $\mathcal{F}_2$, and there are $x_0 \in \mathcal{X}$, $k_0 \in \{1, \ldots, Q\}$ and a couple (i, j) in $\{-n, \ldots, n\}^2$ with $|j - i| \geq 2$, such that for each $f^{(1)} \in \mathcal{F}_1$, $f_{k_0}^{(1)}(x_0) = i$ and for each $f^{(2)} \in \mathcal{F}_2$, $f_{k_0}^{(2)}(x_0) = j$. Obviously, the members of $\mathcal{F}_1$ are pairwise M-separated on $\mathcal{X} \setminus \{x_0\}$ and the same holds for the members of $\mathcal{F}_2$. Hence, by the definition of function t, $\mathcal{F}_1$ M-strongly shatters at least $t(2m, |\mathcal{X}| - 1)$ triplets (s, v_b, v_c) with $s \subseteq \mathcal{X} \setminus \{x_0\}$, and the same holds for $\mathcal{F}_2$. Clearly, $\mathcal{F}_0$ M-strongly shatters all triplets M-strongly shattered by $\mathcal{F}_1$ or $\mathcal{F}_2$. Moreover, if the same triplet (s, v_b, v_c) is M-strongly shattered both by $\mathcal{F}_1$ and by $\mathcal{F}_2$, then $\mathcal{F}_0$ also M-strongly shatters the triplet $(s \bigcup \{x_0\}, \bar{v}_b, \bar{v}_c)$, where $\bar{v}_b$ is deduced from v_b by adding a component corresponding to the point x_0, component taking the value $\left\lfloor \frac{i+j}{2} \right\rfloor$ (which belongs to $\{-n, \ldots, n\}$), and $\bar{v}_c$ is deduced from v_c by adding a component also corresponding to the point x_0 and taking the value k_0. Since, by construction, neither $\mathcal{F}_1$ nor $\mathcal{F}_2$ M-strongly shatters $s \bigcup \{x_0\}$ (whatever the couple (v_b, v_c) may be), it follows that $t(2m|\mathcal{X}|Q(2n+1)^Q, |\mathcal{X}|) \geq 2t(2m, |\mathcal{X}| - 1)$, which is precisely (14).

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

$$l = 2(Q(2n+1)^Q)^r \, \Pi_{u=0}^{r-1}(|\mathcal{X}| - u)$$

Applying (14) iteratively and eventually (13), it appears that $t(l, |\mathcal{X}|) \geq 2^r$. Since t is clearly monotone in its first argument, and $2(|\mathcal{X}|Q(2n+1)^Q)^r \geq l$, this implies

$$t(2(|\mathcal{X}|Q(2n+1)^Q)^r, |\mathcal{X}|) \geq 2^r$$

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

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

and consequently

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

which is precisely (12). If on the contrary $|\mathcal{X}| \leq \lceil \log_2(\phi(d,|\mathcal{X}|)) \rceil$, then

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

Since the total number of functions from $\mathcal{X}$ into $\mathcal{S}$ is precisely $(2n+1)^{Q|\mathcal{X}|}$, there is no set of pairwise M-separated functions of cardinality larger that this included in $\mathcal{F}$ and hence, by definition of t,

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

$t(2(|\mathcal{X}|Q(2n+1)^Q)^{\lceil \log_2(\phi(d,|\mathcal{X}|)) \rceil}, |\mathcal{X}|)$ is consequently once more superior to $\phi(d,|\mathcal{X}|)$, which completes the proof of (12) and thus concludes the proof of the lemma. ∎

4.3 Application to the covering number of $\Delta\mathcal{H}^\gamma$

Let

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

and

$$\mathcal{M}_{\infty,\infty}(2, (\Delta\mathcal{H}^\gamma)^{(\gamma/4)}, 2m) = \max_{s_{2m} \in \mathcal{X}^{2m}} \mathcal{M}(2, (\Delta\mathcal{H}^\gamma)^{(\gamma/4)}, d_{l_\infty, l_\infty(s_{2m})})$$

where $(\Delta\mathcal{H}^\gamma)^{(\gamma/4)}$ designates the set of the functions $(\Delta h^\gamma)^{(\gamma/4)}$, $\gamma/4$-discretizations of the functions $\Delta h^\gamma = [\pi_\gamma \circ \Delta h_k]$ of $\Delta\mathcal{H}^\gamma$. Applying Lemma 2 to $\Delta\mathcal{H}^\gamma$ gives:

$$\mathcal{N}_{\infty,\infty}(\gamma/2, \Delta\mathcal{H}^\gamma, 2m) \leq \mathcal{M}_{\infty,\infty}(\gamma/2, \Delta\mathcal{H}^\gamma, 2m)$$

Setting $\epsilon = \gamma/2$ ($\eta = \gamma/4$) in Proposition (2) of Lemma 3, one establishes that:

$$\mathcal{M}_{\infty,\infty}(\gamma/2, \Delta\mathcal{H}^\gamma, 2m) \leq \mathcal{M}_{\infty,\infty}(2, (\Delta\mathcal{H}^\gamma)^{(\gamma/4)}, 2m)$$

Similarly, the packing numbers of the discretized set of functions can be bounded thanks to Lemma 4, by setting $\mathcal{F} = (\Delta\mathcal{H}^\gamma)^{(\gamma/4)}$ and $|\mathcal{X}| = 2m$. To make use of this lemma, the nature of the range $\mathcal{S}$, and more precisely the value of parameter n, must first be established. By definition of π_γ, each component Δh_k^γ of a function Δh^γ of $\Delta\mathcal{H}^\gamma$ takes its values in $[-\gamma, \gamma]$.

As a consequence, its $\gamma/4$-discretization takes its values in $\{-4, \ldots, 4\}$, i.e. in a set of cardinality 9. Thus $n = 4$, $2n + 1 = 9$ and we get:

$$\mathcal{M}_{\infty,\infty}(2, (\Delta\mathcal{H}^{\gamma})^{\gamma/4}, 2m) \leq 2\left(2mQ9^{Q}\right)^{\lceil \log_2(\phi(d,2m)) \rceil} \tag{15}$$

In the right-hand side of (15), $\phi(d, 2m) = \sum_{i=1}^{d} \binom{2m}{i} (9Q)^i$, where d is the M-strong dimension of $(\Delta\mathcal{H}^{\gamma})^{(\gamma/4)}$. Since we are interested in upperbounding the capacity measure, d can also be set equal to M-fat$_{\Delta\mathcal{H}^{\gamma}}(\gamma/8)$, due to Proposition (1) in Lemma 3. To find an upper bound of $\phi(d, |\mathcal{X}|)$, we take our inspiration from Vapnik's version of "Sauer's lemma". For all couples $(d, |\mathcal{X}|)$ of integers satisfying $1 \leq d < |\mathcal{X}|$, let

$$\Phi(d, |\mathcal{X}|) = \sum_{i=0}^{d} \binom{|\mathcal{X}|}{i} (9Q)^i$$

i.e. $\Phi(d, |\mathcal{X}|) = \phi(d, |\mathcal{X}|) + 1$ for $1 \leq d < |\mathcal{X}|$. Function Φ satisfies the following recurrence formula:

$$\forall d \geq 2,\ \forall |\mathcal{X}| > d,\ \Phi(d, |\mathcal{X}|) = \Phi(d, |\mathcal{X}| - 1) + 9Q\Phi(d - 1, |\mathcal{X}| - 1) \tag{16}$$

We will now prove the following lemma, which extends Lemma 4.5 in [48] (see also the appendix of Chapter 6 in [46]).

Lemma 5 *For all couple $(d, |\mathcal{X}|)$ of positive integers such that $d < |\mathcal{X}|$, the following bound is true:*

$$\Phi(d, |\mathcal{X}|) < \left(\frac{9eQ|\mathcal{X}|}{d}\right)^{d} \tag{17}$$

Proof First, note that $1.5\,(9Q|\mathcal{X}|)^d/d!$ is a lower bound of $(9eQ|\mathcal{X}|/d)^d$. 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 (16), proving (17) can also be performed by recurrence. This amounts to proving three separate results: that (17) stands for $d = 1$, irrespective of the value of

$|\mathcal{X}| > d$, 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,

$$\Phi(1, |\mathcal{X}|) = 1 + 9Q|\mathcal{X}| < 1.5\,(9Q|\mathcal{X}|)$$

As for the "general case", making use of (16) gives:

$$\left(\Phi(d, |\mathcal{X}|) < 1.5\frac{(9Q|\mathcal{X}|)^d}{d!}\right) \wedge \left(\Phi(d+1, |\mathcal{X}|) < 1.5\frac{(9Q|\mathcal{X}|)^{d+1}}{(d+1)!}\right) \Longrightarrow$$

$$9Q\Phi(d, |\mathcal{X}|) + \Phi(d+1, |\mathcal{X}|) < 9Q\,(d+1+|\mathcal{X}|)\,1.5\frac{(9Q|\mathcal{X}|)^d}{(d+1)!} \Longrightarrow$$

$$\Phi(d+1, |\mathcal{X}|+1) < 9Q\,(d+1+|\mathcal{X}|)\,1.5\frac{(9Q|\mathcal{X}|)^d}{(d+1)!} \tag{18}$$

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 (18), this leads to

$$\Phi(d+1, |\mathcal{X}|+1) < 1.5(9Q)^{d+1}\frac{(|\mathcal{X}|+1)^{d+1}}{(d+1)!}$$

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

$$\Phi(d, d+1) = (9Q+1)^{d+1} - (9Q)^{d+1} < (d+1)(9Q+1)^d$$

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

$$2(9eQ)^d \leq \left(\frac{9eQ(d+1)}{d}\right)^d$$

It is obvious that the relation

$$(d+1)(9Q+1)^d < 2(9eQ)^d$$

stands for all the couples (d, Q) satisfying $d >= 1$, $Q >= 3$. By transitivity, we thus get

$$\Phi(d, d+1) < \left(\frac{9eQ(d+1)}{d}\right)^d$$

which is exactly (17) 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. In the case of interest, $|\mathcal{X}| = 2m$, this implies that

$$\Phi(d, 2m) < (18emQ/d)^d$$

and consequently

$$\log_2(\phi(d, 2m)) < \log_2(\Phi(d, 2m)) < d\log_2(18emQ/d)$$

The following theorem is a direct consequence of this last bound and the three lemmas of Subsection 4.2:

Theorem 4 *Let $\mathcal{H}$ be a set of functions from $\mathcal{X}$ into $\mathbb{R}^Q$. For every value of γ in $(0,1]$ and every value of m satisfying $M\text{-}fat_{\Delta\mathcal{H}^\gamma}(\gamma/8) < 2m$, the following bound is true:*

$$\mathcal{N}_{\infty,\infty}(\gamma/2, \Delta\mathcal{H}^\gamma, 2m) \leq 2\left(2mQ9^Q\right)^{d\log_2(18emQ/d)} \tag{19}$$

where $d = M\text{-}fat_{\Delta\mathcal{H}^\gamma}(\gamma/8)$.

To sum up, thanks to the four lemmas of this section, it has been possible to derive Theorem 4, which states that the problem of bounding the covering numbers of interest can be reduced to the problem of bounding the M-fat-shattering dimension of $\Delta\mathcal{H}^\gamma$. Note that we have systematically used the hypothesis that twice the size of the sample available was superior to the extended VC dimension considered, although it was by no means compulsory. This was done in order to highlight the fact that if this hypothesis is not satisfied, then different (simpler) results can be derived, which will give birth to tighter bounds.

5 Bounds on Error Expectation for M-SVMs

5.1 Motivations and hypotheses

In this section, the bound on the covering numbers derived in the former section is applied to the case of kernel machines taking their values in $\mathbb{R}^Q$. The goal is primarily to study existing training algorithms and make it possible to specify new ones in the framework of the (*data-dependent*) SRM inductive principle. In view of Theorem 4, this can be done by studying the behaviour of the corresponding M-fat-shattering dimension as a function of the constraints on the parameters. With the aforementioned aim in mind, we do not attempt to establish the tightest possible bound, or even to present a single master theorem, but rather

to sketch a simple pathway highlighting the incidence of these constraints on the capacity.

We make no specific hypothesis regarding the set $\mathcal{X}$ of covariates. On the contrary, the *feature space* $E_{\Phi(\mathcal{X})}$ is assumed to be a Hilbert space endowed with the Euclidean dot product. This standard hypothesis is a prerequisite to compute linear boundaries. $E_{\Phi(\mathcal{X})}$ can be infinite dimensional, so that no restriction is induced on the nature of the kernel used, which can for instance be Gaussian. Furthermore, $\Phi(\mathcal{X})$ is supposed to be bounded in $E_{\Phi(\mathcal{X})}$, which will be needed to bound the M-fat-shattering dimension.

5.2 Architecture and training of the M-SVMs

The problem of performing multi-class discriminant analysis with SVMs was initially tackled through decomposition schemes [40, 32, 48]. The first multi-class SVM published was proposed independently and under different forms by several teams (see for instance [53, 48, 11, 21]). Variants of this machine can be found in [19, 26]. Recently, two new models became available. The first one, described in [14] (see also [13]), uses an original expression of the empirical risk. The bound on the generalization error provided is directly borrowed from a tree-based decomposition approach called DAGSVM [34]. In [31], the machine is devised to asymptotically implement the Bayes rule. All these SVMs only differ in their training algorithm. They share the same architecture. Precisely, they are obtained by combining a multivariate affine model with the nonlinear mapping Φ into the feature space. Formally, the functions $h = [h_k]$ of the family $\mathcal{H}$ considered are defined by:

$$\forall k \in \{1, \ldots, Q\}\,,\ h_k(x) = w_k^T \Phi(x) + b_k \tag{20}$$

As usual, the mapping Φ does not appears explicitely in the computations. Thanks to the "kernel trick", it is replaced with the *reproducing kernel function* K, which computes the Euclidean dot product in the feature space, i.e.:

$$\forall (x^{(1)}, x^{(2)}) \in \mathcal{X}^2,\ K(x^{(1)}, x^{(2)}) = \langle \Phi(x^{(1)}), \Phi(x^{(2)}) \rangle \tag{21}$$

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 [38, 51, 52]). The kernel satisfies Mercer's conditions [1].

In its primal formulation, training thus consists in finding the values of the vectors w_k and the reals b_k. This amounts to solving a quadratic programming (QP) problem. However, for both theoretical and technical reasons, linked for instance to the use of the kernel trick, this problem is solved in its Wolfe dual form. The parameters to be optimized are then the coefficients $\beta_{i,k}$ appearing in the following expansions:

$$\forall k \in \{1, \ldots, Q\}\,,\ h_k(x) = \sum_{i=1}^{m} \beta_{i,k} K(x_i, x) + b_k \tag{22}$$

where the x_i, $(1 \leq i \leq m)$ are the covariates of the points in the training set.

5.3 Principle of the estimation of the M-fat-shattering dimension

In a nutshell, the strategy implemented is of the "divide and conquer" type. It consists in deriving progressively a bound on M-fat$_{\Delta\mathcal{H}^\gamma}(\epsilon)$ in terms of simpler dimensions of basic discriminant models, for which sharp bounds have already been found. The main idea is to observe that the "margin discriminant funtions" computed by the model of interest are also computed by the multiple-output multi-layer perceptron (MLP) with threshold units depicted in Figure 1.

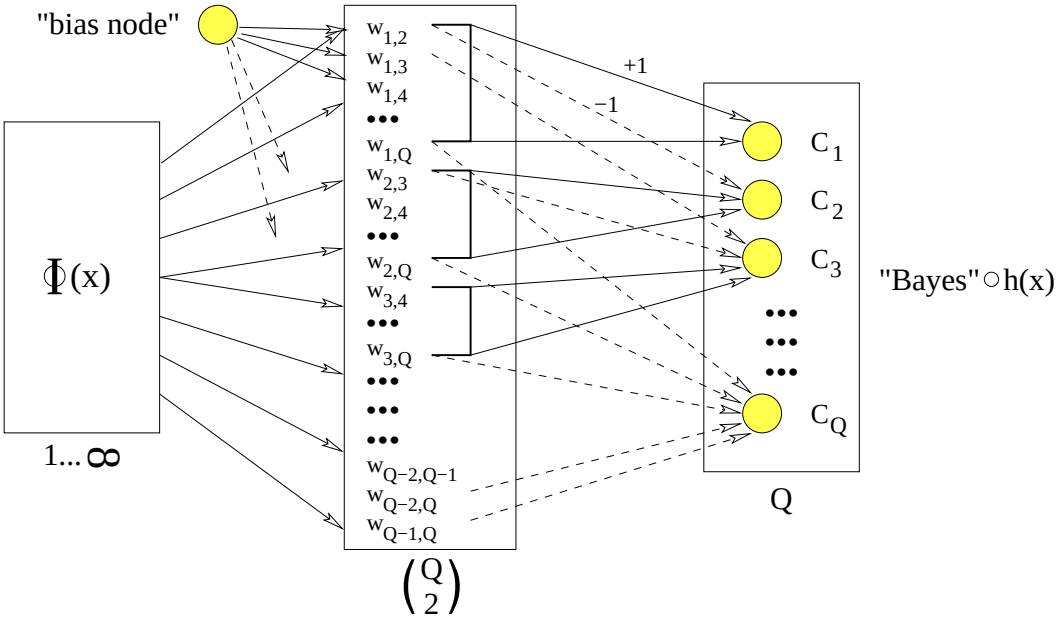

Figure 1: Architecture of a MLP with threshold units computing the ϵ-margin discriminant functions associated with the functions h of $\mathcal{H}$. Its graph dimension is superior to the M-fat-shattering dimension of $\Delta\mathcal{H}^\gamma$ under the constraint $v_b = 0$. Each hidden unit computes a function $\tilde{h}_{k,l}(\Phi(x)) = t_h\left(w_{k,l}^T\Phi(x) + b_{k,l}\right)$, $(1 \leq k < l \leq Q)$, where the threshold function t_h is defined as follows: $t_h(z) = 1$ if $z \geq \epsilon$, $t_h(z) = -1$ if $z \leq -\epsilon$ and $t_h(z) = 0$ otherwise. The real values $b_{k,l}$, $(1 \leq k < l \leq Q)$, correspond to the values coming from the bias node. The weights of the output layer, either $+1$ (solid lines) or -1 (dashed lines), and the biases are chosen so that the output units compute a logical AND ($t_o(z) = 1$ if $z = Q - 1$, $t_o(z) = -1$ otherwise).

The input of this two-layer MLP is vector $\Phi(x)$, which, as was stated before, can be infinite dimensional. The hidden layer is made up of $\binom{Q}{2}$ threshold units, one for each pair of categories. The unit associated with the categories C_k and C_l, with $k < l$, computes

a function $\tilde{h}_{k,l}$ such that $\tilde{h}_{k,l}(x) = t_h\left(h_{k,l}(x)\right)$, where $h_{k,l}(x) = w_{k,l}^T\Phi(x) + b_{k,l}$ and the threshold function t_h outputs 1 if its argument is superior or equal to ϵ, -1 if it is inferior or equal to $-\epsilon$, and 0 otherwise. This last value must not be seen as a third state, but rather as an absence of output. The output layer is made up of Q threshold units, one for each category. The weights of the upper layer of connections, either $+1$ (solid lines) or -1 (dashed lines), and the biases are chosen so that the output units compute a logical AND. Precisely, let $\tilde{h} = \left[\tilde{h}_k\right]$, $(1 \leq k \leq Q)$ be the function computed by the MLP when the parameters are set to the values $(w_{k,l}, b_{k,l})$, $(1 \leq k < l \leq Q)$. Then the value of the k-th output computed on x is given by:

$$\tilde{h}_k(x) = \begin{cases} 1 & if \begin{cases} w_{k,l}^T\Phi(x) + b_{k,l} \geq \epsilon, & (k < l \leq Q) \\ w_{l,k}^T\Phi(x) + b_{l,k} \leq -\epsilon, & (1 \leq l < k) \end{cases} \\ -1 & otherwise \end{cases}$$

This architecture is specifically conceived so that its graph dimension is directly related to M-fat$_{\Delta\mathcal{H}^\gamma}(\epsilon)$. In order to highlight this relationship, additional definitions must first be introduced.

Definition 23 (Vapnik dimension [47]) *Let $\mathcal{H}$ be a set of real-valued functions on a set $\mathcal{X}$. The Vapnik dimension of $\mathcal{H}$, V-dim$(\mathcal{H})$, is simply P-dim$(\mathcal{H})$ in the case where all the components of vector v_b are supposed to take the same value.*

Definition 24 (V_γ dimension [2, 22]) *Let $\mathcal{H}$ be a set of real-valued functions on a set $\mathcal{X}$. Its V_γ dimension, V_γ-dim$(\mathcal{H})$, is defined as the scale-sensitive version of its Vapnik dimension. It is thus its fat-shattering dimension in the case where all the components of vector v_b are supposed to take the same value.*

Several results are available which relate the V_γ dimension of a set of functions to its fat-shattering dimension. More precisely, each of these dimensions can be bounded in terms of the other one (see for instance Lemma 2.2 in [2] or Claim 1 in [22]). Building upon these results, we consider here the following variant of the M-fat-shattering dimension:

Definition 25 (Uniform M-fat-shattering dimension) *Let $\mathcal{H}$ be a set of functions on a set $\mathcal{X}$ taking their values in $\mathbb{R}^Q$. For $\gamma > 0$, the* uniform M-fat-shattering dimension *of $\mathcal{H}$, UM-fat$_{\mathcal{H}}$, is simply M-fat$_{\mathcal{H}}$ in the case where the components of vector v_b are constrained to take only Q different values, one for each category. In other words, if two components of the vector v_c are equal, then the correponding components of the vector v_b are also equal.*

In the same way as the V_γ dimension can be used to bound the fat-shattering dimension, the uniform M-fat-shattering dimension can be used to bound the M-fat-shattering dimension. To that end, one can for instance make use of the following lemma.

Lemma 6 *Let $\mathcal{H}$ be a class of functions on a set $\mathcal{X}$ taking their values in $[-1, 1]^Q$. For any positive real ϵ,*

$$M\text{-}fat_{\mathcal{H}}(\epsilon) \leq 2(1-\epsilon)/\epsilon \; UM\text{-}fat_{\mathcal{H}}(\epsilon/2)$$

Proof Let s_m be a subset of $\mathcal{X}$ of cardinality $m = \text{M-fat}_{\mathcal{H}}(\epsilon)$ M-ϵ-shattered by $\mathcal{H}$ and let (v_b, v_c) be a couple witnessing this shattering. k being any integer in the set $\{1, \ldots, Q\}$, let us consider the subset of s_m made up of the points associated with the components of v_c equal to k. Let m_k be the cardinality of this subset. For the sake of simplicity, and without loss of generality, we assume that the indexes of these points are the integers from 1 to m_k. From the definition of the M-ϵ-shattering we get that $-1 + \epsilon \leq b_i \leq 1 - \epsilon$ holds true for all values of i between 1 and m_k. The interval $[-1 + \epsilon, 1 - \epsilon]$ can be covered by $2(1 - \epsilon)/\epsilon$ consecutive intervals of lenght ϵ. According to the pigeonhole principle, at least one of them contains at least $\lceil \epsilon m_k / (2(1 - \epsilon)) \rceil$ points in the set $\{b_1, b_2, \ldots, b_{m_k}\}$. Let $\mathcal{I}(k)$ denote one interval satisfying this property and let $\tilde{b}_k$ be its center. For a given binary vector $v_y = [y_i] \in \{-1, 1\}^m$, let $h_y = [h_{yk}]$, $(1 \leq k \leq Q)$, be a function of $\mathcal{H}$ such that $(h_{yk}(x_i) - b_i)\, y_i \geq \epsilon$, $(1 \leq i \leq m)$. Then, the triangle inequality implies that any point x_j in s_m such that $c_j = k$ and $b_j \in \mathcal{I}(k)$ satisfies $\left(h_{yk}(x_j) - \tilde{b}_k\right) y_j \geq \epsilon/2$. By definition, the set of all such points is thus uniformly M-$\epsilon/2$-shattered by $\mathcal{H}$. The claim then directly springs from the fact that one can exhibit subsets of s_m of this type (and the corresponding components $\tilde{b}_k$) for each of the Q categories. The sum of their cardinalities constitutes a lower bound of $\text{UM-fat}_{\mathcal{H}}(\epsilon/2)$. We have thus: $\text{UM-fat}_{\mathcal{H}}(\epsilon/2) \geq \epsilon/(2(1 - \epsilon)) \sum_{k=1}^{Q} m_k = \epsilon/(2(1 - \epsilon))m$. ∎

This Lemma can be readily extended to sets of functions taking their values in different products of bounded intervals. In the case of interest, this means that for any choice of the parameters γ and ϵ such that $0 < \epsilon < \gamma$, there exists a constant $K_{\gamma,\epsilon}$, depending only on γ and ϵ, such that $\text{M-fat}_{\Delta\mathcal{H}^\gamma}(\epsilon) \leq K_{\gamma,\epsilon} \text{UM-fat}_{\Delta\mathcal{H}^\gamma}(\epsilon/2)$. As a consequence, bounding the first dimension boils down to bounding the second one.

The M-fat-shattering dimension can be computed for any kind of vector-valued model, not only discriminant ones. In the specific case of multi-class supervised learning, the quantity of interest is the difference between the scores associated with the different labels. The difference between the output corresponding to the category of the example and the second highest output must be as large as possible. There is *a priori* no need for the degree of freedom provided by the vector v_b appearing in the definition of the fat-shattering dimension and its variations. As a consequence, in what follows, the study deals with the uniform M-fat-shattering dimension of the M-SVMs computed under the additional constraint $v_b = 0$. The adaptation of the results derived so far to this specific situation rises no difficulty. For the sake of simplicity, we omit details here.

5.4 Uniform M-fat-shattering dimension of M-SVMs and graph dimension of the MLP

In what follows, $\mathcal{H} = [h]$ will designate the set a functions computed by a M-SVM, and $\tilde{\mathcal{H}} = \left[\tilde{h}\right]$ the set of functions computed by the MLP of Section 5.3. Furthermore, in this section and the next one only, no hypothesis will be made regarding constraints on the parameters of any of the two architectures. The following bound can be derived very simply:

$$\forall(\gamma,\epsilon) \ / \ 0<\epsilon<\gamma,\ \text{UM-fat}_{\Delta\mathcal{H}^\gamma}(\epsilon) \leq \text{UM-fat}_{\Delta\mathcal{H}}(\epsilon) \tag{23}$$

Given a set of functions h in $\mathcal{H}$ such that the set of functions Δh uniformly M-ϵ-shatters a subset s_m of $\mathcal{X}$, our goal is to exhibit a corresponding set of functions computed by the MLP and also shattering s_m, but according to the graph dimension this time.

Let $s_m = \{x_i\}$, $(1 \leq i \leq m)$ be a subset of $\mathcal{X}$ uniformly M-ϵ-shattered by $\Delta\mathcal{H}$ and let $v_c = [c_i] \in \{1,\ldots,Q\}^m$ be a vector witnessing this shattering for $v_b = 0$. Let v_y be any vector of $\{-1,1\}^m$ and let h_y be a function of $\mathcal{H}$ such that

$$\Delta h_{yc_i}(x_i) y_i \geq \epsilon,\ (1 \leq i \leq m)$$

Let (w_{yk}, b_{yk}), $(1 \leq k, \leq Q)$, be the set of parameters characterizing the hyperplanes of h_y. Consider now the following parameterization of the MLP:

$$w_{k,l} = 1/2\ (w_{yk} - w_{yl}),\ (1 \leq k < l \leq Q)$$

$$b_{k,l} = 1/2\ (b_{yk} - b_{yl}),\ (1 \leq k < l \leq Q)$$

i.e. $h_{k,l}(x) = 1/2\ \ (h_{yk}(x) - h_{yl}(x))$, $(1 \leq k < l \leq Q)$. If $y_i = 1$, then by definition of the delta function, we have:

$$\Delta h_{yc_i}(x_i) \geq \epsilon \Longrightarrow \min_{k \neq c_i} \{1/2\,(h_{yc_i}(x_i) - h_{yk}(x_i))\} \geq \epsilon \Longrightarrow \tilde{h}_l(x_i) = \left\{ \begin{array}{cc} 1 & if\ l = c_i \\ -1 & otherwise \end{array} \right.$$

In short, the MLP outputs the canonical coding of the category of x_i. A difficulty springs from the fact that on the contrary, if $y_i = -1$, $\Delta h_{yc_i}(x_i) \leq -\epsilon$ does not implies that the MLP will output the coding of a category different from $C(x_i) = C_{c_i}$. Indeed, the output can be simply -1_Q (no output of the M-SVM exceeds the others by a large enough margin). To overcome this difficulty, a simple possibility consists in duplicating the units of the output layer. Whereas the first unit associated with category C_k had to fire if and only if $\Delta h_{yk}(x_i) \geq \epsilon$, the second one will fire if and only if $\Delta h_{yk}(x_i) > -\epsilon$. This can be easily obtained by setting the values of the new weights of the upper layer adequately. Note that this addition has no incidence on the capacity (graph dimension) of the MLP, since the values of all the weights in the upper layer are set once for all. Then, in the case when the Q first output units do not select any category ($\tilde{h}(x_i) = -1_Q$), simply selecting among the units firing in the second set the one of smallest index provides us with a unique label which is different from $C(x_i)$, as required by the definition of the graph dimension. We have thus established that a subset of $\mathcal{X}$ uniformly M-ϵ-shattered by $\Delta\mathcal{H}$ is also shattered by $\tilde{\mathcal{H}}$. The following theorem, which makes it possible to reformulate once more the problem of bounding the covering numbers of interest, is a direct consequence of this result and (23).

Theorem 5 *Let $\mathcal{H}$ be the set of functions computed by a M-SVM and let $\tilde{\mathcal{H}}$ be the set of functions computed by the MLP described in Section 5.3. Then, the uniform M-fat-shattering dimension of the M-SVM, computed under the additional constraint $v_b = 0$, satisfies:*

$$\textit{UM-fat}_{\Delta\mathcal{H}^\gamma}(\epsilon) \leq \mathcal{G}\textit{-dim}(\tilde{\mathcal{H}})$$

5.5 Graph dimension of the MLP

The content of this section is made up of straightforward extensions of results exposed in [7] (see also [42]). Let $\mathcal{H}_{k,l}$ (resp. $\tilde{\mathcal{H}}_{k,l}$) be the set of functions $h_{k,l}$ (resp. $\tilde{h}_{k,l}$) computed by the hyperplane (resp. hidden unit) of the MLP associated with the categories C_k and C_l. The growth function of $\tilde{\mathcal{H}}$ is trivially bounded above by:

$$\Pi_{\tilde{\mathcal{H}}}(m) \leq \Pi_{k<l} \Pi_{\tilde{\mathcal{H}}_{k,l}}(m)$$

By application of Sauer's lemma, the right-hand side of this formula is bounded by:

$$\Pi_{k<l} \Pi_{\tilde{\mathcal{H}}_{k,l}}(m) < \Pi_{k<l} \left(\frac{em}{\text{VC-dim}\left(\tilde{\mathcal{H}}_{k,l}\right)} \right)^{\text{VC-dim}\left(\tilde{\mathcal{H}}_{k,l}\right)}$$

Furthermore, by construction of the MLP, $\text{VC-dim}\left(\tilde{\mathcal{H}}_{k,l}\right) \leq \text{fat}_{\mathcal{H}_{k,l}}(\epsilon)$. Indeed, if $s_m \subset \mathcal{X}$ is shattered by $\tilde{\mathcal{H}}_{k,l}$, then $v_b = 0$ witnesses the ϵ-shattering of s_m by $\mathcal{H}_{k,l}$. Let $d_\epsilon = \sum_{k<l} \text{fat}_{\mathcal{H}_{k,l}}(\epsilon)$. Using the fact that $\sum_{i=1}^{N} -\theta_i \ln(\theta_i) \leq \ln(N)$, whenever $\theta_i > 0$, $(1 \leq i \leq N)$ and $\sum_{i=1}^{N} \theta_i = 1$, it is easily verified that:

$$\Pi_{k<l} \left(\text{fat}_{\mathcal{H}_{k,l}}(\epsilon)^{\text{fat}_{\mathcal{H}_{k,l}}(\epsilon)} \right) \geq \left(\frac{d_\epsilon}{\binom{Q}{2}} \right)^{d_\epsilon}$$

By way of consequence:

$$\Pi_{k<l} \left(\frac{em}{\text{VC-dim}\left(\tilde{\mathcal{H}}_{k,l}\right)} \right)^{\text{VC-dim}\left(\tilde{\mathcal{H}}_{k,l}\right)} \leq \left(\binom{Q}{2} \frac{em}{d_\epsilon} \right)^{d_\epsilon}$$

A simple condition to ensure that $\mathcal{G}\text{-dim}(\tilde{\mathcal{H}}) < m$ is $\Pi_{\tilde{\mathcal{H}}}(m) < 2^m$. Indeed, if $\tilde{\mathcal{H}}$ cannot compute 2^m different functions on a set s_m of size m, then it cannot shatter it either. Given the former results, this is satisfied when $\left(\binom{Q}{2} \frac{em}{d_\epsilon} \right)^{d_\epsilon} \leq 2^m$. An appropriate value of m is $m = 2d_\epsilon \log_2 \left(\binom{Q}{2} e \right)$. By transitivity, the whole set of upper bounds of this section eventually provides us with the desired bound on $\mathcal{G}\text{-dim}(\tilde{\mathcal{H}})$.

Theorem 6 *Let $\tilde{\mathcal{H}}$ be the set of functions computed by the MLP described in Section 5.3 and let $\mathcal{H}_{k,l}$, $(1 \leq k < l \leq Q)$, be the sets of affine functions associated with its hidden units. Let $d_\epsilon = \sum_{k<l} fat_{\mathcal{H}_{k,l}}(\epsilon)$. Then the graph dimension of the MLP is bounded from above by:*

$$\mathcal{G}\text{-}dim(\tilde{\mathcal{H}}) < 2d_\epsilon \log_2 \left(\binom{Q}{2} e \right) \tag{24}$$

This last theorem, in conjunction with Theorem 5, highlights the fact that the M-fat-shattering dimension of a M-SVM can be simply bounded in terms of the fat-shattering dimension of linear classifiers (sets of hyperplanes). This is very satisfactory indeed, since sharp bounds have already been derived for these models, which take into account the constraints on the parameters. This is the subject of the next section.

5.6 Fat-shattering dimension of linear classifiers with bounded $\|w\|$

Many pathways can be followed to derive bounds on the error expectation of standard support vector machines (see for instance [49]). Estimating the fat-shattering dimension of these machine has been the subject of several studies. One of them can be found in [22], where the author extends results published in [23]. Theorem 4.6 in [6] (see also Theorem 4.16 in [15]) provides the tighest result available so far. Note that the proof make use of the same probabilistic argument used by Gurvits, involving a Rademacher's sequence (see for instance [30]). Both bounds represent an improvement over Corollary 5.4 in [43]. We discuss here the theorem of Bartlett and Shawe-Taylor, in order to characterize more specifically the incidence of a constraint on $\|w\|$. It is dealing with a linear (not affine) model. However, this restriction does not rise any difficulty, since an affine model can always be transformed into a linear one by adding a component to both the vector w and the vector of features $\Phi(x)$. Note however that the consequences of the change of feature space must be taken into account in the subsequent computations.

Theorem 7 (Bartlett and Shawe-Taylor, Theorem 4.6 in [6]) *Suppose that $\Phi(\mathcal{X})$ is the ball of radius $\Lambda_{\Phi(\mathcal{X})}$ in $E_{\Phi(\mathcal{X})}$ and consider the set $\mathcal{H}$ of linear functions h such that $h(x) = w^T\Phi(x)$ with $\|w\| \leq 1$. Then, for all $\epsilon > 0$,*

$$fat_{\mathcal{H}}(\epsilon) \leq \left(\frac{\Lambda_{\Phi(\mathcal{X})}}{\epsilon}\right)^2$$

This theorem can be readily generalized to take into account a parametrized constraint on $\|w\|$. In practice, changing $\|w\| \leq 1$ for $\|w\| \leq \Lambda_w$ only affects Lemma 4.2 in [6], whereas Lemma 4.3 remains unchanged. We then get:

Theorem 8 *Suppose that $\Phi(\mathcal{X})$ is the ball of radius $\Lambda_{\Phi(\mathcal{X})}$ in $E_{\Phi(\mathcal{X})}$ and consider the set $\mathcal{H}$ of linear functions h such that $h(x) = w^T\Phi(x)$ with $\|w\| \leq \Lambda_w$. Then, for all $\epsilon > 0$,*

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

5.7 Discussion

This last theorem highlights the fact that the functional $\max_{k<l} \|w_k - w_l\|^2$, or alternatively $\sum_{k<l} \|w_k - w_l\|^2$, play for M-SVMs a role similar to the one played by $\|w\|^2$ for the standard

binary SVMs. 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 [21, 19], is thus once more justified. In [53] as well as in [48], Chapter 10, Section 10, the functional selected to perform the capacity control is slightly different, since it is $\sum_{k=1}^{Q} \|w_k\|^2$, whereas in [11], the authors used instead $\sum_{k<l}^{Q} \|w_k - w_l\|^2 + \sum_{k=1}^{Q} \|w_k\|^2$. Can the theorems derived in this report justify these choices as well? This is the case indeed, since it was proved in [19] (see also [26]) that the last three aforementioned choices are equivalent (correspond to the same optimal solution), provided the value of the soft margin parameter C is appropriately modified. As for the M-SVMs described in [14, 31], relating their objective function to the framework described above requires additional work. However, the guaranteed risk which can be derived from the four major theorems of the paper can be directly used to compare their generalization performance with the one of the other M-SVMs. We have thus endowed all the M-SVMs published so far with a well founded theoretical justification which makes it possible to compare on a fair basis their performance.

6 Alternative Approaches

In the three preceding sections, the guaranteed risk of interest has been studied according to a standard strategy, which can be summarized as follows. First, express the confidence interval in terms of a capacity measure (Theorem 2). Second, relate this capacity measure to an extended notion of VC dimension, by means of a generalized Sauer's lemma (Theorem 4). Third, characterize the behaviour of this VC dimension as a function of the constraints on the model parameters (Theorems 5, 6 and 8). Recently, Williamson and co-workers have introduced an alternative approach in [55] (see also [45, 54, 41]). It is based on functional analysis results on the compactness of operators (see for instance [12]). The covering numbers are determined via the entropy numbers of a linear operator. The main advantage of this strategy rests on the fact that it makes no use of a combinatorial dimension, and is thus more "direct". With fewer partial bounds, the confidence interval should *a priori* be tighter. We already made use of this work in [17], to pave the way for a theoretical study of M-SVMs. A comparison with the results of this paper is currently underway. A more diverging possibility consists in deriving bounds based on data dependent capacity measures such as the empirical VC entropy. In this field, the most promising studies are probably those dealing with concentration inequalities, and especially [9, 10]. More generally, the study of model selection based on penalized empirical loss minimization, as presented for instance in [5], should also prove particularly fruitful.

7 Conclusions and Future Work

This paper has highlighted a pathway to bound the covering numbers of sets of vector-valued functions used to perform multi-class discriminant analysis. The resulting bound, involving an extended notion of fat-shattering dimension, has been applied to the architecture shared

by the different multi-class SVMs developed so far. This has enabled us to cast them into a unified theoretical framework. From there, one could compare them or put forward new arguments to justify *a posteriori* the choice of the structure on which these machines are based, i.e. the choice of their objective functions. Our results could also be used to design new machines.

However, a precise bound on the M-fat-shattering dimension of M-SVMs is still to be derived. Furthermore, the use of less conservative guaranteed risks should tell us more about the precise behaviour of these machines. In that respect, significant benefits should be expected from extending to the multi-class case the approaches evoqued in the preceding subsection. This extension, and the subsequent comparison, are the subject of an ongoing work.

Acknowledgements
The author gratefully acknowledges the support of the ESPRIT funded Working Group N. 27150 "Neural Networks and Computational Learning Theory". Thanks are also due to E. Domenjoud 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] N. Alon, S. Ben-David, N. Cesa-Bianchi, and D. Haussler. Scale-sensitive dimensions, uniform convergence, and learnability. *J. ACM*, 44:615–631, 1997.

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

[4] 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.

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

[6] 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*. The MIT Press, Cambridge, 1999.

[7] E.B. Baum and D. Haussler. What size net gives valid generalization? *Neural Computation*, 1:151–160, 1989.

[8] 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.

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

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

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

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

[13] 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.

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

[15] N. Cristianini and J. Shawe-Taylor. *An Introduction to Support Vector Machines and other kernel-based learning methods*. Cambridge University Press, 2000.

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

[17] 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).

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

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

[20] 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.

[21] 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.

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

[23] L. Gurvits and P. Koiran. Approximation and learning of convex superpositions. *Journal of Computer and System Sciences*, 55(1):161–170, 1997.

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

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

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

[27] 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.

[28] 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.

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

[30] M. Ledoux and M. Talagrand. *Probability in Banach Spaces.* Springer-Verlag, Berlin, 1991.

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

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

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

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

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

[36] 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.

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

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

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

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

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

[42] J. Shawe-Taylor and M. Anthony. Sample sizes for multiple-output threshold networks. *Network: Computation in Neural Systems*, 2:107–117, 1991.

[43] 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.

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

[45] A.J. Smola. *Learning with Kernels.* PhD thesis, Technische Universität Berlin, 1998.

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

[47] 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.

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

[49] V.N. Vapnik and 0. Chapelle. Bounds on error expectation for support vector machines. *Neural Computation*, 12(9):2013–2036, 2000.

[50] 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*, 16:264–280, 1971.

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

[52] 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.

[53] 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.

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

[55] 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.

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 : Domaine de Voluceau - Rocquencourt - BP 105 - 78153 Le Chesnay 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 - 38330 Montbonnot-St-Martin (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